UNIVERSITA' DEGLI STUDI DI VERONA

SCUOLA DI DOTTORATO IN SCIENZE DELLA VITA E DELLA SALUTE

DIPARTIMENTO SCIENTIFICO E TECNOLOGICO

DOTTORATO DI RICERCA IN BIOTECNOLOGIE MOLECOLARI, INDUSTRIALI ED AMBIENTALI

Curriculum: biologia molecolare vegetale

CICLO XXII

Genetic selection for flowering time traits during speciation of tomato

Coordinatore: Ch.mo Prof. Roberto Bassi

Tutor interno: Ch.mo Prof. Roberto Bassi

Tutor esterno: Ch.mo Prof. Giovanni Giuliano

Dottorando: Dott. Elio Fantini

SUMMARY

In the last decade, a coordinated network of knowledge about the Solanaceae family has been created by the International Solanaceae Genome Project (SOL), whose final purpose is to explain how a common set of genes/proteins can result in a wide range of morphologically and ecologically distinct organisms like those in the Solanaceae family. This taxon includes more than 3000 species many of which evolved in the Andean/Amazonian regions of South America. Their habitats vary dramatically, from rain forests to deserts to high mountains. Moreover, the Solanaceae is the third most valuable crop family exceeded only by the grasses (e.g. rice, maize, wheat) and legumes (e.g. soybean), and the most valuable in terms of vegetable crops. It includes, among others, tomato, potato, eggplant, pepper, petunia, tobacco.

Among Solanaceae, tomato was selected as a reference since it provides the smallest diploid genome (950 Mb) for which homozygous inbreds are available, as well as an advanced BAC-based physical map to start the sequencing. It also offers the vantage of short generation time, routine transformation technology, and availability of rich genetic and genomic resources.

In this project, we merged structural and functional genomic approaches, in order to study in tomato one of the most important characters in terms of fitness and adaptation: flowering. We have focused our efforts in the study of genes involved in the photoperiodic regulatory pathway, since some wild tomato species, that growth between 0 and -25 degrees of latitude and between 0 and 3700 m of elevation, show different photoperiodic responses. These genes are members of three gene families, whose orthologs in Arabidopsis and rice play a key role in the regulation of flowering in dependence of day length. They are the *TCOL* (tomato *CONSTANS*-like), the *CRYPTOCHROME* and the *GIGANTEA* gene families.

In chapter 2 we provide structural information on the organization and expression of the *CRY, COL* and *GI* gene families in tomato and we analyse the microsynteny with Arabidopsis. We also identify four new *COL* and one new *GI* gene.

In chapter 3 we investigate the sequence diversification of the three gene families during speciation in the tomato clade, using both sequencing and expression profiling approaches. In particular, we observed a high degree of diversifying selection for *TCOL3*, that presents also a geographical cline in the frequency of synonymous mutations, suggesting a role of this gene in adaptation to low latitudes.

Finally, in chapter 4, we use four reverse genetic approaches (VIGS, RNAi, TILLING and overexpression) for a functional characterization of the tomato Cryptochrome gene family. Novel *cry1b-* and *cry2-* mutants and RNAi plants and *CRY1a* overexpressors are described for the first time and novel developmental and flowering phenotypes are attributed to the various genes.

CONTENTS

CHAPTER 1

INTRODUCTION TO TOMATO AND FLOWERING

1.1 The wolf peach

1.2 Domestication

1.3 Nutritional and economic importance

1.4 Tomato as model species

1.5 Tomato wild species

1.6 Transition to flowering

1.7 Flowering and photoperiod

    1.7.1 Do plants measure daylength or nightlength?

    1.7.2 LDP model: Arabidopsis

    1.7.3 SDP model: rice

1.1 The wolf peach

Tomato is a perennial plant, which is cultivated as an annual crop. It belongs to the *Solanaceae* family, like potato, eggplant, pepper, tobacco and petunia. Morphological descriptions made from herbarium specimens placed tomato as belonging to the genus *Lycopersicon.* In 1694, Tournefort gave to cultivated tomatoes the name Lycopersicon, "wolf peach" in Greek, because in old German folklore, witches used plants of the nightshade family to evoke werewolves, a practice known as lycanthropy. In the 18th century Carl Linnaeus conjured up binomial nomenclature to name species and chose for tomato the name *Solanum lycopersicum* [1]. In 1768, Philip Miller changed the name in *Lycopersicon esculentum* (literally "edible wolf peach"), supporting that the differences from other *Solanum* species were relevant and enough to justify the different nomenclature [2]. A recent study, based on genetic and molecular markers, has shown that it is deeply nested in the *Solanum* genus, forming the sister clade to potato [3] . For this reason, *Lycopersicon esculentum* Mill. has been renamed *Solanum lycopersicum* L.

1.2 Domestication

Although wild tomato species are distributed in a relatively small habitat, the Andean region, the site of domestication remain uncertain. Two hypotheses have been proposed: Peru [4] and southern Mexico [5] [6]. The Mexican domestication hypothesis suggests that wild cherry tomatoes (*S. lycopersicum* var. cerasiforme) migrated from Peru into Meso-America, and became domesticated in Mexico (Fig.1) [5] [7].
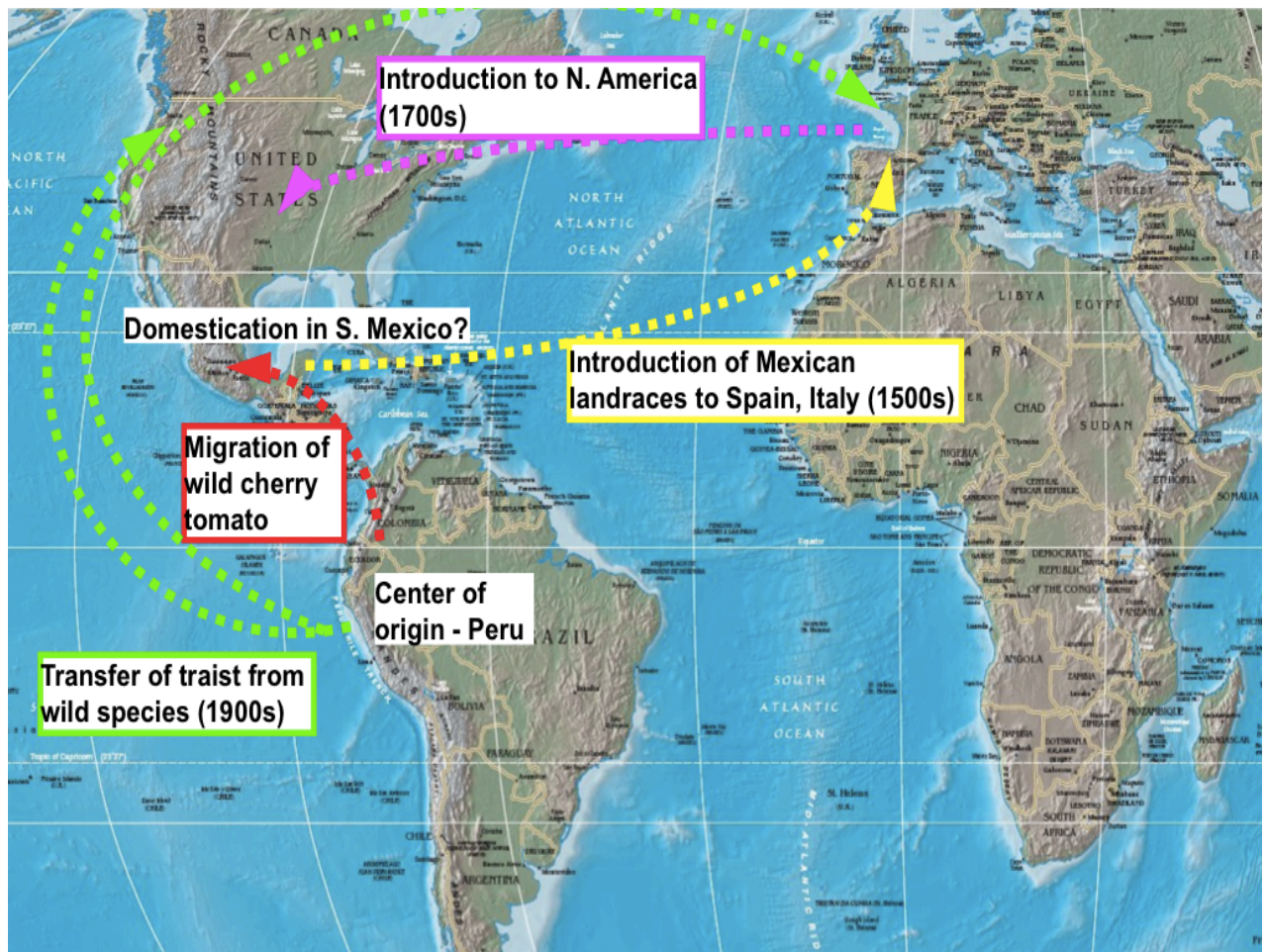
Figure 1. Origins and domestication of tomato. (Courtesy of Prof. Roger Chetelat – Department of Plant Sciences – UC Davis – USA)

However, there is little information on when domestication occurred; during the conquest of Mexico in 1521, the Spanish observed that large-fruited types, a sign of human selection, were already being grown for food [8]. Linguistic records also support the theory of domestication in Central rather than South America. The modern name "tomato" is derived from the word "tomatl", that defines this plant in the native language of the Aztecs [9]. In addition to these cultural considerations, genetic evidence also supports Central American domestication. Data from allozymes indicate that older European cultivars, that descend from tomatoes introduced by the Spanish explorers, are extremely homogeneous at most loci and are closely related to cultivars and wild cherry tomatoes from southern Mexico and Central America, but differ from those of South America [10].

1.3 Nutritional and economic importance

The tomato fruit and its products (e.g. ketchup, juices, soups and sauces) are widely utilized in the Western diet (8.21 kg of fruits and 32.6 kg of products per capita in 2003) [11]. Therefore, they have become the principal source of vitamin C (19 mg/100 g of fresh weight), provitamin A (623 IU/100 g of fresh weight), and lycopene (3.0 mg/100 g of fresh weight) in the Western diet. The main flavonoids found in tomatoes are quercetin, kaempferol, and naringenin, with quercetin levels ranging from 0.03−2.77 mg/100 g in fresh tomatoes to 4.77 mg/100 g in processed tomato products. Fresh and processed tomato products supply about 2.0 g of quercetin annually per capita [12]. It also contains significant amounts of dietary fiber, iron, magnesium, niacin, potassium, phosphorus, riboflavin, sodium and thiamine [13].

Tomato is low in saturated fat, cholesterol and sodium [14]. And unlike most foods, cooking or processing of tomato is beneficial to health. It increases the bioavailability of lycopene (e.g. in tomato paste, catsup, tomato soup, tomato sauce). This is because heating up tomato breaks down its cell walls and releases more lycopene. Eating tomatoes has more benefits (with all of its other ingredients) than taking lycopene alone [15].

The tomato is now grown worldwide for its edible fruits, with thousands of cultivars having been selected with varying fruit types, and for optimum growth in differing growing conditions. About 130 million tons of tomatoes were produced in the world in 2008 [16] (Tab. 1).

Top Tomato Producers — 2008 (in tonnes)	
China	33 811 702
United States	12 575 900
Turkey	10 985 400
India	10 260 600
Italy	5 976 912
World Total	129 649 883

Table 1. Tomato top producers in 2008 according to FAOSTAT data.

1.4 Tomato as a model plant species

Together with its economic importance, tomato has also become a model plant for research purposes [17] [18] [19] [20]:

1. It is easy to cultivate in a wide range of environmental conditions, has a short life cycle and lends itself to horticultural manipulation including grafting or cutting. It grows as an indeterminate plant due to reiterate switches from vegetative to reproductive stages.

2. Various types of explants can be cultured *in vitro* and plant regeneration is feasible, allowing efficient transformation procedures.

3. It has several features that distinguish it from other model plant species: it is phylogenetically distant from maize, *Arabidopsis*, snapdragon, rice, *Medicago* or poplar.

4. It is genetically the best characterized species bearing a fleshy fruit..

5. The presence of related, wild interfertile species affords a great wealth of readily accessible germplasm and can be used as a source of genetic variability.

6. It is a basic diploid with a small genome size (0.9 pg or 950 Mbases per haploid genome) [21].

7. Stock collections and a large set of mutants are available (http://tgrc.ucdavis.edu/). Moreover, induced mutations, mainly generated through chemical ethylmethane sulfonate (EMS) and irradiation, are providing available screening populations and the possibility to identify new developmental genes (http:// zamir.sgn.cornell.edu/mutants/ - http://www.agrobios.it/tilling/index.html - http://tilling.ucdavis.edu/index.php/Tomato_Tilling).

8. The development of genomic and sequencing resources (genetic and physical maps, ESTs, microarrays and a genome sequence (see chapter 2), have favoured an international sequencing project and are contributing to the current progress in understanding the biological bases of plant development (http://www.sgn.cornell.edu/).

1.5 Tomato wild species

The tomato clade is evolutionarily a young group that has diversified to occupy a great variety of habitats. The age of the genus *Solanum* is estimated at ~12 million years (My) based on nuclear (18S rDNA) and chloroplast markers ribulose-bisphosphate carboxylase large subunit (rbcL) and ATP Synthase B (atpB) [22], and the radiation of the tomato clade has been estimated as ~7 My, based on four nuclear genes [23]. During this time, tomato species have evolved and occupied various habitats of the western coast of South America, from central Ecuador to northern Chile, including the Galapagos Islands. They range from sea level to above 3,000m in altitude, within various grades of xeric to mesic environments [24].

Based on morphological characters, phylogenetic relationships, and geographic distribution, 13 species (Tab. 2), including the cultivated tomato (*Solanum lycopersicum*), and four closely related species have been recognized as part of the tomato clade (Fig. 2) [25].
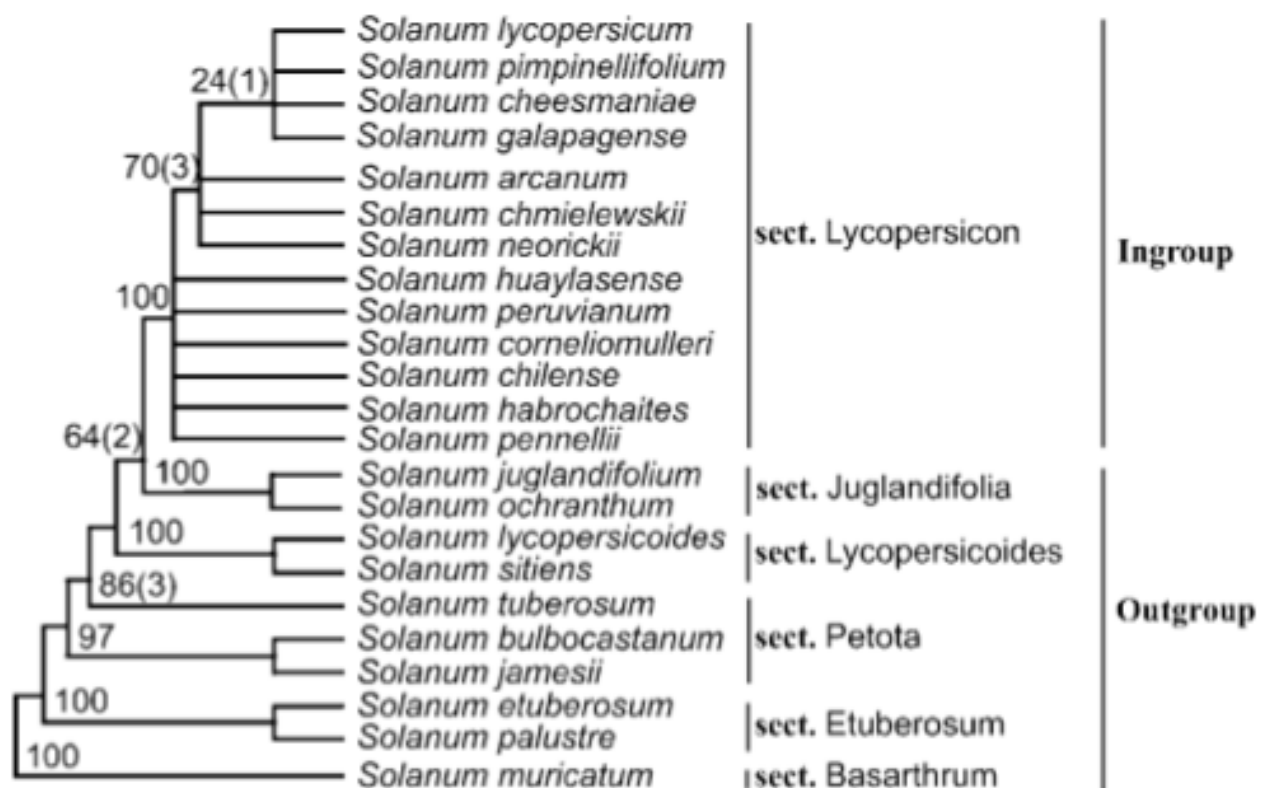

Figure 2. Abstracted cladistic results of the 65 accessions of the 13 tomato species (ingroup sect. Lycopersicon) and ten outgroup taxa examined in the phylogenetic analysis of the GBSSI gene sequences by Peralta and Spooner (2001). Numbers indicate bootstrap values, and decay values are indicated between parentheses [26].

10

Name in Peralta et al. (2007)	Lycopersicon equivalent	Distribution and habitats	Breeding system	Fruit color
S. lycopersicum	*L. esculentum*	Apparently native to Perú; the domesticated form of S. lycopersicum now occurs worldwide. The cherry tomato, commonly known as S. lycopersicum var. cerasiforme, has been suggested as the ancestor of cultivated tomato and can often be found growing as a weed in temperate habitats and the edges of cultivated fields, where it is not necessarily native; recent studies suggest S. lycopersicum var. cerasiforme may be a mixture of wild and cultivated tomatoes, rather than an ancestor (Nesbitt and Tanksley 2002)	SC, autogamous, facultative allogamous	Red
S. pimpinellifolium	*L. pimpinellifolium*	Apparently native to coastal areas from Central Ecuador to southern Peru, 0–500 m, parts of Chile2, adventive in N. America. Grows in humid places and on the edges of cultivated fields throughout its native range and has apparently escaped from cultivation in the Galápagos	SC, autogamous, facultative allogamous	Red
S. cheesmaniae	*L. cheesmaniae*	Endemic to the Galápagos Islands (Ecuador) from sea level to 1,300 m	SC, exclusively autogamous	Yellow, orange
S. galapagense	Part of *L. cheesmaniae*; also known as L. glandulosum f. minor	Endemic to the Galápagos Islands, particularly the western and southern islands, mostly occurring on coastal lava and on volcanic slopes, sea-level to 650 m, exceptionally up to 1,500 m in Fernandina and Santiago Islands	SC, exclusively autogamous	Yellow, orange
S. arcanum	Part of *L. peruvianum*	Northern Peru. Coastal and inland Andean valleys, on dry rocky slopes, 100 to 2,500 m	Typically SI, allogamous, rare populations SC, autogamous, facultative allogamous	Typically green with dark green stripes
S. chmielewskii	*L. chmielewskii*	Southern Peru to northern Bolivia (Sorata). In high dry Andean valleys, 2,300–3,000 m	SC, facultative allogamous	Typically green with dark green stripes
S. neorickii	*L. parviflorum*	Southern Ecuador to southern Perú. In dry Andean valleys, 1,950–3,000 m, often growing over rocky banks and roadsides. Sometimes found in sympatry with S. chmielewskii	SC, highly autogamous	Typically green with dark green stripes
S. huaylasense	Part of *L. peruvianum*	Northern Perú (Department of Ancash). On the rocky slopes along rivers, 1,700–3,000 m	Typically SI, allogamous	Typically green with dark green stripes
S. peruvianum	*L. peruvianum*	Central Peru to northern Chile. In lomas formations and occasionally in coastal deserts from sea level to 600 m, sometimes growing as a weed at field edges in coastal river valleys	Typically SI, allogamous	Typically green to greenish-white, sometimes flushed with purple
S. corneliomuelleri	Part of *L. peruvianum*; also known as *L. glandulosum*	Central to southern Perú. On western slopes of the Andes, (400) 1,000–3,000 m, and on lower slopes on the edges of landslides (huaycos) towards the southern range of the species distribution	Typically SI, allogamous	Typically green with dark green or purple stripes, sometimes flushed with purple
S. chilense	*L. chilense*	Southern Peru to northern Chile. On western slopes of the Andes, hyper-arid rocky plains, dry river beds, and coastal deserts, from sea level to 3,000 m	SI, allogamous	Green to whitish green with purple stripes
S. habrochaites	*L. hirsutum*	Central Ecuador to Central Peru. In premontane forests to dry forests on the western slopes of the Andes, occasionally in lomas formations in northern Peru, 400–3,600 m	Typically SI, with SC populations in N and S of species range	Green with darker green stripes
S. pennellii	*L. pennellii*	Northern Peru to northern Chile, in dry rocky hillsides and sandy areas, from sea level to 2,850 m	Usually SI, some SC in South of species range	Green
S. lycopersicoides	*L. lycopersicoides*	Southern Peru to northern Chile on the western slopes of the Andes on dry rocky hillsides, 1,500–3,700 m	SI, allogamous	Green-yellow when maturing, black when ripe
S. sitiens	*L. sitiens*	Northern Chile, western Andean slopes on rocky hillsides and dry quebradas, 2,500–3,500 m	SI, allogamous	Green-yellow when maturing, brown and dry when ripe
S. juglandifolium	*L. juglandifolium*	Northeastern Colombia to southern Ecuador, on the edges of forest clearings, open areas and roadsides, 1,200–3,100 m	SI, allogamous	Green to yellow-green
S. ochranthum	*L. ochranthum*	Central Colombia to southern Peru, in montane forests and riparian sites, 1,400–3,660 m	SI, allogamous	Green to yellow-green

Table 2. Species list for tomatoes and wild relatives (with equivalents in the previously recognized genus Lycopersicon, now part of a monophyletic Solanum), along with characteristic fruit colour, breeding system, and distribution [26].

Wild tomato species are herbaceous plants, although they can also undergo secondary growth at the base of the stems and the main root. They are perennial and most of them flower regardless of day length, although in their natural habitats some wild tomatoes behave as annuals, probably because frost or drought kills the plants after the first growing season [27]. Some species flower only in short days if growth at mid-latitudes (Tab. 3).

Species	Daylength Preference	Sowing Date (Davis)*	Mating System	Pollination Method	# Plants / Gener.	# Plants per gal. pot	Notes
L. cheesmanii	short day	Nov - wk 4	autogamous (SC)	self	10	2	Seed produced under low light conditions is of poor quality.
L. chilense	short day	July - wk 2	allogamous (SI)	mass sib	50	5	
L. chmielewskii	day neutral	May - wk 2	facultative (SC)	mass sib	50	5	
L. esculentum var. cerasiforme	day neutral	April - wk 2	autogamous (SC)	self	6	(field)	
L. hirsutum	short day	July - wk 4				3	Forms edema on leaves under high humidity.
f. glabratum			facultative (SC)	self	15		
f. typicum			allogamous (SI) or facultative	mass sib	50		
L. parviflorum	day neutral	May - wk 2	autogamous (SC)	self	15	3	
L. pennellii	day neutral	June - wk 1	allogamous (SI) or facultative (SC)	mass sib	50	5	Use well-drained soil and water sparingly.
L. peruvianum	mostly day neutral	June - wk 4	allogamous (SI) or facultative (SC)	mass sib	50	5	f. glandulosum and mountain races are short day.
L. pimpinellifolium							
selfing pops:	day neutral	April - wk 2	autogamous (SC)	self	6	(field)	
--------------------	---------------	-------------	----------------------	------------	-------	------------	---------------------------------------
outcrossing pops:	mostly short day	Feb - wk 2	facultative (SC)	mass sib	50	5	Regenerate in greenhouse to limit outcrossing.
S. juglandifolium	short day	Aug - wk 2	allogamous (SI)	mass sib	50	3 (2 gal pot)	Graft onto LA4135; water heavily; 8 mo. seed maturation.
S. lycopersicoides	short day	Aug - wk 2	allogamous (SI)	mass sib	50	3	Pollinate in spring months; 6 mo. seed maturation.
S. ochranthum	short day	Aug - wk 2	allogamous (SI)	mass sib	50	2 (2 gal pot)	Graft onto LA4135; water heavily; 10 mo. seed maturation.
S. sitiens	short day (almost neutral)	Aug - wk 2	allogamous (SI)	mass sib	50	3	Graft onto LA4135; 6 mo. seed maturation.

Table 3. Recommendations for flowering and reproducing wild tomato species from http://tgrc.ucdavis.edu/spprecommed.html

Tomato leaves are characterized as pinnate with a continuum of leaf dissection (Fig. 3a). All members of the genus have perfect flowers (hermaphroditic) (Fig. 3b) and the breeding system varies from allogamous self-incompatible to facultative allogamous and self-compatible, to autogamous and self-compatible. The self-incompatibility system in tomatoes is gametophytic and controlled by a single, multiallelic S locus [28]. Fruits of all species are globose, bilocular berries, with the exception of few wild populations and many cultivated forms of *S. lycopersicum*, which are multilocular and can occur in many unusual shapes [29] (Fig. 3c).

Figure 3. Leaves (a), flowers (b) and fruits (c) of *Solanum* sect. *Lycopersicon*, sect. *Juglandifolia* and sect. *Lycopersicoides*. **A**. *S. lycopersicum*; **B**. *S. pimpinellifolium*; **C**. *S. cheesmaniae*; **D**. *S. galapagense*; **E**. *S. neorickii*; **F**. *S. chmielewskii*; **G**. *S. arcanum*; **H**. *S. huaylasense*; **I**. *S. peruvianum*; **J**. *S. corneliomulleri*; **K**. *S. chilense*; **L**. *S. habrochaites*; **M**. *S. pennellii*; **N**. *S. ochranthum*; **O**. *S. juglandifolium*; **P**. *S. lycopersicoides*; **Q**. *s. sitiens*. Scale bars 3a: A, B, E-K, M, P, Q, 2 cm; C, D, 1 cm; L, N, O, 3 cm. Scale bars 3b: A-D, H-Q, 1 cm; E-G, 0,5 cm. Scale bars 3c: 1cm. [30]

13

1.6 Transition to flowering

In recent years, a great deal of progress has been made in understanding the molecular mechanisms that regulate flowering time, particularly in Arabidopsis. The transition to flowering is influenced by both endogenous and exogenous signals. A genetic regulatory network that integrates these signals has been elucidated in Arabidopsis using two complementary approaches: comparisons between naturally occurring ecotypes and the genetic analysis of mutations that result in early or late flowering phenotypes [31]. Genes have been positioned in several pathways that promote or repress flowering, depending on environmental or autonomous conditions, and how these pathways interact is an area of active study. Many models have been constructed and refined to obtain a network for flowering time control [32] [33] [34]. An integrated view [35] is shown in fig. 4. Four principal pathways (photoperiod, autonomous, vernalization, and gibberellin) converge to few key genes that integrate signals on the meristem identity genes *LEAFY* (*LFY*) and *APETALA1* (*AP1*). Under appropriate conditions, their activities lead promote *LFY* and *AP1* expression, which then mediate the switch to reproductive development in the shoot meristem.

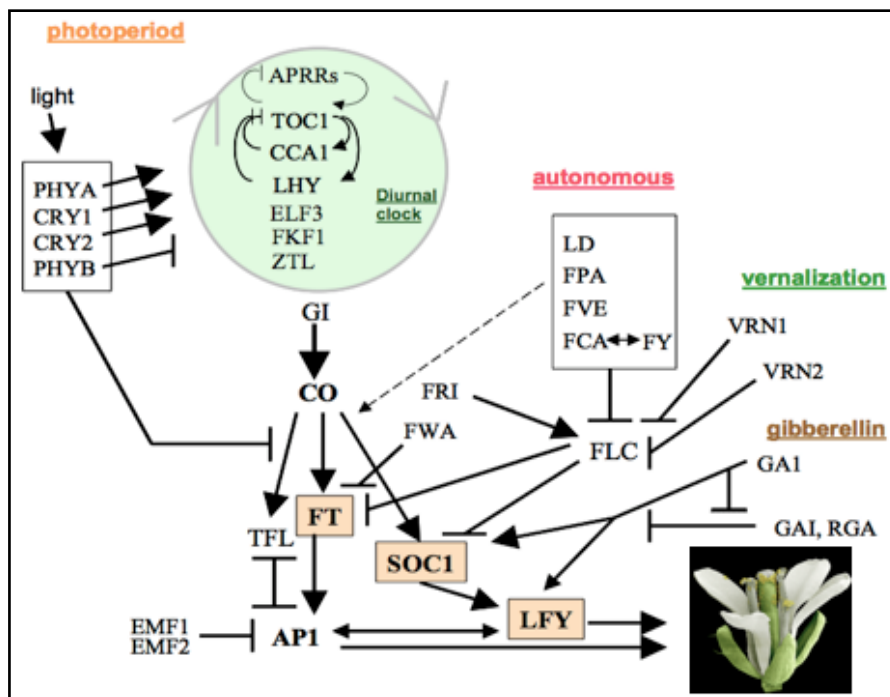**Figure 4.** The genetic regulatory network controlling flowering time in Arabidopsis. The positive (arrows) and negative (bars) regulatory relationships are from reviews cited in the text. Genes in tinted boxes (FT, LFY and SOC1) integrate signals from multiple pathways. The dashed line depicts a possible connection between the autonomous pathway and the photoperiod pathway that is independent of FLC. [35]

In tomato, the main axis of the young plant is monopodial but, once the shoot apex is induced to flower and forms an inflorescence, further growth is sympodial. This latter growth is characterized by 3-4 nodes of vegetative growth before a terminal inflorescence is formed [36] (Fig. 5).
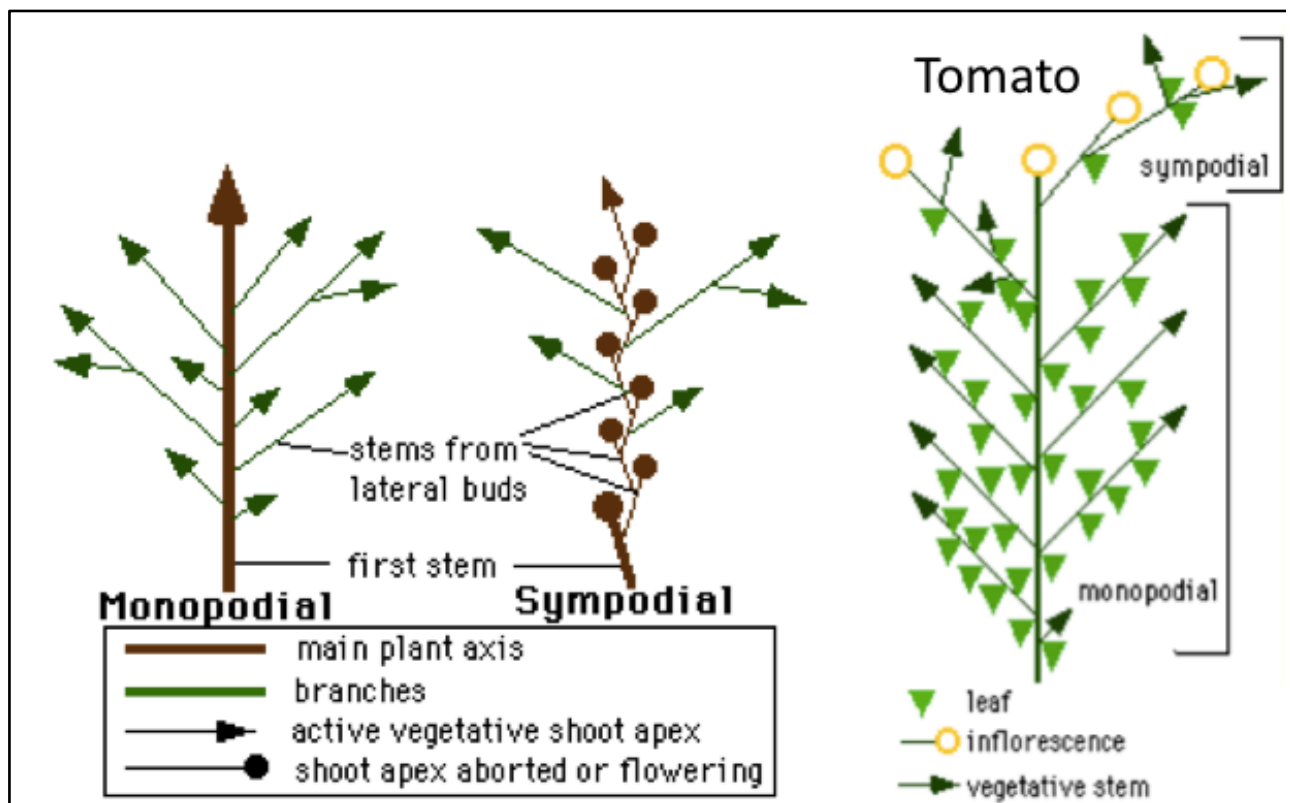
Figure 5. Branching in Tomato [36].

The development of the first inflorescence is under the control of environmental and endogenous factors, the latter being of genetic and hormonal nature. Cultivated tomato is considered a day-neutral plant since the time to flowering, as measured by the number of leaves developed before floral transition, is not affected by photoperiod [37]. In fact, some genes have been identified and characterized as members of an autonomous pathway controlling floral transition. In the initial monopodial segment, *FALSIFLORA* (*FA*) and *SINGLE FLOWER TRUSS* (*SFT*) promote floral transition [38] [39] [40], while *SP* regulates this process in the sympodial segments [41]. Mutations at either *FA* or *SFT* loci result in a photoperiod-independent late flowering phenotype and in abnormalities affecting inflorescence development. *FA* is orthologous to *LFY* [38]. *SFT* is orthologous to the Arabidopsis *FT* and triggers systemic signals that regulate floral transition and sympodial growth in tomato [42]. Double *sft fa* mutants are unable to flower, suggesting that *FA*

15

and *SFT* regulate floral transition by independent pathways. This agrees with the fact that *FA* expression is not affected by the *sft* mutation [39]. The *fa* mutation produces the opposite effect to *sp*, i.e. an increased number of vegetative nodes in the first sympodial segments. This feature, and the expression pattern of *FA*, proves its involvement in sympodial development. Although the function of *SP* seems to be antagonistic to *FA*, the vegetative-to-reproductive switch in the sympodial segment may depend on a balance between *FA* and *SP* transcription levels [38]. Likewise, a balance in the activities of *SFT* and *SP* could be responsible for the floral transition of both initial and sympodial segments [40].

Although tomato flowers autonomously, environmental cues can modify this developmental pattern. Low temperatures (10- 15°C) reduce the number of nodes up to the first inflorescence, and also the rate at which these leaves are produced. Similarly, a scarce but significant reduction in flowering time has been observed in many cultivars grown under short day conditions [43]. Most importantly, high irradiance accelerates flowering, an effect associated with a higher rate of leaf initiation and an increased assimilate availability in the meristem [37] [44]. Late flowering mutants *uniflora* (*uf*) and *compound inflorescence* (*s*) show enhanced phenotypes under winter conditions, namely low irradiance (daily light energy integral) and poor assimilate availability perceived by the apical meristem [45] [46]. As the *fa sft* double mutant, introducing *sft* into the *uf* background completely suppresses floral transition, which suggests that *UF* and *SFT* promote flowering but participating in parallel regulatory pathways [40]. Although the genetic interactions between *UF* and *FA* remain to be clarified, *uf* is epistatic over most of flowering mutations [47], indicating that *UF* is a key regulator of tomato flowering. Moreover, *UF* might function upstream to *SFT* as is indicated by the fact that constitutive expression of *SFT* rescues the flowering time phenotype of the *uniflora* mutant, substituting its high light requirements [42].

Light is perceived by the plant through photoreceptors, particularly red/far red light PHYTOCHROME (PHY) and blue light CRYPTOCHROME (CRY) receptors. Overexpression of tomato *CRY2* does not alter the number of vegetative segments before the first flower but does

increase the number of days to the first floral anthesis [48]. Interestingly, a QTL mapping approach has revealed that the *PHY2B* gene, as well as *FALSIFLORA*, co-localize with major QTLs responsible for floral transition in this species*,* making them candidate genes for the domestication process of tomato [49].

Flowering of tomato is also modulated by hormones, though their roles during floral transition have been poorly studied. Gibberellins (GAs) promote tomato flowering since GA-deficient mutants require exogenous gibberellins to flower [50]. Elevated GA contents increase the number of leaves before flowering and the rate of leaf initiation [37]. Furthermore, *in vitro* experiments show that cytokinin-mediated stimulating effect on floral initiation could be inhibited by GA. Most probably, plant hormones modulate tomato flowering by gene interactions involving several regulatory pathways although the nature of such interactions is still unknown.

Most flowering mutants characterized in tomato flower later than the corresponding wild type plants. Of fourty one flowering mutants identified by an "*in silico*" screening, only four showed an early flowering phenotype [51]. These observations, together with the influence of environmental cues on flowering time, suggest that selection of favourable combinations of flowering genes has played a crucial role throughout the tomato domestication process, enabling early flowering and ensuring fruit yield under several conditions.

1.7 Flowering and photoperiod

The time for reproduction in plants, including flowering, tuberization, and bulbing, is often controlled by photoperiod or daylength [52] [53]. In perennial plants, the photoperiodic control of the annual growth cycle is evenly important for survival and fitness, which in temperate areas is strongly controlled by photoperiod [54] [55] [56] [57].

During the last 15 years, great progress has been made in the molecular understanding of the photoperiodic induction of flowering, in particular in the model species *Arabidopsis thaliana*. An

obvious question is to what extent the molecular mechanisms of photoperiodic induction of flowering in *A. thaliana* are also shared by other species and other traits controlled by photoperiod. Within a species, accessions may respond to different daylengths, in relation to their geographical distribution and this suggests that this trait is associated with adaptation to growth at particular latitudes. Examples of such distributions include induction of flowering by daylength in cultivated populations of soybean (*Glycine max*) [58] and natural populations of *Xanthium strumarium* [59] or repression of bud growth in poplar (Populus spp.) [60].

Already at the beginning of the 20th century, Julien Tournois and Hans Klebs suggested that daylength was more important than light quantity for the induction of flowering [61] [62]. Few years later, it was clearly demonstrated that flowering and other responses were induced by long days in some species and by short days in others [63] [52]. The term photoperiodism was coined to identify this phenomenon. Garner and Allard classified plants into three main categories based on photoperiodic response, short-day plants (SDPs), long-day plants (LDPs), and day-neutral plants (DNPs). Soon, it became clear that the photoperiodic signal was perceived in leaves [64]. In several species, leaves exposed to inductive photoperiod induced flowering when they were grafted onto plants growing in non-inductive light conditions [65]. This discovery implied that the vector of the received photoperiod signal is a transmissible signal, a floral hormone or florigen [66].

1.7.1 Do plants measure daylength or nightlength?

Even though daylength and nightlength are perfectly correlated in the 24 h cycle, plants could in principle measure the length of either the night or the day [67]. It is possible to distinguish these alternatives by varying them independently. In the SDP *Xanthium strumarium*, long nights induced flowering even if these were coupled with a long day, but short days did not induce flowering if the night was also short (Fig. 6) [68]. Further studies also showed that flowering could be prevented if the long night was interrupted with a light pulse (a night break), supporting the importance of the dark period.
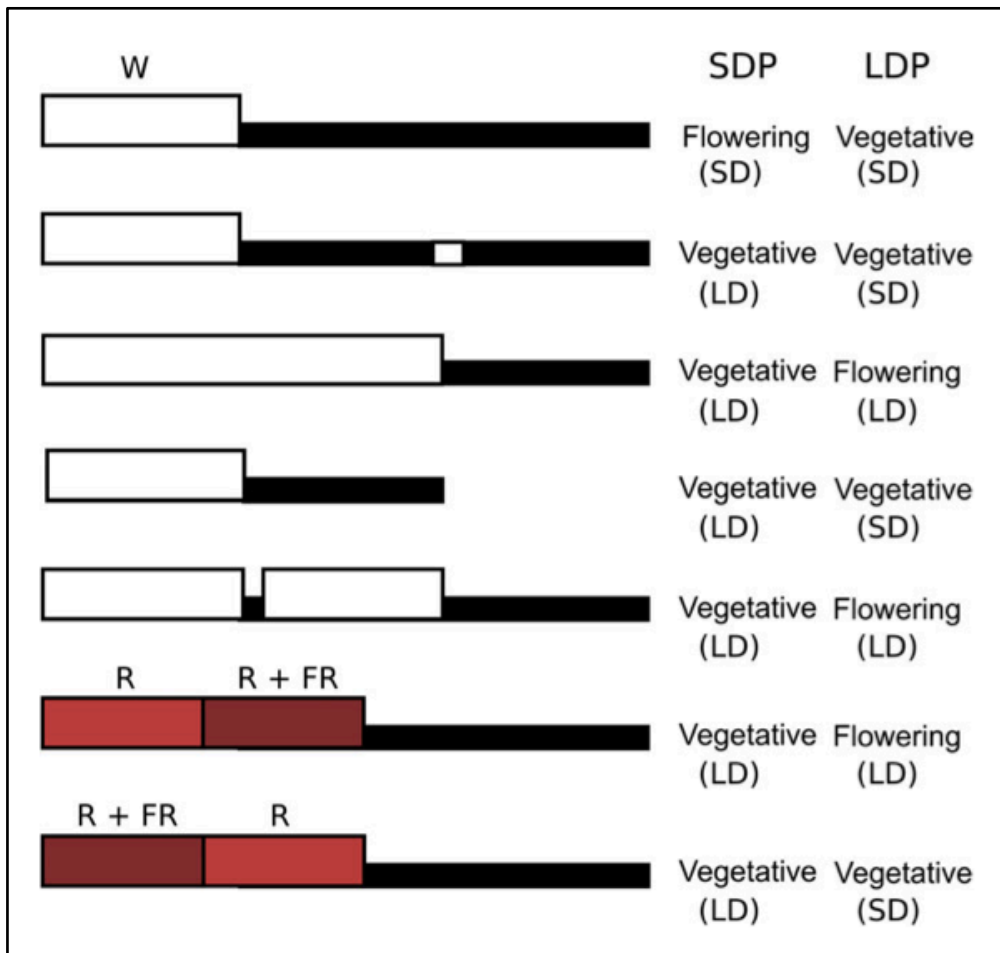
Figure 6. Schematic illustration of typical responses of SDPs and LDPs to various combinations of light and dark. W, white light; R, red light; FR, far-red light; SDP, short day (dark-dominant) plant; LDP long day (light-dominant) plant. (SD) and (LD) indicated the photoperiod perceived by SDPs and LDPs [67].

In LDPs the situation is different. Generally they are less influenced by a night break treatment [53]. In contrast to SDPs, longer night interruptions are needed, and the flowering response is often semi-quantitative in nature. Furthermore, in contrast to SDPs, most LDPs require far-red light at the end of the light period to interpret the light period as a long day [69] [53]. If instead far-red light is followed by red light in the second part of the day, promotion of flowering is poor or absent (Fig. 6). In SDPs, the light quality given in different parts of the light period has little effect on flowering.

The mechanisms controlling photoperiodic flowering are best understood in *Arabidopsis thaliana* [70] [71], but no comprehensive analysis of quantitative variation in photoperiod response within this species has been reported. Although genetic differences at loci involved in vernalization are responsible for much of the variation in flowering time among Arabidopsis accessions, allelic

differences at genes contributing to the photoperiodic response can also have important effects on flowering time. Arabidopsis is a quantitative long-day plant that flowers earlier under long days (LDs) of spring and early summer than during short days (SD) of winter. This variation was mainly characterized by comparing flowering time under extreme LDs of 16 h and SDs of 8 or 10 h [72] [73] [74]. Commonly used laboratory accessions such as Columbia (Col) and Landsberg erecta (Ler) show a marked flowering response to daylength and were used to screen for mutations that impair photoperiodic flowering [75] [76]. The genes identified by these mutations defined a pathway that promotes flowering in response to LDs. *GIGANTEA* (*GI*), *CONSTANS* (*CO*), and *FT* are central to this pathway [77] [78] [79] [80] [81]. Transcription of each of these genes is regulated by the circadian clock [78] [81] [82], while CO activity is promoted by exposure to light both at the transcriptional and posttranscriptional levels [83] [84]. This complex regulation ensures that CO activates *FT* transcription only under LDs.

Comparative studies in Arabidopsis, rice and *Ipomoea nil* have demonstrated that there is a highly conserved network of genes involved in the photoperiod pathway in both LD and SD plants [85] [86]. The basis for the differential flowering responses to photoperiod derive from duplications and changes in the expression of gene family members [87] [88], development of alternative flowering induction pathways that are unique to each species [89] [90] [88] [91] [92] and a reversal in the function of at least three genes in rice and Arabidopsis, *GIGANTEA* (*GI*), *CONSTANS* (*CO* or *HD1*), and *FLOWERING TIME LOCUS T* (*FT* or *HD3A*) [93] [94] [86] [95].

Although tremendous progress has been achieved in understanding which genes are involved in the regulation of flowering time and how different genes and gene families interact at the molecular level, our understanding of how flowering time is regulated under field conditions and how allelic variation and copy number variation in natural populations affects the flowering response of our major crop plants remains largely unknown.

1.7.2 LDP model: Arabidopsis

CONSTANS is a key protein in photoperiod sensing in Arabidopsis: A key component of this model is the *CONSTANS* gene (*CO*) [77]. The expression of *CO* is controlled by the circadian clock, with a diurnal peak of expression during the night in short days [82]. However, the CO protein is degraded in darkness, so CO function can only be obtained if *CO* mRNA is also expressed before darkness. This is achieved in long days when light is present closer to the peak of *CO* mRNA expression but also because *CO* mRNA displays a broader peak of expression in long days. The broader peak of expression results from the activity of FLAVIN-BINDING, KELCH REPEAT, F-BOX (FKF1), which degrades a repressor of *CO* [84]. The end result is that CO protein accumulates only in long days, and then activates transcription of *FT* in vascular tissue. A number of recent studies strongly support that the FT protein then moves from the leaves to the shoot apical meristem (SAM), where it activates transcription factors that induce flowering [67].

A number of photoreceptors are important for the control of CO stability in light [83]. Phytochrome A and Cryptochrome 2 that are responsive to far-red and blue light, respectively, promote the stability of CO at the end of the day by repression of degradation. In contrast, Phytochrome B reduces the abundance of CO in red light, specifically in the morning. Recent models of the circadian clock of A. thaliana include at least three interlocked negative feedback loops [96] [97] [98]. One output of the circadian clock controls the transcription of CO through several genes including *GI*, *FKF1*, and *CYCLING DOF FACTOR1* (*CDF1*). CDF1 exhibits its peak of expression early in the day, and acts as a repressor of *CO* transcription [84].

1.7.3 SDP model: rice

The progress in understanding photoperiodic induction of flowering in the LDP Arabidopsis has inspired similar studies also in SDPs. Most of this work has so far focused on the monocotyledonous plant rice. It seems as if the function of the proposed florigen FT might be highly conserved in a wide range of plant species, but that the regulation of its expression

determines the various responses to photoperiod evident in LDPs, SDPs, and DNPs. Very early flowering is seen when *FT* homologues are overexpressed in rice [86] and Morning Glory (*Ipomoea nil*) [99], as well as in tobacco, tomato [40], and poplar [60] [100]. In rice, the induction of *HD3A* (an *FT* homologue) is mediated through a *CO* homologue *HD1*, but a different function of the rice *HD1* homologue explains the short day response. In contrast to CO, HD1 represses *HD3A* in long days, but promotes *HD3A* expression and subsequent flowering in short days [101]. The reason for this divergent function of HD1 and CO is still unclear. The expression pattern of *HD1* shows a similar diurnal pattern to *CO*, and HD1 function is probably controlled by phytochrome, as a mutant lacking functional phytochromes (*photoperiod sensitivity 5*, *se5*) is early flowering under all photoperiods, although it retains the diurnal expression pattern of *HD1* [102]. These observations suggest that light activates HD1 to function as a repressor of *HD3A*, while in darkness HD1 instead acts as an activator of *HD3A*.

The similar diurnal expression patterns of *CO* and *HD1* suggest a similar regulation of their expression by a circadian clock. This hypothesis is supported by the presence of rice homologues to several genes in the Arabidopsis clock, such as *GI*, *CCA1/LHY*, *TOC1*, *ZTL*, and *ELF3* [103]. The rice *GI* homologue *OsGI* has also been shown to regulate *HD1* expression similarly to how *GI* regulates *CO* in Arabidopsis [94].

References

1.   Kriemhild, C.O. and F.K. Kenneth, *The Cambridge history of food.* Cambridge university press, 2000: p. 1654-1656.

2.   Cutler, K.D., *Tantalizing Tomatoes: Smart Tips and Tasty Picks for Gardeners Everywhere - From Wolf Peach to Outer Space.* Gardening Series, ed. B.B.G. 21st-Century. 1998.

3.   Peralta, I.E., S.K. Knapp, and D.M. Spooner, *New species of wild tomatoes (Solanum section Lycopersicon: Solanaceae) from Northern Peru.* Syst. Bot., 2005. **30**: p. 424-434.

4.   De Candolle, A., *Origin of cultivated plants.* Hafner Publishing Company, New York, 1886: p. 468 (reprint 1959).

5.   Jenkins, J.A., *The origin of the cultivated tomato.* Econ. Bot., 1948. **2**: p. 379-392.

6.   Rick, C.M. and J.F. Fobes, *Allozyme variation in the cultivated tomato and closely related species.* Bull Torrey Bot Club, 1975. **102**: p. 376-384.

7.   Rick, C.M., *Tomato: Lycopersicon esculentum (Solanaceae)*, in *Evolution of crop plants*, J.a.S. Smartt, N.W., Editor. 1995. p. 452-457.

8.   Hancock, J.F., *Plant evolution and the origin of crop species*, ed. C. Publ. 2004, Wallingford, Oxon, UK. 313.

9.   Gould, W.A., *Tomato production, processing and quality*. AVI Publ. 1983, Westport, CT, USA.

10.   Rick, C.M. and M. Holle, *Andean Lycopersicon esculentum var. cerasiforme - genetic variation and its evolutionary significance.* Econ Bot, 1990. **44**: p. 69-78.

11.   USDA, U.S.D.o.A.-. 2006.

12.   Chassy, A.W., et al., *Three-year comparison of the content of antioxidant microconstituents and several quality characteristics in organic and conventionally managed tomatoes and bell peppers.* J Agric Food Chem, 2006. **54**(21): p. 8244-52.

13.   Fraser, P.D. and P.M. Bramley, *The biosynthesis and nutritional uses of carotenoids.* Prog

Lipid Res, 2004. **43**(3): p. 228-65.

14.   Liu, Y., et al., *Manipulation of light signal transduction as a means of modifying fruit nutritional quality in tomato.* PNAS, 2004. **101**(26): p. 9897-9902.

15.   PhilippineHerbalMedicine, *Tomato Nutrition & Tomato Soup Nutrition Fact*, P.H. Medicine, Editor. 2005-2010.

16.   FAOSTAT, *FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS.* 2008.

17.   Rick, C.M. and J.I. Yoder, *Classical and molecular genetics of tomato: highlights and perspectives.* Annu Rev Genet, 1988. **22**: p. 281-300.

18.   Emmanuel, E. and A.A. Levy, *Tomato mutants as tools for functional genomics.* Curr. Op. Plant Biol., 2002. **5**: p. 112-117.

19.   Van der Hoeven, R., et al., *Deductions about the number, orga- nization and evolution of genes in the tomato genome based on analysis of a large expressed sequence tag collection and selective genomic sequencing.* Plant Cell, 2002. **14**: p. 1441-1456.

20.   Giovannoni, J.J., *Fruit ripening mutants yield insights into ripening control.* Curr. Op. Plant Biol., 2007. **10**: p. 283-289.

21.   Arumuganathan, K. and E. Earle, *Estimation of nuclear DNA content of plants by flow cytometry.* Plant Mol Biol Rep, 1991. **9**: p. 208-218.

22.   Wikstrom, N., V. Savolainen, and M.W. Chase, *Evolution of the angiosperms: calibrating the family tree.* Proc Roy Soc Lond Ser B Biol Sci, 2001. **268**: p. 2211-2220.

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

24.   Taylor, I.B., *Biosystematics of the tomato*, in *The tomato crop: a scientific basis for improvement*, J.G.a.R. Atherton, J., Editor. 1986, Chapman and Hall: London. p. 1-34.

25.   Peralta, I.E. and D.M. Spooner, *Granule-bound starch synthase (GBSSI) gene phylogeny of*

wild tomatoes (Solanum L. section Lycopersicon [Mill.] Wettst. subsection Lycopersicon). Am J Bot, 2001. **88**: p. 1888-1902.

26. Labate, J.A.e.a., *Tomato*, in *Genome Mapping and Molecular Breeding in Plants*, C. Kole, Editor. 2007, Springer-Verlag: Berlin Heidelberg.

27. Muller, C.H., *A revision of the genus Lycopersicon.* USDA Mis Publ, 1940a. **382**: p. 1-28.

28. Tanksley, S.D. and F. Loaiza-Figueroa, *Gametophytic self-incompatibility is controlled by a single major locus on chromosome 1 in Lycopersicon peruvianum.* Proc Natl Acad Sci U S A, 1985. **82**(15): p. 5093-5096.

29. van der Knaap, E. and S.D. Tanksley, *The making of a bell pepper-shaped tomato fruit: identification of loci controlling fruit morphology in Yellow Stuffer tomato.* Theoretical and applied genetics, 2003. **107**: p. 139-147.

30. Peralta, I.E., D.M. Spooner, and S.K. Knapp, *Taxonomy of Wild Tomatoes and their Relatives.* Systematic Botany Monographs - Solanum (Solanaceae). Vol. 84. 2008.

31. Levy, Y. and C. Dean, *The transition to flowering.* Plant Cell, 1998. **10**: p. 1973-1989.

32. Martinez-Zapater, J., et al., *The transition to flowering in Arabidopsis*, a.C.S. E. Meyerowitz, Editor. 1994, Cold Spring Harbor Lab Press: Cold Spring Harbor, NY. p. 403-433.

33. Haughn, G., E. Schultz, and T. Martinez-Zapater, *The regulation of flowering in Arabidopsis thaliana: meristems, morphogenesis, and mutants.* Canadian J Botany, 1995. **73**: p. 959-981.

34. Blazquez, M., *Flower development pathways.* J Cell Science, 2000. **113**: p. 3547-3548.

35. Welch, S.M., Z. Dong, and J.L. Roe, *Modelling gene networks controlling transition to flowering in Arabidopsis.* "New directions for a diverse planet". Proceedings of the 4th International Crop Science Congress, 2004. **Published on CDROM**.

36. Rost, T.L., *TOMATO - An Anatomy Atlas.* 1996.

37. Kinet, J.M. and M.M. Peet, *Tomato*, in *The Physiology of Vegetable Crops*, H.C. Wien,

Editor. 1997, CAB International: Wallingford. p. 207-258.

38.   Molinero-Rosales, N., et al., *FALSIFLORA, the tomato orthologue of FLORICAULA and LEAFY, controls flowering time and floral meristem identity.* Plant J, 1999. **20**(6): p. 685-93.

39.   Molinero-Rosales, N., et al., *SINGLE FLOWER TRUSS regulates the transition and maintenance of flowering in tomato.* Planta, 2004. **218**(3): p. 427-34.

40.   Lifschitz, E. and Y. Eshed, *Universal florigenic signals triggered by FT homologues regulate growth and flowering cycles in perennial day-neutral tomato.* J Exp Bot, 2006. **57**(13): p. 3405-14.

41.   Pnueli, L., et al., *The SELF-PRUNING gene of tomato regulates vegetative to reproductive switching of sympodial meristems and is the ortholog of CEN and TFL1.* Development, 1998. **125**(11): p. 1979-89.

42.   Lifschitz, E., et al., *The tomato FT ortholog triggers systemic signals that regulate growth and flowering and substitute for diverse environmental stimuli.* Proc Natl Acad Sci U S A, 2006. **103**(16): p. 6398-403.

43.   Samach, A. and H. Lotan, *The transition to flowering in tomato.* Plant Biotechnology, 2007. **24**: p. 71-82.

44.   Dielen, V., et al., *UNIFLORA, a pivotal gene that regulates floral transition and meristem identity in tomato (Lycopersicon esculentum).* New Phytologist, 2004. **161**: p. 393-400.

45.   Dielen, V., D. Marc, and J.M. Kinet, *Flowering in the uniflora mutant of tomato (Lycopersicon esculentum Mill.): description of the reproductive structure and manipulation of flowering time.* Plant Growth Regulation, 1998. **25**: p. 149-157.

46.   Quinet, M., et al., *Characterization of tomato (Solanum lycopersicum L.) mutants affected in their flowering time and in the morphogenesis of their reproductive structure.* J Exp Bot, 2006b. **57**(6): p. 1381-90.

47.   Quinet, M., et al., *Genetic interactions in the control of flowering time and reproductive structure development in tomato (Solanum lycopersicum).* New Phytol, 2006a. **170**(4): p.

701-10.

48.   Giliberto, L., et al., *Manipulation of the blue light photoreceptor cryptochrome 2 in tomato affects vegetative development, flowering time, and fruit antioxidant content.* Plant Physiol, 2005. **137**(1): p. 199-208.

49.   Jimenez-Gomez, J.M., et al., *Quantitative genetic analysis of flowering time in tomato.* Genome, 2007. **50**(3): p. 303-15.

50.   Koornneef, M.K., et al., *The isolation and characterization of gibberellin-deficient mutants in tomato.* Theor. Appl. Genet, 1990. **80**: p. 852-857.

51.   Menda, N., et al., *In silico screening of a saturated mutation library of tomato.* Plant J, 2004. **38**(5): p. 861-72.

52.   Garner, W.W. and H.A. Allard, *Effect of the relative length of day and night and other factors of the environment of growth and reproduction in plants.* Journal of Agricultural Research, 1920. **18**: p. 553-603.

53.   Thomas, B. and D. Vince-Prue, *Photoperiodism in plants.* 1997, Academic Press: London.

54.   Wareing, P.F., *Photoperiodism in woody plants.* Annual Review of Plant Physiology, 1956. **7**: p. 191-214.

55.   Dormling, I., *Photoperiodic control of growth and growth cessation in Norway spruce seedlings*, in *Dormancy in trees*, U.D. Symposium, Editor. 1973: Kornik.

56.   Vince-Prue, D., *Photoperiodism in plants.* 1975: London: McGraw-Hill.

57.   Eriksson, G., et al., *Inheritance of bud-set and bud-flushing in Picea abies (L.) Karst.* Theoretical and Applied Genetics, 1978. **52**: p. 2-19.

58.   Borthwick, H.A. and M.W. Parker, *Photoperiodic responses of several varieties of soybean.* Bot Gaz, 1939. **101**: p. 341-365.

59.   Ray, P.M. and W.E. Alexander, *Photoperiodic adaptation to latitude in Xanthium strumarium.* Am J Bot, 1966. **53**: p. 806-816.

60.   Bohlenius, H., et al., *CO/FT regulatory module controls timing of flowering and seasonal*

growth cessation in trees. Science, 2006. **312**(5776): p. 1040-3.

61. Tournois, J., *Infuence de la lumiŽre sur la florasion du houblon japonais et du chanvre dŽterminŽ es par des semis haitifs.*, in *Comptes Rendus de lÕAcademie des Sciences*. 1912: Paris. p. 297-300.

62. Klebs, G., *Uber das Verhaltnis der aussenwelt zur entwicklung der pflanze*, in *Heidelberger Academie der Wissenschaften*. 1913. p. 1-47.

63. Garner, W.W. and H.A. Allard, *Further studies on photoperiodism, the response of plants to relative length of day and night.* Journal of Agricultural Research, 1923. **23**: p. 871-920.

64. Knott, J.E., *Effect of a localized photoperiod on spinach.* Proceedings of the American Society of Horticultural Science, 1934. **31**: p. 152-154.

65. Lang, A., *Physiology of flower initiation*. Encyclopedia of plant physiology, ed. W. Ruhland. 1965, Berlin: Springer-Verlag. 1380-1536.

66. Chailakhyan, M.K., *GormonalÕnaya Teoriya Razvitiya Rastenii*. 1937, Akad Nauk: Moscow, Russia.

67. Lagercrantz, U., *At the end of the day: a common molecular mechanism for photoperiod responses in plants?* J Exp Bot, 2009. **60**(9): p. 2501-15.

68. Hamner, K.C. and J. Bonner, *Photoperiodism in relation to hormones as factors in floral initiation and development.* Botanical Gazette, 1938. **100**: p. 388-431.

69. Thomas, B., *Photoperiodism: an overview. In:, eds..:*, in *Biological rhythms and photoperiodism in plants*, P.J. Lumsden, Millar, A. J., Editor. 1988, BIOS Scientific Publishers: Oxford. p. 151-165.

70. Kobayashi, Y. and D. Weigel, *Move on up, itÕs time for change: mobile signals controlling photoperiod-dependent flowering*. Genes Dev, 2007. **21**: p. 2371-2384.

71. Turck, F., F. Fornara, and G. Coupland, *Regulation and identity of florigen: FLOWERING LOCUS T moves center stage.* Annu Rev Plant Biol, 2008. **59**: p. 573-94.

72. Alonso-Blanco, C., et al., *Analysis of natural allelic variation at flowering time loci in the*

Landsberg erecta and Cape Verde Islands ecotypes of Arabidopsis thaliana. Genetics, 1998. **149**(2): p. 749-64.

73. Lempe, J., et al., *Diversity of flowering responses in wild Arabidopsis thaliana strains.* PLoS Genet, 2005. **1**(1): p. 109-18.

74. Werner, J.D., et al., *Quantitative trait locus mapping and DNA array hybridization identify an FLM deletion as a cause for natural flowering-time variation.* Proc Natl Acad Sci U S A, 2005b. **102**(7): p. 2460-5.

75. Redei, G.P., *Supervital mutants of Arabidopsis.* Genetics, 1962. **47**: p. 443-460.

76. Koornneef, M., C.J. Hanhart, and J.H. van der Veen, *A genetic and physiological analysis of late flowering mutants in Arabidopsis thaliana.* Mol Gen Genet, 1991. **229**(1): p. 57-66.

77. Putterill, J., et al., *The CONSTANS gene of Arabidopsis promotes flowering and encodes a protein showing similarities to zinc finger transcription factors.* Cell, 1995. **80**(6): p. 847-57.

78. Fowler, S., et al., *GIGANTEA: a circadian clock-controlled gene that regulates photoperiodic flowering in Arabidopsis and encodes a protein with several possible membrane-spanning domains.* Embo J, 1999. **18**(17): p. 4679-88.

79. Kobayashi, Y., et al., *A pair of related genes with antagonistic roles in mediating flowering signals.* Science, 1999. **286**(5446): p. 1960-2.

80. Kardailsky, I., et al., *Activation tagging of the floral inducer FT.* Science, 1999. **286**(5446): p. 1962-5.

81. Park, D.H., et al., *Control of circadian rhythms and photoperiodic flowering by the Arabidopsis GIGANTEA gene.* Science, 1999. **285**(5433): p. 1579-82.

82. Suarez-Lopez, P., et al., *CONSTANS mediates between the circadian clock and the control of flowering in Arabidopsis.* Nature, 2001. **410**(6832): p. 1116-20.

83. Valverde, F., et al., *Photoreceptor regulation of CONSTANS protein in photoperiodic flowering.* Science, 2004. **303**(5660): p. 1003-6.

84.   Imaizumi, T., et al., *FKF1 F-box protein mediates cyclic degradation of a repressor of CONSTANS in Arabidopsis.* Science, 2005. **309**(5732): p. 293-7.

85.   Izawa, T., Y. Takahashi, and M. Yano, *Comparative biology comes into bloom: genomic and genetic comparison of flowering pathways in rice and Arabidopsis.* Curr Opin Plant Biol, 2003. **6**(2): p. 113-20.

86.   Kojima, S., et al., *Hd3a, a rice ortholog of the Arabidopsis FT gene, promotes transition to flowering downstream of Hd1 under short-day conditions.* Plant Cell Physiol, 2002. **43**(10): p. 1096-105.

87.   Chardon, F. and C. Damerval, *Phylogenomic analysis of the PEBP gene family in cereals.* J Mol Evol, 2005. **61**(5): p. 579-90.

88.   Komiya, R., et al., *Hd3a and RFT1 are essential for flowering in rice.* Development, 2008. **135**(4): p. 767-74.

89.   Doi, K., et al., *Ehd1, a B-type response regulator in rice, confers short-day promotion of flowering and controls FT-like gene expression independently of Hd1.* Genes Dev, 2004. **18**(8): p. 926-36.

90.   Kim, S.L., et al., *OsMADS51 is a short-day flowering promoter that functions upstream of Ehd1, OsMADS14, and Hd3a.* Plant Physiol, 2007. **145**(4): p. 1484-94.

91.   Matsubara, K., et al., *Ehd2, a rice ortholog of the maize INDETERMINATE1 gene, promotes flowering by up-regulating Ehd1.* Plant Physiol, 2008. **148**(3): p. 1425-35.

92.   Tsuji, H., et al., *Florigen and the photoperiodic control of flowering in rice.* Rice, 2008. **1**: p. 25-35.

93.   Hayama, R. and G. Coupland, *The molecular basis of diversity in the photoperiodic flowering responses of Arabidopsis and rice.* Plant Physiol, 2004. **135**(2): p. 677-84.

94.   Hayama, R., et al., *Adaptation of photoperiodic control pathways produces short-day flowering in rice.* Nature, 2003. **422**(6933): p. 719-22.

95.   Putterill, J., R. Laurie, and R. Macknight, *It's time to flower: the genetic control of flowering*

time. Bioessays, 2004. **26**(4): p. 363-73.

96.   McClung, C.R., *Plant circadian rhythms.* Plant Cell, 2006. **18**(4): p. 792-803.

97.   Locke, J.C., et al., *Experimental validation of a predicted feedback loop in the multi-oscillator clock of Arabidopsis thaliana.* Mol Syst Biol, 2006. **2**: p. 59.

98.   Zeilinger, M.N., et al., *A novel computational model of the circadian clock in Arabidopsis that incorporates PRR7 and PRR9.* Mol Syst Biol, 2006. **2**: p. 58.

99.   Hayama, R., et al., *A circadian rhythm set by dusk determines the expression of FT homologs and the short-day photoperiodic flowering response in Pharbitis.* Plant Cell, 2007. **19**(10): p. 2988-3000.

100.   Hsu, C.Y., et al., *Poplar FT2 shortens the juvenile phase and promotes seasonal flowering.* Plant Cell, 2006. **18**(8): p. 1846-61.

101.   Yano, M., et al., *Hd1, a major photoperiod sensitivity quantitative trait locus in rice, is closely related to the Arabidopsis flowering time gene CONSTANS.* Plant Cell, 2000. **12**(12): p. 2473-2484.

102.   Izawa, T., et al., *Phytochrome mediates the external light signal to repress FT orthologs in photoperiodic flowering of rice.* Genes Dev, 2002. **16**(15): p. 2006-20.

103.   Murakami, M., et al., *Comparative overviews of clock-associated genes of Arabidopsis thaliana and Oryza sativa.* Plant Cell Physiol, 2007. **48**(1): p. 110-21.

CHAPTER 2

THE TOMATO GENOME: MAP POSITION AND SHARED MICROSYNTENY OF GENES INVOLVED IN FLOWERING

2.1 Introduction

    2.1.1 Sequencing of the tomato genome

    2.1.2 IL mapping in tomato

2.2 Aim of the work

2.3 Materials and methods

2.4 Results and discussion

2.1 Introduction

2.1.1 Sequencing of the tomato genome

The haploid genome of tomato comprises approximately 950 Mb of DNA, divided in 12 chromosomes. More than 75% of the genome is constituted by heterochromatin and largely devoid of genes [1]. About 90% of genes are found in long contiguous stretches of gene-dense euchromatin located on the distal portions of each chromosome arm [2]. The heterochromatic fraction consists mostly of repetitive sequences. The genome is being sequenced by an international consortium of 10 countries (including Italy) as part of a larger initiative called the "International Solanaceae Genome Project (SOL): Systems Approach to Diversity and Adaptation", launched in November 2003.

Due to the high level of macro and micro-synteny in the Solanaceae, the first aim of the SOL project was to use an ordered BAC approach to generate a high quality sequence for the euchromatic portions of the tomato as a reference for the shotgun sequencing of other Solanaceae. The starting point chosen for sequencing the tomato genome were BACs anchored to the genetic map by overgo hybridization and AFLP technology. The overgos derived from approximately 1500 markers from the tomato high density genetic map (http://solgenomics.net/index.pl), based on a single, common *S. lycopersicum* x *S. pennellii* F2 population (referred to as the F2-2000). These seed BACs were used as anchors from which to radiate the tiling path using BAC end sequence data available from an end sequence database, and to produce a fingerprint contig physical map. Validation of the physical map was carried out using fluorescent in-situ hybridization (FISH) on pachytene complements with entire BAC clones as probes [3] [4]. This sequencing approach was chosen because it provides the highest possible sequence quality. However, the main drawback of the BAC by BAC approach is that it is remarkably more expensive and slower than the a whole genome shotgun sequencing (WGS) approach, since novel "next generation" sequencing technologies have become applicable for the sequencing of complex genomes.

The BAC-by-BAC sequencing approach revealed itself to be slower and more challenging than expected. In 2008, after five years, several chromosomes had large gaps and the efforts to identify novel seed BACs within several of these gaps remained unsuccessful. Moreover, the gene density in heterochromatin was found to be higher than expected. For this reason, at the end of 2008 a "Next Generation Sequencing Initiative" was launched by 4 countries, including Italy, in order to produce a whole genome shotgun draft sequence of the tomato genome, anchored to a novel physical map, as a complement to the ongoing BAC-by-BAC sequencing effort of the euchromatin. The aim of the project was to obtain a Whole Genome Sequence based on:

1: ~20x genome coverage in 454 shotgun and paired-end reads;

2: ~30x genome coverage in SOLID paired-end reads;

3: ~3x genome coverage in Sanger paired-end reads (plasmid, fosmid and BAC ends);

4: A whole genome physical map based on 10X Genome Analyzer (Illumina).

In little more than one year, a first assembly version of the tomato genome was released, with more than 90 Gb of data. Moreover, iTAG (International Tomato Annotation Group), a multinational consortium funded in part by the EU-SOL project, released a comprehensive and high quality annotation. The tomato genome shotgun sequence and annotation are available on http://solgenomics.net/index.pl and are constantly updated.

2.1.2 IL mapping in tomato

An introgression line (IL) library, consists of a set of lines, each of which carries a single, defined chromosome segment (introgressed segment) that originates from a donor species, in the genetic background of a closely related species. Chromosome segments are traced through crosses by genotyping the lines with a genome-wide panel of polymorphic markers that can distinguish between parental alleles (Fig 1). The production of such a library takes about ten generations. These lines provide an efficient tool for detecting and mapping valuable agronomic traits [5].

In tomato, RFLP (restriction fragment length polymorphism) markers have been used to develop a full-coverage IL library from a cross between the wild green-fruited species *S. pennellii* and the cultivated tomato *S. lycopersicum* cv. M82 [6] (Fig. 2). This IL library is successfully used for fine gene mapping in tomato (http://solgenomics.net/index.pl).

a   b
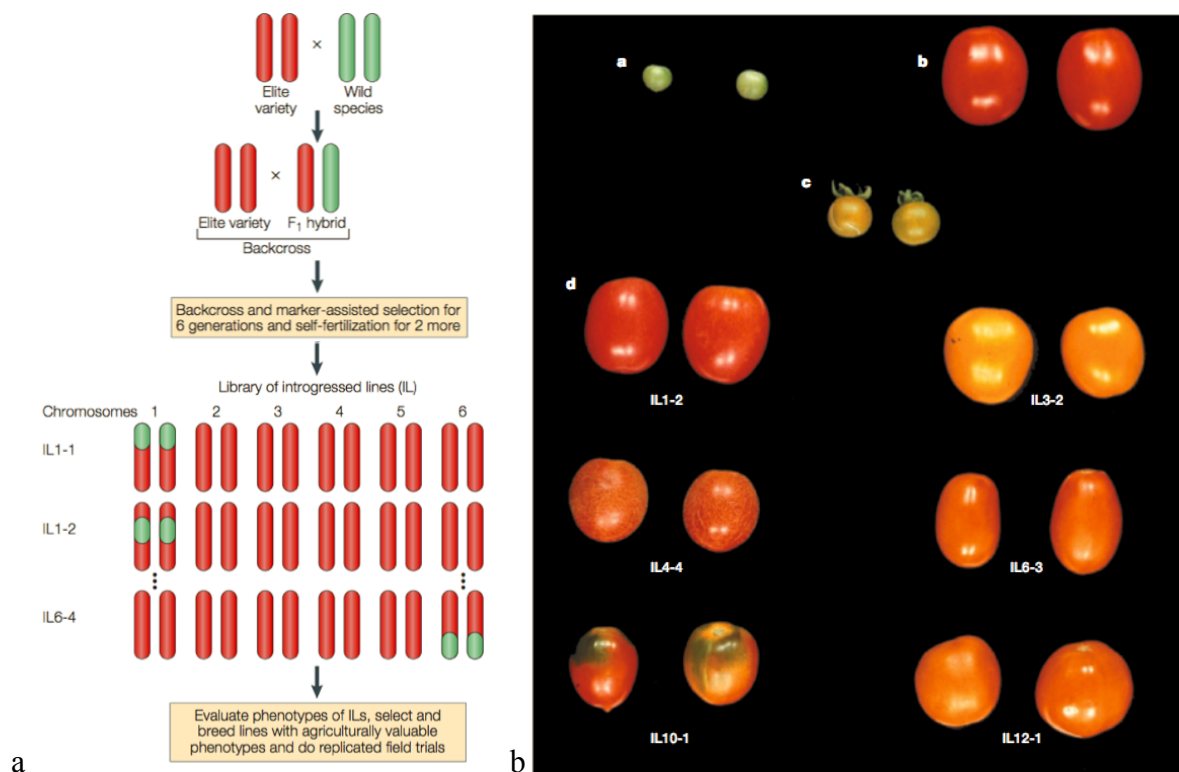
Figure 1a The wild species (green) is crossed as a male parent to an elite cultivated variety (red), and the F1 hybrid is backcrossed to the elite parent. Through recurrent backcrosses, the average proportion of the wild species genome is reduced by 50% in each generation. By the sixth backcross generation, independent plants can be isolated, each heterozygous for a different segment of the wild-species genome. Further self-pollination and selection leads to homozygosity at targeted introgressed segments [5]
Figure 1b The *S. pennellii* library. **a**: Green fruits of the wild species *S. pennellii*. **b**: the lycopene-rich red fruits of *S. lycopersicum*, c | their F1 hybrid progeny, and d | six introgression lines (ILs) that show their different fruit colour and carotenoid-content phenotypes. Each line contains a single, marker-defined, wild-species-derived chromosome segment that harbours a gene that affects fruit phenotype. [5]

2.2 Aim of the work

Having contributed to the genomic sequencing, our first goal has been to produce structural genomic data on candidate genes controlling flowering in tomato. More specifically, a search for unknown paralogs and a study on microsynteny with other species was undertaken. Shared synteny can reflect merely co-inheritance, or selection for functional relationships between syntenic genes,

such as combinations of alleles that are advantageous when inherited together, or shared regulatory mechanisms. The genes selected for this thesis project are the *TCOL* (tomato *CONSTANS*-like), the *CRYPTOCHROME* and the *GIGANTEA* gene families. Their role in flowering induction is described in detail in chapters 1 and 3. While members of the *TCOL* and *CRY* families have been previously published in tomato [9] [10], the *GIGANTEA* gene(s) still await detailed characterization.

2.3 Materials and methods

2.3.1 Gene mapping via IL

Primer pairs were designed to amplify, via PCR, about 800-1000 bp of the gene of interest, of both the IL library parentals: *S. lycopersicum* and *S. pennellii.* The "Oligo – Primer" analysis software was used from primer design. Amplification results were verified via agarose gel electrophoresis. PCR products were sequenced with an ABI 3730 DNA Analyser. Sequences of *S. lycopersicum* and *S. pennellii* were aligned with the "CodonCode Aligner" software in order to detect SNPs or indels. One primer was designed to match the SNP/indel with its 3' terminus in order to discriminate, among the chromosome-specific DNA pools, the one that contains the *S. pennellii* sequence. The other primer was distant at least 70 bases (fragments that are too short are not detectable on agarose gels). Mutations that destabilize the double helix (like transversions or small indels) are preferable, but the method works well also with transitions on both primers. Parental DNAs (*S. pennellii* and *S. lycopersicum*) and chromosome-specific pools are the substrates for the PCR with the *S. pennellii*-specific primers. Successful reactions are characterized by only two positive samples: *S. pennellii* and one of the IL chromosome pools. The procedure was repeated with single DNA sample of the positive pool in order to identify chromosomal region in which the gene maps.

2.3.2 Paralog identification

The "SGN Blast" tool (http://solgenomics.net/tools/blast/) allows searching for nucleotide and amino acid sequences within the tomato genome assembly. Nucleotide sequences of candidate genes were submitted using the "BLASTN" algorithm, while the "TBLASTX" algorithm was used to submit amino acid sequences. The sequence database used was the "Tomato Whole Genome prerelease" (http://solgenomics.net/tomato/). For each submitted sequence, the tool identified the scaffold harbouring the sequences that produce significant alignments. A scaffold is a series of contigs that are in the right order and orientation, but are separated by gaps whose size is known. Sequences with significant score value (estimation based on sequence similarity) and e-value (expectation value, a measure of the reliability of the score value) were submitted to the NCBI database (http://www.ncbi.nlm.nih.gov/) in order to verify if the identified putative paralogs are already known.

2.3.3 Shared Microsynteny

Candidate genes can be associated to specific scaffolds as described above. The code of the scaffold and the position of the gene in the scaffold were used to visualize flanking annotated genes in the "SGN Genome Browser" (http://solgenomics.net/gbrowse/). A region of 100 Kb surrounding the gene was analysed and annotated flanking genes were listed.

Microsynteny with *Arabidopsis thaliana* was verified, comparing the order of genes flanking the Arabidopsis ortholog. The Arabidopsis genome browser is available on AtEnsembl website (http://atensembl.arabidopsis.info/index.html). The Blast tool allows to localize the submitted sequence and to display flanking genes. Like for tomato, a region of 100 Kb surrounding the gene was analysed.

2.4 Results and discussion

Many of the candidate genes involved in the tomato flowering pathway are members of multigene families (e.g *PHY* [8], *CRY* [9], *COL* [10], *SP* [11]). We verified, with the newest data available for the tomato genome, that no additional Cryptochrome genes are present in tomato, other than *CRY1a, CRY1b, CRY2* and *CRY3*.

TCOL1, TCOL2 and TCOL3 are known members of the *TCOL* family in tomato [10]. Our sequencing data demonstrate, in agreement with the newest genome sequence data, and unlike previously published [10], that *TCOL2* is a pseudogene containing a frameshift mutation that produces a truncated polypeptide (143 amino acids instead of 341). In total, 6 putative *TCOL* genes and 1 *TCOL* pseudogene are present in tomato. compared to 17 *COL* genes in Arabidopsis [12], and 16 in rice [13].

Gene	Scaffold (V1.0)	Position	Gene code	Peptide length (aa)	N° of exons	chromosome
putative *TCOL*	SL1.00sc02102	1751102..1752269	281.1	358	2	12
putative *TCOL*	SL1.00sc02156	63642..65248	8.1	359	2	8
putative *TCOL*	SL1.00sc07059	3310214..3311904	287.1	386	2	7
putative *TCOL*	SL1.00sc07266	951557..953339	80.1	373	4	9
TCOL1	SL1.00sc01656	2204785..2207308	269.1	391	3	2
TCOL2	SL1.00sc01656	2220749..2222194	272.1	143*	2*	2
TCOL3	SL1.00sc01656	2212025..2216135	271.1	409	3	2

Table 1. Putative *TCOL* genes identified with the scaffold code and with the putative gene code produce by iTAG annotation. The chromosome position was provided by Shusei Sato at Kazusa (Japan). *: truncated polipeptide.

We performed a phylogenetic analysis comparing the 6 tomato TCOL proteins with the 17 Arabidopsis COL proteins. In Arabidopsis, the *COL* gene family is subdivided into three groups [12]. The first comprises *CO* and *COL1* to *COL5* (two B-box genes), the second comprises *COL6* to *COL8* and *COL16* (one B-box genes), and the third comprises *COL9* to *COL15* (one CO-like B-box and one more diverged zinc finger domain). The phylogenetic distances were calculated using "MEGA4" software. A dendrogram is showed in figure 3. The results confirmed what has been previously observed: TCOL1 and TCOL3 belong to group I [10]. According to this analysis, five of the active tomato *TCOL* genes, including *TCOL1-3*, belong to group I and one to group III.
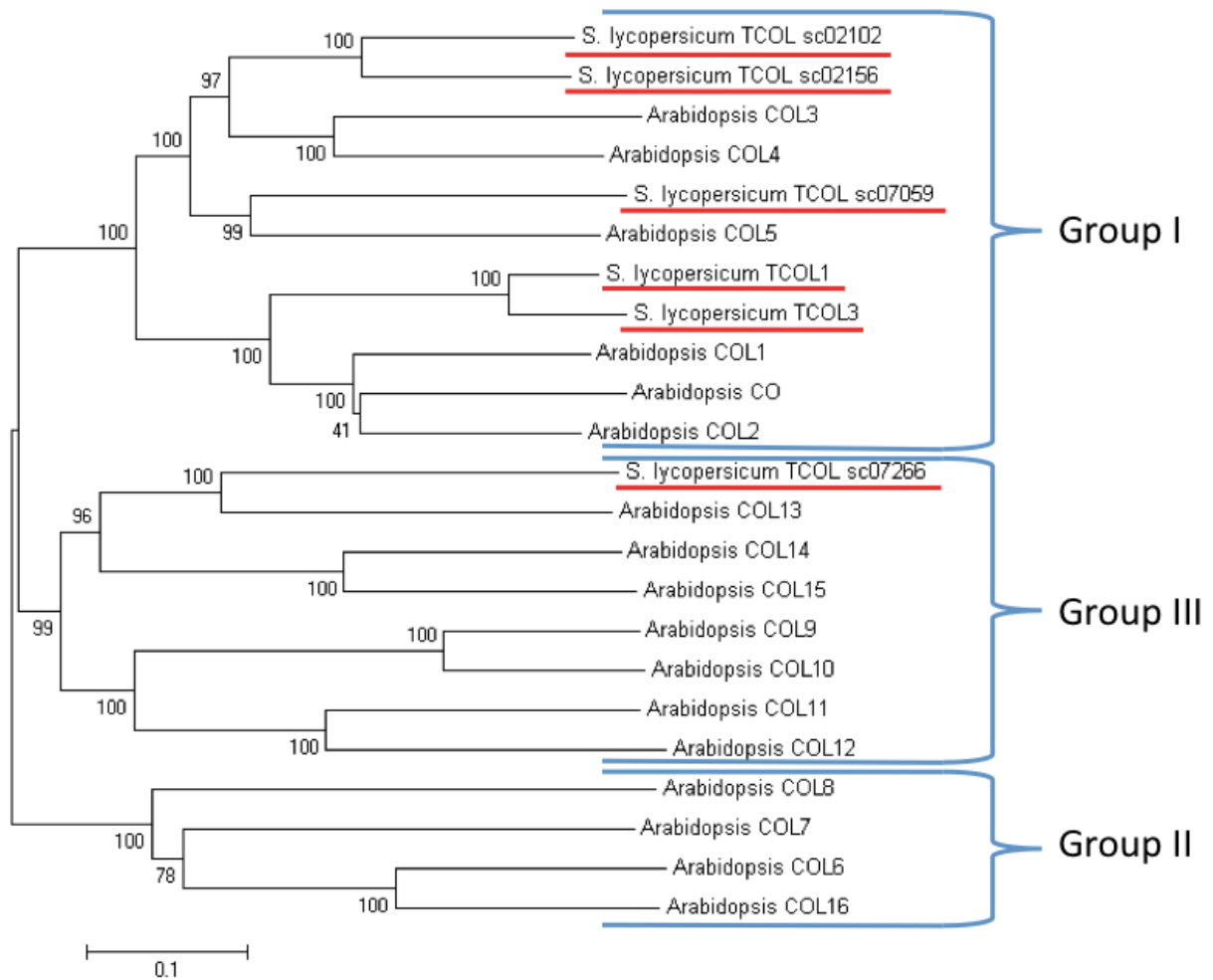
Figure 2. Dendrogram of Arabidopsis and tomato COL proteins. The dendrogram respect the classification made by Robson et al. (2001) [12]. The evolutionary history was inferred using the Neighbor-Joining method [14]. The optimal tree with the sum of branch length = 5.74460019 is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates) are shown next to the branches [15]. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. All positions containing alignment gaps and missing data were eliminated only in pairwise sequence comparisons (Pairwise deletion option). There were a total of 563 positions in the final dataset. Phylogenetic analyses were conducted in MEGA4 [16].

We were also able to identify the tomato *GIGANTEA* (*TGI*) gene before the release of the genome sequence and annotation. The Arabidopsis *GIGANTEA* mRNA sequence was utilized to identify a homologous unigene in the tomato SGN database, used to design primers for direct amplification from tomato cDNA. This led to the identification of two distinct *TGI* paralogs, named *TGI1* and *TGI2*. The data was confirmed once the genome sequence and the annotation were released. No other *TGI* paralogs were identified. The TGI1 and TGI2 proteins are equally distant from their Arabidopsis homolog (Fig 4).
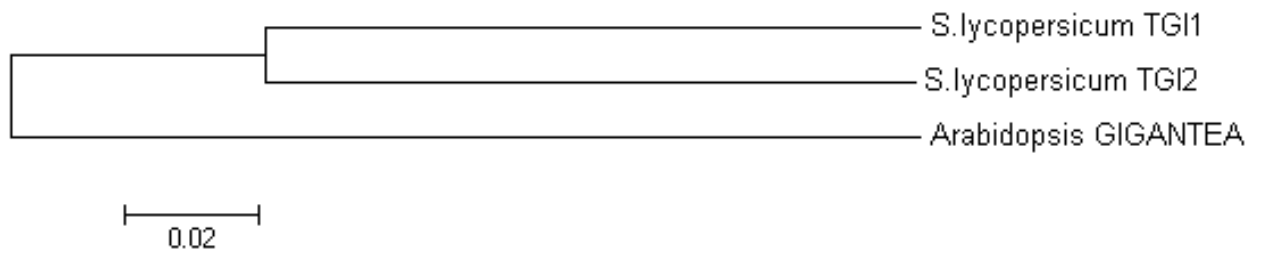
Figure 3. Dendrogram of Arabidopsis and tomato GIGANTEA proteins. The evolutionary history was inferred using the Neighbor-Joining method [14]. The optimal tree with the sum of branch length = 0.36471962 is shown. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. All positions containing alignment gaps and missing data were eliminated only in pairwise sequence comparisons (Pairwise deletion option). There were a total of 1185 positions in the final dataset. Phylogenetic analyses were conducted in MEGA4 [16].

A phylogenetic analysis was performed also for the tomato and Arabidopsis Cryptochromes (Fig 5).

The dendrogram provides evidence for a *CRY1* gene duplication in the tomato clade.
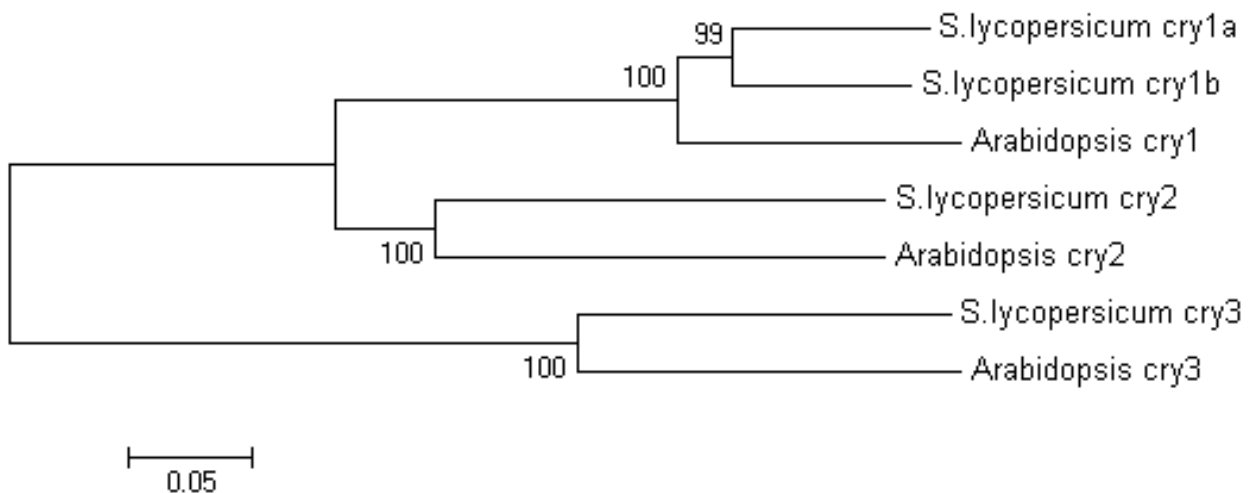
Figure 4. Dendrogram of Arabidopsis and tomato Cryptochrome proteins. The evolutionary history was inferred using the Neighbor-Joining method [14]. The optimal tree with the sum of branch length = 1.49819792 is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates) are shown next to the branches [15].. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. All positions containing alignment gaps and missing data were eliminated only in pairwise sequence comparisons (Pairwise deletion option). There were a total of 790 positions in the final dataset. Phylogenetic analyses were conducted in MEGA4 [16].

The four *CRY* genes, as well as *TCOL1, TCOL3, TGI1* and *TGI2* were all mapped using the IL

mapping protocol and these data were compared with the scaffold mapping produced by the Shusei

Sato at Kazusa (Japan). These mapping results are listed in the table 2.

Gene	ver 1.0 scaffold (iTAG data)	ver1.03 scaffold (mapping data)	Strand	Flanking marker	Position (cM)	Chromosome	IL mapping
CRY1a	scaffold03736	scaffold02256	-	TEI0139	65,553	chr04	chr04 BIN G
				T737	63,473		
CRY1b	scaffold03785	scaffold00087	-	TES1686	73,764	chr12	chr12 BIN F
				TES0743	73,03		
CRY2	scaffold01402	scaffold02760	-	TGS0560	78,873	chr09	chr09 BIN J
				TG421	81,9		
CRY3	scaffold00777	scaffold01291	-	TES0611	50,852	chr08	chr08 BIN E
TGI1	scaffold03736	scaffold00103	+	T1050	58,33	chr04	chr04 BIN G
				CT178	58,064		
TGI2	scaffold03785	scaffold00087	+	T1266	71.00	chr12	chr12 BIN F
				TES0743	73,03		
TCOL1 TCOL2 TCOL3	scaffold01656	scaffold01403	+	TGS1770	123,698	chr02	chr02 BIN J
				cLES3G11	122,997		

Table 2. Map position of candidate genes. For each gene, the table provides the code of the scaffold of the genome release V1.0 (used for the annotation) and V1.03 (used for the scaffold mapping). Genetic marker that are close to the gene were used to estimate its genetic position (cM). The genetic positions were confirmed via IL mapping.

Two interesting observations arise from the results. The first is that the three *TCOL1-3* genes map in the same genomic region, in disagreement with previous work of Ben-Naim et al (2006), which placed *TCOL2* and *TCOL3* in tandem, and *TCOL1* on a different genome region. The second is that *TGI1* and *CRY1a* map in the same bin in chromosome 4 and *TGI2* and *CRY1b* map in the same bin in chromosome 12. Data from the potato genome prerelease allow the identification of two *CRY1* and two *GIGANTEA* also in this species. The mapping data available for the potato genome are still too preliminary to evaluate synteny with tomato. However, we were able to verify that potato *CRY1b* and one of the two potato *GIGANTEA* belong to the same genomic region, and their relative distance is similar to that of their tomato orthologs (data not shown). The same analysis has been performed for Arabidopsis, which contains only one *CRY1* and one *GI*, located on different chromosomes. Sequence data from the "Rice annotation project database" (http://rapdb.dna.affrc.go.jp/) indicates that also rice contains only one *GI*. Our working hypothesis is that *CRY1* and *GI* moved together, and then the whole region was duplicated, in a common ancestor of tomato and potato.

Concerning the microsynteny shared between tomato and Arabidopsis, a region of 100 Kb containing the three *TCOL* genes was compared to a similar size that contains *CONSTANS* in the Arabidopsis genome. In this region, two *COL* genes are present in Arabidopsis, while 3 are present in tomato. They are all classified as member of the group I of the *COL* gene family, as mentioned above. Furthermore, protein sequence alignment allows the identification of three additional genes that are shared between the two species in this region (Table 3). The relative position and orientation of these three genes in Arabidopsis is inverted respect to their orthologs in tomato. Our hypothesis is that one of the *TCOL* genes was subjected to duplication and that a second inversion event generated a rearrangement between *COL* genes and the other three genes. A schematic representation is provided in figure 5.
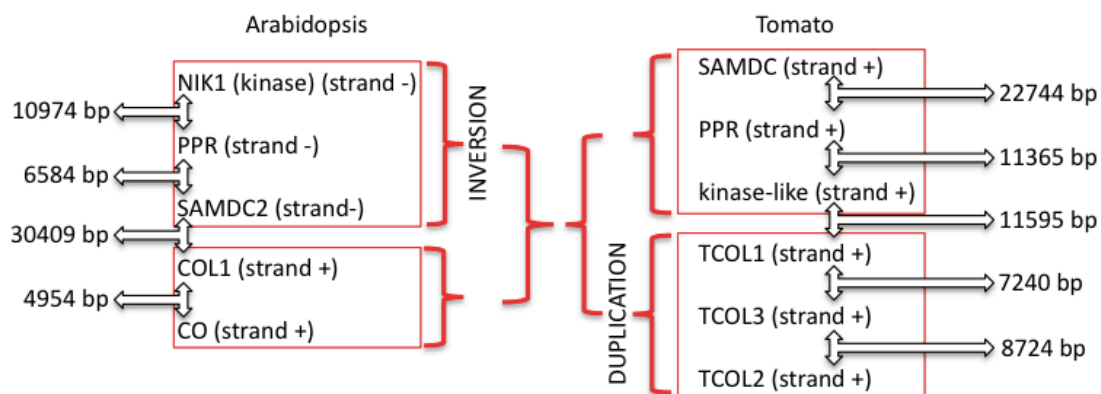

Figure 5. Schematic representation of microsynteny shared between Arabidopsis and tomato in the region harbouring *TCOL* genes belonging to the group I. Strand and distance between genes are indicated.

ARABIDOPSIS			
Chr	chromosome position	strand	protein
5	5141452 5142704	+	RNA-binding S4 domain-containing protein
	5143103 5143908	+	plastid-specific 30S ribosomal protein 3, putative
	5144096 5144545	-	GCN5-related N-acetyltransferase (GNAT) family protein
	5144901 5146300	-	pollen Ole e 1 allergen and extensin family protein
	5149803 5150967	+	zinc finger (C3HC4-type RING finger) family protein
	5151597 5153770	-	SEP1 (SEPALLATA1); DNA binding / transcription factor
	5157801 5161124	+	N2,N2-dimethylguanosine tRNA methyltransferase family protein
	5161790 5162836	+	zinc finger (C3HC4-type RING finger) family protein
	5168594 5169154	+	bZIP transcription factor family protein
	5171346 5172700	-	CO (CONSTANS)
	5176300 5177476	-	COL1 (CONSTANS-LIKE 1)
	5178679 5181266	-	ATPCME (PRENYLCYSTEINE METHYLESTERASE)
	5182644 5184881	-	glycosyl hydrolase family 81 protein
	5193210 5195205	+	dehydrogenase-related; similar to TAIR:AT5G10730.1
	5196213 5199562	+	structural maintenance of chromosomes (SMC) family protein (MSS2)
	5200332 5202253	+	PAM1 (PLANT ADHESION MOLECULE 1); RAB GTPase activator
	5202804 5204718	+	short-chain dehydrogenase/reductase (SDR) family protein
	5206709 5207797	+	S-adenosylmethionine decarboxylase proenzyme 2 (SAMDC2)
	5209995 5210476	+	KIN1; cold and ABA inducible protein
	5211969 5212444	+	KIN2 (COLD-RESPONSIVE 6.6)
	5213293 5215299	+	pentatricopeptide (PPR) repeat-containing protein
	5224267 5227006	+	NIK1 (NSP-INTERACTING KINASE 1); kinase
	5227985 5229015	+	3-oxo-5-alpha-steroid 4-dehydrogenase family protein
	5232734 5232889	-	DVL1/RTFL18 (ROTUNDIFOLIA LIKE 18)

TOMATO			
Chr	scaffold position	strand	protein
2	2159081 2162270	+	S-adenosylmethionine decarboxylase proenzyme (SAMDC)
	2163564 2168348	-	Ribosomal RNA-processing protein 7 homolog A
	2169873 2170793	+	Pollen coat-like protein
	2172733 2175706	+	Unknown Protein
	2181825 2184453	+	Pentatricopeptide (PPR) repeat-containing protein
	2184672 2184962	-	Unknown Protein
	2193190 2196416	+	Receptor protein kinase-like protein
	2204785 2207308	+	TCOL1
	2209897 2212023	-	Unknown Protein
	2212025 2216135	+	TCOL3
	2220749 2222194	+	TCOL2
	2222840 2226290	-	Unknown Protein
	2233868 2234233	+	AtGDU2 Arabidopsis thaliana GLUTAMINE DUMPER 2
	2260585 2263394	-	GAUT9 Galacturonosyltransferase 9
	2268329 2270579	+	Unknown Protein
	2270910 2272095	-	Transcription factor HBP-1a

Table 3. Detailed information about the regions flanking Group A *COL* genes in Arabidopsis and tomato. The positions in the genome (or in the scaffold, for tomato) are plotted with the respective orientation of the genes. The encoded proteins that are syntenic are marked with specific colours.

43

Preliminary expression levels of selected genes were obtained from 454 cDNA libraries from 5 different tissues (Fig. 6). The read number is actually too low to have statistical significance but some observations can be done. *TCOL2* seems to be not expressed in all tissues, expect for some reads in two berry stages, in agreement with the fact that it is a pseudogene. *TCOL1* and *TCOL3* are expressed predominantly in leaves. *TGI1* and *TGI2* present unexpected expression profiles, with lowest values in leaves, even if they are involved in circadian regulation. The *CRY1a*, *1b* and *2* genes are expressed in most tissues, especially in stem, leaf and mature green berry, while *CRY3* shows the lowest expression levels among all cryptochromes.

Gene	reads per million						
	root	stem	leaf	flower	berry		
					mature green	breaker	10 DPB
CRY1a	72.13	159.02	120.51	12.23	73.04	60.71	76.51
CRY1b	2.19	49.98	166.70	0.00	208.39	54.64	7.65
CRY2	28.41	38.62	20.08	12.23	75.19	39.46	56.10
CRY3	2.19	4.54	6.03	0.00	10.74	3.04	2.55
TGI1	61.20	29.53	10.04	27.51	30.08	33.39	30.60
TGI2	45.90	11.36	10.04	24.46	47.26	27.32	38.25
TCOL1	2.19	13.63	108.46	0.00	8.59	12.14	7.65
TCOL2	0.00	0.00	0.00	0.00	2.15	0.00	5.10
TCOL3	2.19	2.27	28.12	0.00	4.30	6.07	10.20

colour	reads
	0
	>0-10
	>10-20
	>20-50
	>50-100
	>100

Figure 6. Expression level of candidate genes normalized per million of reads. Data were obtained blasting the CDS vs the reads obtained by the 454 sequencing of the cDNA library, with a threshold identity of 95%.

References:

1.   Peterson, D.G., et al., *DNA content of heterochromatin and euchromatin in tomato (Lycopersicon esculentum) pachytene chromosomes.* Genome, 1996. **39**(1): p. 77-82.

2.   Wang, Y., et al., *Euchromatin and pericentromeric heterochromatin: comparative composition in the tomato genome.* Genetics, 2006. **172**(4): p. 2529-40.

3.   Chang, S.B., et al., *Predicting and testing physical locations of genetically mapped loci on tomato pachytene chromosome 1.* Genetics, 2007. **176**(4): p. 2131-8.

4.   Szinay, D., et al., *High-resolution chromosome mapping of BACs using multi-colour FISH and pooled-BAC FISH as a backbone for sequencing tomato chromosome 6.* Plant J, 2008. **56**(4): p. 627-37.

5.   Zamir, D., *Improving plant breeding with exotic genetic libraries.* Nat Rev Genet, 2001. **2**(12): p. 983-9.

6.   Eshed, Y. and D. Zamir, *An introgression line population of Lycopersicon pennellii in the cultivated tomato enables the identification and fine mapping of yield-associated QTL.* Genetics, 1995. **141**(3): p. 1147-62.

7.   Rick, C.M. and J.I. Yoder, *Classical and molecular genetics of tomato: highlights and perspectives.* Annu Rev Genet, 1988. **22**: p. 281-300.

8.   Alba, R., et al., *The phytochrome gene family in tomato and the rapid differential evolution of this family in angiosperms.* Mol Biol Evol, 2000. **17**(3): p. 362-73.

9.   Perrotta, G., et al., *Tomato contains homologues of Arabidopsis cryptochromes 1 and 2.* Plant Mol Biol, 2000. **42**(5): p. 765-73.

10.   Ben-Naim, O., et al., *The CCAAT binding factor can mediate interactions between CONSTANS-like proteins and DNA.* Plant J, 2006. **46**(3): p. 462-76.

11.   Carmel-Goren, L., et al., *The SELF-PRUNING gene family in tomato.* Plant Mol Biol, 2003. **52**(6): p. 1215-22.

12.   Robson, F., et al., *Functional importance of conserved domains in the flowering-time gene CONSTANS demonstrated by analysis of mutant alleles and transgenic plants.* Plant J, 2001. **28**(6): p. 619-31.

13.   Griffiths, S., et al., *The evolution of CONSTANS-like gene families in barley, rice, and Arabidopsis.* Plant Physiol, 2003. **131**(4): p. 1855-67.

14.   Saitou, N. and M. Nei, *The neighbor-joining method: a new method for reconstructing phylogenetic trees.* Mol Biol Evol, 1987. 4(4): p. 406-25.

15.   Felsenstein, J., *Confidence limits on phylogenies: An approach using the bootstrap.* Evolution, 1985. 39: p. 783-791.

16.   Tamura, K., et al., *MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0.* Mol Biol Evol, 2007. 24(8): p. 1596-9.

CHAPTER 3

**MOLECULAR AND FUNCTIONAL ANALYSIS OF FLOWERING GENES IN TOMATO
AND ITS WILD RELATIVES**

3.1 Introduction

    3.1.1 CONSTANS

    3.1.2 Cryptochromes

    3.1.3 GIGANTEA

3.2 Aim of the work

3.3 Material and methods

3.4 Results and discussion

3.1 Introduction

3.1.1 CONSTANS

<u>*CONSTANS* in Arabidopsis</u>

As previously discussed in chapter 1, flower development is initiated at the shoot apex in response to environmental cues. Photoperiod is one of the most important of these cues. Genetic analysis in Arabidopsis identified a pathway of genes responsible for the initiation of flowering that is influenced by day length. The nuclear zinc-finger protein CONSTANS (CO, or AtCO) plays a central role in this pathway, and activates the transcription of *FT* in response to long days. CO acts non-cell autonomously to trigger flowering [1]. Although *CO* mRNA is expressed at low levels in wild-type plants [2], it is able to induce early flowering and complement the *co* mutation when expressed from phloem-specific promoters, but not from meristem-specific promoters [1]. Genetic approaches indicate that CO activates flowering through both FT-dependent and FT-independent processes and egulates the synthesis or transport of a systemic flowering signal [1]. Recently, CO was proposed to be part of a complex mechanism that allows Arabidopsis to distinguish between long and short days, through a combination of circadian-clock regulation and direct responsiveness to exposure to light [3] [4] [5] [6].

In situ hybridisation allowed to detect *CO* mRNA in the Shoot Apical Meristem and in young leaf primordia [7], while classical grafting experiments suggested that the perception of photoperiod occurs in the leaf, which is consistent with CO acting in phloem cells to promote flowering. This conclusion may imply that CO function is beyond flowering-time control, as heterologous expression of *CO* in potato delayed tuberisation, and this effect was graft transmissible [8].

<u>*CONSTANS-Like* genes</u>

The Arabidopsis *CONSTANS* (*AtCO*) gene, belong to *CONSTANS-Like* (*COL*) genes group. These genes encode proteins containing two highly conserved domains: an N-terminal zinc-finger B- box

[9] domain, which is also found in animals, and a plant-specific C-terminal CCT domain [10] domain. *COL* genes have been identified in different plant species [11] and the genome of each of these plants seems to host large families of these genes. In Arabidopsis there are 17 *COL* genes [12], and at least 16 of them are present in rice [11].

The best-characterized COL proteins are AtCO [2] and rice Heading Date 1 [13]. They involve mediation of the photoperiodic induction of flowering. These proteins act by regulating transcription levels of floral integrators [14]. One such common integrator is encoded by *FLOWERING LOCUS T* (*FT*), [15] [16] in Arabidopsis and *Heading Date 3A* (*HD3A*), [17] in rice. Accumulation of FT or HD3A promotes flowering in both species, but AtCO activates *FT* expression [1] [18] [19] while HD1 seems to repress *HD3A* expression [20]. The accumulation of AtCO and HD1 occurs under long photoperiods [3] [6] and, as a result, extended photoperiods delay the transition to flowering in rice and promote it in Arabidopsis [14].

No DNA-binding activity has been demonstrated for AtCO or for any other COL protein but some of them function as transcription factors [21] [18], probably by interacting with other DNA-binding factors that target promoters [22].

Tomato *CONSTANS-Like* genes

In Tomato, three *COL* genes (*TCOL1, TCOL2* and *TCOL3)* were previously identified by screening, under low stringency, of a tomato genomic library with the AtCO gene as a probe [23]. According to these data, *TCOL2* and *TCOL3* are arranged in tandem while *TCOL1* is located in a different region of the genome. We demonstrated in chapter 2 that this result is incorrect and that all three *TCOL* genes are arranged in tandem in the same genomic region.

Furthermore, sequence data published in the same work [23] indicate that all three genes contain sequences encoding both B-boxes and a CCT domain and, *TCOL2* encodes a polypeptide of 341 amino acids (391 and 409 are the respective lengths of TCOL1 and TCOL3). This result is incorrect

too, and we demonstrate in this chapter that TCOL2 is composed of 143 amino acids, due to a frameshift mutation.

The three *TCOL* genes belong to group I *COL* genes, which are represented in Arabidopsis by *AtCO* and *COL1*-5, and like them, an intron is located before the sequence encoding the conserved C-terminal CCT domain. However, the three tomato genes have an additional intron at the 3' end of the B-box domain, creating a second exon that precisely defines the region encoding the activation domain (AD) of COL proteins [23]. Unlike Arabidopsis and rice, transition to flowering in cultivated tomato is not affected by photoperiod. Nevertheless, in all three species, flowering is accelerated by increased expression of *FT*-like genes [14] [24]. *35S:TCOL1* and *35S:TCOL3* tomato transgenic lines seem to have no alteration of the flowering time, while *35S:TCOL3* Arabidopsis transgenic lines consistently flowered slightly later under long days and earlier under short days [23]. This may imply that the regulation of the tomato *FT* ortholog might not be linked to levels of group I *COL* factors, or perhaps that the tomato ortholog of *FT* has lost promoter motifs required for recognition by COL transcriptional complexes, but not necessarily that *TCOL* genes are not involved in flowering induction in photoperiodic tomato wild species, since in Arabidopsis CO activates flowering through both FT-dependent and FT-independent pathways [1]. Because both TCOL1 and TCOL3 are under circadian regulation, they might mediate other photoperiodic or time-of-day-specific processes, like adaptation to different latitudes.

3.1.2 Cryptochrome genes

Cryptochromes in Arabidopsis

Cryptochromes are flavin-containing blue/UV-A light photoreceptors, first discovered in plants. They share sequence similarity to photolyases, flavoproteins that catalyze the repair of UV light-damaged DNA [25], but do not have photolyase activity [26] [27]. The first cryptochrome gene was isolated through the insertional cloning of an Arabidopsis mutant allelic to *hy4* [28]. When Arabidopsis seedlings are grown under light, they have a shortened hypocotyl respect to seedlings

grown in the dark. The *hy4 / cry1* mutant has hypocotyl longer than wild type when grown under blue or UV-A light [28] [29]. By contrast, the mutant appears similar to wild-type when grown under red or far-red light, where inhibition of hypocotyl growth is induced by the phytochrome photoreceptors [30].

The Arabidopsis cry1 protein affects anthocyanin production and the expression of chalcone synthase gene. *cry1* mutant plants are deficient in these responses [31]. *CRY1* overexpressor plants are hyper sensible to blue / UV-A light, with transgenic seedlings exhibiting unusually short hypocotyls and high levels of anthocyanin [32]. *CRY2*, the second member of the Arabidopsis cryptochrome family, also affects hypocotyl elongation [33]. Both *cry1* and *cry2* mutations affect flowering time [34, 35]. Indeed *cry2* is allelic to *fha*, first characterized as a late-flowering Arabidopsis mutant [36]. There is now evidence for a third cryptochrome gene (*CRY3*) in Arabidopsis, whose sequence is similar to Synechocystis *CRY DASH* (for Drosophila–Arabidopsis–Synechocystis–Human), a transcriptional repressor [37]. Unlike classical plant cryptochromes, cry3 lacks a C-terminal extension and it contains a transit peptide that directs cry3 into the mitochondria and chloroplast [38]. cry3 catalyses light-driven DNA repair like conventional photolyases, but lacks an efficient flipping mechanism for interaction with cyclobutane pyrimidine dimer (CPD) lesions within duplex DNA [39].

Genetic studies have demonstrated that cryptochromes interact with phytochromes in the regulation of photomorphogenic development, floral initiation, and in the entrainment of the circadian clock in Arabidopsis [40] [41] [42] [43] [44]. Concerning the floral initiation, cryptochromes act to enhance CO stability, whereas phytochrome B acts to promote CO degradation [6]. This antagonistic action of cryptochrome and phyB results in the generation of daily rhythms in CO abundance. A naturally found, gain-of-function allele of *CRY2* is responsible for the early flowering in short days of the Cape Verde ecotype of Arabidopsis, with respect to its Northern European counterparts [45] [46].

Cryptochromes in Tomato:

S. lycopersicum is the second plant species in which cryptochromes were cloned and characterized [47] [48] . Four *CRY* genes have been identified in tomato: *CRY1a, CRY1b, CRY2* and *CRY3*. Amino acid sequences of *CRY1a*/*CRY1b* and *CRY2* are more similar to their Arabidopsis counterparts than to each other [48]. The functions of *CRY1a* and *CRY2* have been characterized. Anti-sense transgenic tomato plants expressing reduced level of *CRY1a* and *cry1a* mutant had reduced response of blue light inhibition of long hypocotyls and blue light stimulation of anthocyanin accumulation [47] [49]. Analysis of single, double, and triple mutants of *cry1a*, *phyA* and *phyB* indicated that *CRY1a* acts with *PHYA* and *PHYB* to promote photomorphogenesis and increase anthocyanin accumulation, which is consistent with data from Arabidopsis. More recently, the function of *CRY2* has been characterized through a combination of transgenic overexpression and virus-induced gene silencing (VIGS) [50]. Tomato *CRY2* overexpressors exhibited shortened hypocotyl and internodes under both low- and high fluence blue light, while Arabidopsis *cry2* mutants showed enhanced hypocotyl elongation only under low fluence blue light [33]. Moreover, overexpression of *CRY2* in tomato resulted in overproduction of anthocyanins and chlorophyll in leaves and of flavonoids and lycopene in fruits. It is interesting to note that the *CRY2* overexpressors displayed an unexpected delay in development, and in subsequent flowering, under both SD and LD conditions, and an increased outgrowth of axillary branches [50].

3.1.3 GIGANTEA

GIGANTEA and the circadian clock in Arabidopsis

Physiological experiments identified in the circadian clock the timekeeping mechanism that enables the measurement of photoperiod [18] [51]. The photoperiodic induction of flowering in Arabidopsis thaliana identified with forward genetics approach and the role of the circadian clock in photoperiodic time measurement was confirmed by demonstrating that transcription of the genes that act in this pathway is circadian clock controlled [52]. Mutations in one of these genes,

GIGANTEA (*GI*), both impair circadian rhythms and delay flowering. Arabidopsis *GI* encodes a protein of 1173 amino acids that has no homology with other biochemically characterized proteins [53] [54]. GI is highly conserved in seed plants, also monocotyledonous, such as rice [55], and in gymnosperms, such as loblolly pine (*Pinus taeda*).

gi mutants are characterized by evident alterations in the circadian system regulation. The *gi-1* and *gi-2* mutations have reduced period length in circadian leaf movements, and *gi-1* causes a similar effect in expression of the *CHLOROPHYLL a/b BINDING PROTEIN* (*CAB*) gene, whereas *gi-2* lengthens the period of the latter rhythm [54]. All *gi* alleles cause late flowering under LDs, in contrast with other mutations that cause short period rhythms, such as *timing of cab expression1-1* (*toc1-1*) or *late elongated hypocotyl-11* (*lhy-11*), that induce early flowering under SDs [56] [57]. Some *gi* mutant alleles also cause a long hypocotyl phenotype in deetiolated seedlings, particularly under red light, suggesting impaired PHYB signaling [58]. Similarly, in *gi-1* mutants, circadian period length is less sensitive to increasing light intensity than in wild-type plants [54].

Role of *GI* in promoting flowering in Arabidopsis

The abundance of *CO* mRNA is reduced in *gi* mutants and the 35S:*CO* transgene suppresses the late flowering of *gi* mutants [3]. In early flowering mutants *lhy-11* and *cca1-1, GI* is misexpressed due to impaired control of circadian and diurnal rhythms. The *gi-3* mutation is epistatic to *lhy-11 cca1-1* under SDs and reduced *CO* expression in this background [59]. This supports the idea that GI promotes flowering by acting between the oscillator - which involves *LHY*/*CCA1* - and *CO*. The epistasis of *gi* to the flowering phenotype of *lhy-11 cca1-1* suggests that GI is essential for flowering controlled by circadian clock. Moreover, 35S:*GI* plants show a strong early- flowering phenotype and enhanced expression of *CO* and *FT*, demonstrating that *GI* expression is positively related to flowering-time gene expression. *co-2* and *ft-1* mutations partially suppress the effect on flowering of 35S:*GI* and *lhy-11 cca1-1*. This indicates that the mechanism by which GI promotes early flowering includes CO and its target gene *FT* [59].

GI also appears to promote flowering by a second mechanism that is independent of CO and FT. The delay in flowering of *lhy-11 cca1-1* caused by *co-2* and *ft-1* was weaker than that caused by *gi-3*, suggesting that as well as promoting flowering by activating CO and FT, GI promotes flowering independently of these genes. Similarly, *co-2* and *ft-1* only partially suppressed the early flowering of 35S:*GI* plants [59].

3.2 Aim of the work

As mentioned in the introduction, cultivated tomato (*S. lycopersicum*) is a day neutral plant that can flower regardless of the length of the day. However, some wild species that belong to the *Solanum* section *Lycopersicon,* and four closely related species of section *Juglandifolia* and section *Lycopersicoides*, behave like short day plants when growth at mid-latitudes (see chapter 1). Little information is available on the genetic control of tomato flowering time [60], while no information is available on the genetic mechanisms responsible for the photoperiodic induction of flowering in wild species. The first aim of this thesis (chapter 2) was to provide structural information on the organization of the *CRY*, *COL* and *GI* gene families in totmato. The second aim (this chapter) is to investigate their sequence diversification during speciation in the tomato clade.

3.3 Materials and methods

3.3.1 Plant material and growth conditions

Greenhouse

The species selected for this thesis project are listed in table 1. Their pictures and geographical localization can be observed in the appendix. Seeds were obtained from the Tomato Genetics Resource Centre (TGRC) of Davis (USA). In order to induce flowering in short day species, plants were growth in greenhouse in winter. Seeds of wild species are more recalcitrant to germination

than seeds of the cultivated species. They were soaked in 2.7% sodium hypochlorite for 30 minutes. After bleaching, seeds were rinsed for 15 minutes. For germination, seeds were placed on wet blotter paper in plastic dishes at 25°C. Plantlets were moved to greenhouse, in pots (diameter: 25 cm) in Vigorplant TN soil. In the first month, in order to encourage vegetative development, the photoperiod was set at 16 hours of light and 8 hours of dark, with temperature set at 25°C during the light period and at 20°C during the dark period. Then, it was set at 8 hours of light and 16 hours of dark to promote flowering also in short day species. For all the species, during fruit production, watering was reduced, while for *S. habrochaites*, it was reduced from the early vegetative stage, in order to prevent leaf edema.

For genomic DNA extraction, leaf from 3 plants for each species were collected and used as fresh tissue or stored at -80°C.

Species	Accession	Latitude (degree)	Elevation (m)	Country	Day length preference
S.lycopersicum cv. Heinz	LA4345	-	-	-	Day neutral
S.lycopersicum v. cerasiforme	LA1542	9.9	700	Costa Rica	Day neutral
S.cheesmaniae	LA0428	-0.63	700	Galapagos I.	Short day
S.arcanum	LA2172	-6.2	1000	Peru	Short day
S.arcanum	LA2550	-7.28	1700	Peru	Short day*
S.habrochaites	LA1353	-7.36	2800	Peru	Short day
S.pimpinellifolium	LA0722	-8.11	50	Peru	Day neutral
S.habrochaites	LA1777	-9.55	3150	Peru	Short day*
S.neorickii	LA1319	-13.38	2500	Peru	Day neutral
S.chmielewskii	LA1028	-13.55	2800	Peru	Day neutral
S.pennellii	LA0716	-16.21	100	Peru	Day neutral
S.chilense	LA2930	-25.4	650	Chile	Day neutral**

Table 1. List of selected species with their relative TGRC accession number, ordered by latitude value of the collection site. For each species it is also provided the elevation of the collection site, and the daylength preference according to TGRC. *: confirmed by photoperiodic experiment (data by Salomè Prat at CNB of Madrid (Spain). **: night break experiment disproves TGRC data that classified S. as short day (data by Salomè Prat at CNB of Madrid (Spain).

Growth chambers

In order to produce a diurnal expression pattern of selected genes, two sets of plants were growth in parallel under two different light conditions in climatic chambers:

short day (SD): 16 hours of light (25°C) and 8 hours of dark (20°C).

long day (LD): 8 hours of light (25°C) and 16 hours of dark (20°C).

Light was provided by 3 Osram "L36W/865 cool day light" and 1 "L36W/77 Fluora" fluorescent tube, giving an irradiance of 130 $\mu E\ m^{-2}\ s^{-1}$. For each light condition, leaf samples were collected

from 3 plants for each species at the stage of 4-5 internodes. Sampling was performed every 4 hours, from 00 a.m. to 20 p.m. and leaves were stored at -80°C.

3.3.1 Protocols

Genomic DNA extraction

Genomic DNA extraction was performed using the DNeasy Plant Kit (Qiagen) from fresh or frozen leaf tissue following the standard protocol (catalog n° 69104 and n°68163).

Analytical PCR Primer design

The "Oligo 7" software was utilized to design primer pairs in order to produce amplicons of 800-1000 bp, suitable for subsequent Sanger sequencing. The annealing temperature varied between 50 and 60°C and the primer size between 18 and 22 bases.

Analytical PCR conditions

Reaction were prepared in a total volume of 20 ul/sample:

1x PCR Buffer,

0.2 mM of dNTPs,

50 ng of forward primer,

50 ng of reverse primer,

0.25 U of Taq polimerase,

1 ng of DNA.

Thermocycler conditions:

94 °C x 2 min

35 cycles: 94°C x 45 sec, Ta °C x 45 sec, 72 °C x 1.5 min.

72°C x 5 min

Amplification results were verified via 1% agarose gel electrophoresis. PCR products were purified on Sephadex G100 (GE, catalog n° 17006002) and quantified using a Nanodrop spectrophotometer.

Sequencing pipeline

Purified PCR products were utilized as substrates for sequencing reactions.

Reaction were prepared in a total volume of 10 ul/sample:

1x BigDye Terminator sequencing buffer (AB, catalog n° 4336699),

2 pmoles of forward OR reverse primer,

0.33 uL of BigDye Terminator Ready Reaction mix (AB, catalog n° 4336611).

Thermocycler conditions:

30 cycles: 94 °C x 30 sec, 48 °C x 30 sec, 60 °C x 4 min.

Sequencing reaction products were purified on Sephadex G50 (GE, catalog n° 17004202). Sample are eluted in 5 ul of ABI Prism Hi-Di Formamide (AB, catalog n° 43311320) and loaded into the ABI 3730 DNA Analyser.

Sequence analysis

All the sequence files were analysed with the "CodonCode Aligner" software in order to produce contigs, define coding regions and produce protein sequences for each selected gene in each species. The "MegAlign" software was used to produce alignments of the proteins with the "Clustal W" method.

The "SIFT" software (Sorting intolerant from tolerant - http://blocks.fhcrc.org/sift/SIFT.html) was used to predict whether an amino acid substitution affects protein function. The prediction is based on database sequence homology and the physical properties of amino acids. SIFT values vary between 0 and 1. Values lower than 0.05 correspond to mutations that, theoretically, affect protein

function. Because SIFT usually under-estimates the number of deleterious mutations, we empirically identified values between 0.05 and 0.1 as possible deleterious mutations. For each SIFT value, the software return a quality value that indicate the goodness of the prediction. This value depends on the quality of the alignment with the sequences present in the database.

Phylogenetic analysis

Phylogeny was constructed with the "Mega4" software [61], using the "Neighbor Joining" method (NJ). The NJ method is a simplified version of the "minimum evolution" method, which uses distance measures to correct for multiple hits at the same sites, and chooses a topology showing the smallest value of the sum of all branches as an estimate of the correct tree. With the NJ method, the examination of different topologies is embedded in the algorithm, so that only one final tree is produced. This method does not require the assumption of a constant rate of evolution so it produces an unrooted tree. To tests of the reliability of the tree we used the bootstrap test with 500 replicas and a random number of seeds. Sequence gaps are treated with the "pairwise deletion" option and the substitution model was the "p-distance".

RNA mini-prep extraction

RNA extraction was performed with the RNeasy Plant mini Kit (Qiagen, catalog n° 74104) from frozen leaf tissue following the standard protocol.

Real Time PCR

The "Primer Express v2.0" software was utilized to design primer pairs in coding region in order to produce amplicons of 150 bp, in regions conserved among all the species analysed. Primer pairs were tested through PCR assay (as described before) except for the PCR cycle that is: 35 cycles: 95°C x 15 sec, 60°C x 1 min. Amplification results were verified via 2% agarose gel electrophoresis. PCR products were purified on Sephadex G100 and quantified with the Nanodrop

spectrophotometer. In order to obtain a standard curve for each amplicon, serial dilutions (factor 10) were prepared starting from 1ng/ul until 10ag/ul.

3 replicas for each dilution from 1pg to 10ag were prepared in a total volume of 15 ul:

1X SYBR Green Master Mix (AB, catalog n° 4364346),

50 nM primer forward,

50 nM primer reverse,

1.5 ul of template.

Thermocycler conditions:

50 cycles: 95°C x 15 sec, 60°C x 1 min

Real time PCR was performed using the ABI PRISM 7000 instrument.

The same protocol has been used for the analysis of the unknown samples.

Real time PCR primer pairs:

Actin:

forward: CCCCGCCATACTGGTGTGA

reverse: CTCCTCAGGGGCAACACGAA

TCOL1:

forward: AGGGTGTGTGACAGTTGCCA

reverse: GCTCGCAGGCTTCGCA

TCOL3:

forward: GTCGATGACTACTTGGATCTTGCA

reverse: GTTTTCCCTGTCCGTCCTGA

CRY1a

forward: AGCTGGGTTCCAATTATCCTCTT

reverse: TCAGCAGAGTCACCATGTCCTT

CRY1b

forward: ATTTTAAGGCGTGGAGGCAA

reverse: CCCCATGTCCAAGGAAGCT

CRY2:

forward: GTCGATGCAGGAATGAGGGA

reverse: CTAAATCCGCGTCCAAGAGTG

CRY3:

forward: CCATGTTTGATGCCTGGAGA

reverse: ATCCGCCAGTCGATGCC

TGI1:

forward: AGGCCGTTGTCTCCATGG

reverse: TTCCAGAGCCACGAGAAGAAG

TGI2

forward: GATCTCAAGCCCCCATCAAG

reverse: AGGTAGCAGTCTCGTAGCGAGAA

3.4 Results and discussion

In order to clarify the nature of the different flowering responses arising during tomato speciation, we studied three gene families involved in the photoperiodic regulatory pathway. For each gene, we performed a phylogenetic analysis, in order to assess the degree of evolutionary divergence between the different species. Selection modulates gene sequence evolution in different ways by constraining potential changes of amino acid sequences (purifying selection) or by favouring new and adaptive genetic variants (diversifying selection). The frequency of synonymoous and non-synonymous changes in a pair of protein-coding sequences can be used to quantify the mode and strength of selection. Substitution rates vary across the genome and between species and this makes difficult a direct comparison solely based on non-synonymous substitutions. A standard approach to

control for variation in the underlying mutation rate is to calculate the ratio of the frequency of non-synonymous changes per total number of possible non-synonymous changes (dN) to the frequency of synonymous changes per total number of possible synonymous changes (dS). This ratio (dN/dS) is then used as a measure of the strength of purifying vs diversifying selection. We calculated dN/dS over all species for each gene, while dS and dN values have been calculated for each species/gene with respect to *S. lycopersicum*.

Furthermore, we analysed the protein sequences in order to identify mutations that separate short day species from day neutral species and we evaluated the strength of each identified amino acid mutation with SIFT. To get a deeper understanding of the role and importance that each gene has in the photoperiodic regulatory pathway, we measured the diurnal expression profiles in day neutral and short day species, in different photoperiodic conditions.

The *CRY* and *TCOL* genes were the first to be analysed in all the 12 accessions selected for this thesis project. *TCOL2* presents a frame-shift mutation (described in chapter 2) in all species, indicating that it is a pseudogene in the whole *Lycopersicon* section.

As mentioned in chapter 2, we recently identified two *GIGANTEA* paralogs in tomato, named *TGI1* and *TGI2*. These two genes were studied in 6 of the 12 accessions.

Dendrograms constructed with MEGA4 present a conserved structure for all the genes (Fig. 1, 3, 5, 7, 9, 11, 13 and 15) and are consistent with data from previous work [62] [63]. No lateral gene transfer events seem to have occurred between the species for all the genes analyzed.

Genes will be treated singularly, as they appear in the photoperiodic regulatory pathway:

- Cryptochromes, the blue light photoreceptors that interact upstream with the circadian clock and in Arabidopsis are involved in the stabilization of CO under long days,

- the tomato *GIGANTEA* gene family, whose ortholog in Arabidopsis acts between the circadian clock and *CO*, enhancing its expression.

- *TCOL1* and *3*, whose Arabidopsis ortholog *CO* promotes flowering by activating the transcription of *FT*.

3.4.1 CRY1a

This protein has the lower level of divergence between species, with the shortest sum of branch length (Fig. 1) and the lower number of amino acid mutations (Tab. 3) among all the genes analysed. Despite this, all 5 wild species that diverge from *S. lycopersicum* present predicted strong mutations (Tab. 3) One of the mutations is shared by the two phylogenetically most distant species, *S. pennelli* and the two accessions of *S. habrochaites*.
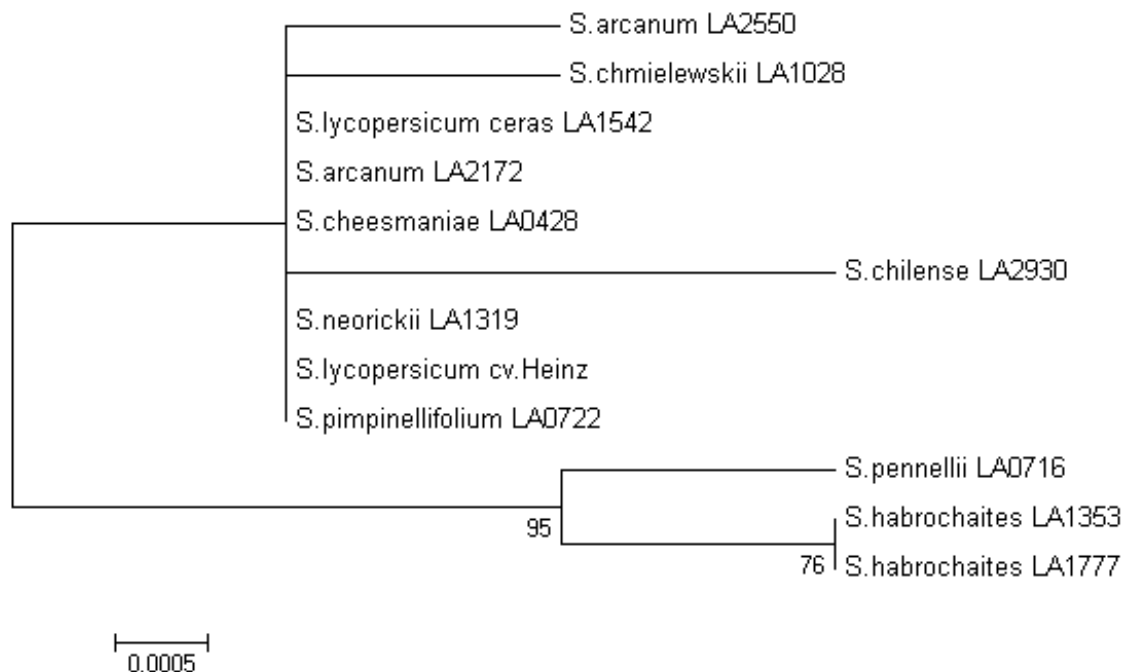
Figure 1. cry1a protein dendrogram. The evolutionary history was inferred using the Neighbor-Joining method [64]. The optimal tree with the sum of branch length = 0.01325479 is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates) are shown next to the branches [65]. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. All positions containing alignment gaps and missing data were eliminated only in pairwise sequence comparisons (Pairwise deletion option). There were a total of 679 positions in the final dataset. Phylogenetic analyses were conducted in MEGA4 [61].

The low level of divergence between species is also observable from the absence of dN variation (Tab. 2). Is interesting to observe that the two accessions of *S. arcanum* diverge phylogenetically, with the accession LA2172 more close to S. *lycopersicum* than the LA2550.

Species	Accession	Latitude (degree)	Elevation (m)	Day length preference	dN	dS
S.lycopersicum v. cerasiforme	LA1542	9.9	700	Day neutral	0	0
S.cheesmaniae	LA0428	-0.63	700	Short day	0	0.002
S.arcanum	LA2172	-6.2	1000	Short day	0	0.017
S. arcanum	LA2550	-7.28	1700	Short day	0	0.015
S.habrochaites	LA1353	-7.36	2800	Short day	0	0.028
S.pimpinellifolium	LA0722	-8.11	50	Day neutral	0	0
S. habrochaites	LA1777	-9.55	3150	Short day	0	0.028
S.neorickii	LA1319	-13.38	2500	Day neutral	0	0.019
S.chmielewskii	LA1028	-13.55	2800	Day neutral	0	0.006
S.pennellii	LA0716	-16.21	100	Day neutral	0	0.026
S.chilense	LA2930	-25.4	650	Day neutral	0	0.017

Table 2. Estimates of Codon-based Evolutionary Divergence between wild species and *S. lycopersicum* for *CRY1a*. The number of dN and dS from analysis between sequences is shown . All results are based on the pairwise analysis of 12 sequences. Analyses were conducted using the Nei-Gojobori method in MEGA4 [66] [61]. All positions containing alignment gaps and missing data were eliminated only in pairwise sequence comparisons (Pairwise deletion option). Data are ordered by latitude value.

Position	Heinz	Subst.	SIFT value	Q. value	Conserved in Arabidopsis	Conserved in Rice	LA1542	LA0722	LA1028	LA1319	LA0716	LA2930	LA0428	LA2172	LA2550	LA1353	LA1777
							DAY NEUTRAL						SHORT DAY				
62	H	Y	0.01	0.94		H					X					X	X
153	A	D	0.54	0.95	A							X					
190	F	I	0.05	0.97	F	F									X		
450	S	I	0.01	1.00	S	S			X								
523	H	R	1.00	0.79		R					X					X	X
558	N	T	0.57	0.81												X	X
566	V	L	0.31	0.65	V						X						
582	V	G	0.38	0.68							X					X	X
623	S	R	0.05	0.46	S							X					

a

Position	Heinz	Subst.	SIFT value	Q. value	Conserved in Arabidopsis	Conserved in Rice	9.9 LA1542	-0.63 LA0428	-6.2 LA2172	-7.28 LA2550	-7.36 LA1353	-8.11 LA0722	-9.55 LA1777	-13.38 LA1319	-13.55 LA1028	-16.21 LA0716	-25.4 LA2930
62	H	Y	0.01	0.94		H					X		X			X	
153	A	D	0.54	0.95	A												X
190	F	I	0.05	0.97	F	F				X							
450	S	I	0.01	1.00	S	S									X		
523	H	R	1.00	0.79		R					X		X			X	
558	N	T	0.57	0.81							X		X				
566	V	L	0.31	0.65	V											X	
582	V	G	0.38	0.68							X		X			X	
623	S	R	0.05	0.46	S												X

b

Table 3. SIFT values of substitution in cry1a. *S. lycopersicum* cv. Heinz has beenused as reference sequence. SIFT values lower than 0.05 (red) indicate that the substitution affect the protein function. Values between 0.05 and 0.1 (orange) indicate possible deleterious substitutions. Values higher than 0.1 (black) indicate tolerated substitution. The prediction is based on sequence homology and the physical properties of amino acids. **a:** data grouped according to day length preference; **b:** data grouped according to latitude value.

Expression of *CRY1a* shows a diurnal rhythm, and differs between day neutral and short day species, although not as expected. In day neutral species the expression peak of *CRY1a* is delayed in LD conditions, from ZT 8 to ZT 12 (ZT=Zeitgeber Time, i.e. hours after light is switched on) while in short day plants the expression peak is synchronized in LD and SD conditions (Fig. 2). This finding is somewhat puzzling, since cry proteins have been hypothesized in Arabidopsis to be positive regulators of flowering. On the other hand, as shown in Chapter 4, tomato cry1a and cry2 are negative regulators of flowering. Thus, the expression of *CRY1a* at ZT8 in LD condition in short

day plants may be in agreement with its functional role as a negative regulator of flowering (chapter 4).
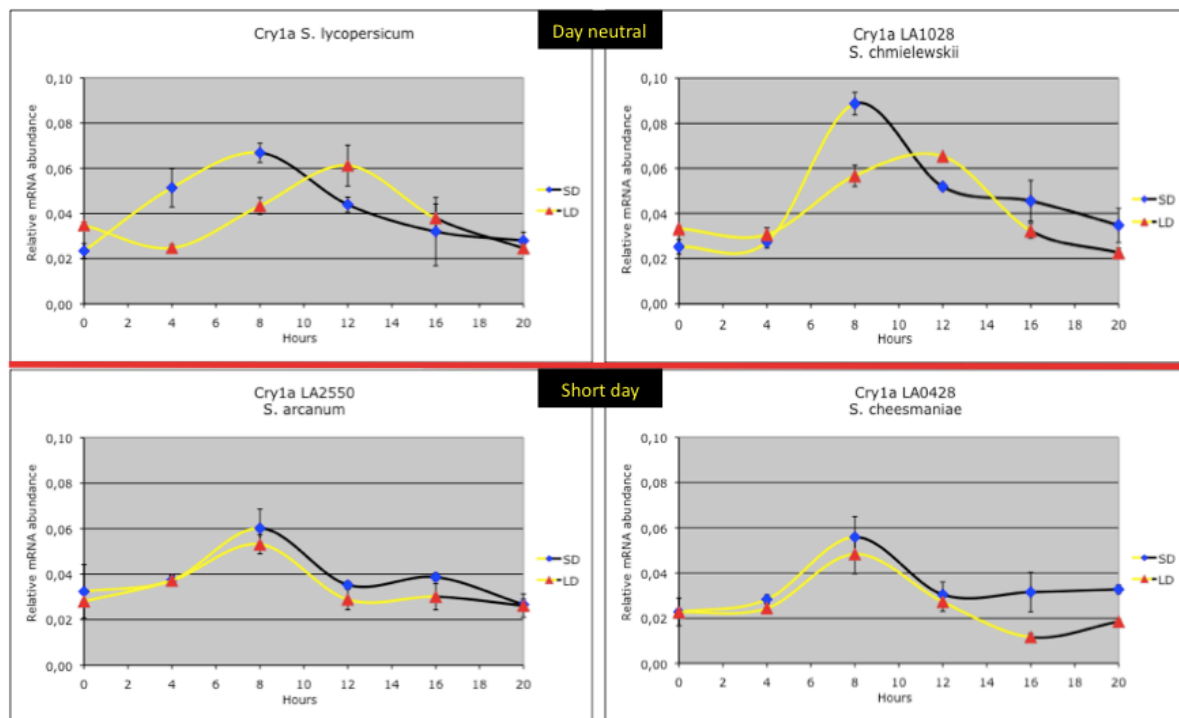
Figure 2. Diurnal oscillations of *CRY1a* transcripts analyzed by QRT-PCR in two different photoperiodic conditions: short day (SD: 8 hours of ligh and 16 hours of dark) and long day (LD: 16 hours of ligh and 8 hours of dark). Four species have been analysed: two day neutral species (*S. lycopersicum* and *S.chmielewskii*) and two short day species (*S. arcanum* and *S. cheesmaniae*). Results are expressed as relative mRNA abundance, after normalization with β-actin. Yellow and black lines represent light and dark periods, respectively. Time points are measured in hours from dawn. Data shown are the average of two biological and 3 technical replicates, with error bars representing standard deviation.

In order to correlate expression and sequence data, we sequenced a part of the 5'UTR/promoter region of each gene. Since the tomato genome has been released recently, we based our work on sequences already present in the database. For *CRY1a*, about 590 bp before the start codon were available and about 500 were sequenced in the four species. According to iTAG data (see chapter 2), 370 of the available 500 bp belong to the 5' UTR region. In the remaining 130 bp only one mutation respect to *S.lycopersicum* has been identified in position -385 and is common to the four wild species.

3.4.2 CRY1b

The protein has a higher level of divergence between species respect to cry1a (Fig. 3) and a higher number of total mutations identified in the amino acid sequence (Tab. 5), although the number of predicted strong mutations is lower than in cry1a. Again, the *S. arcanum* accessions diverge from each other.
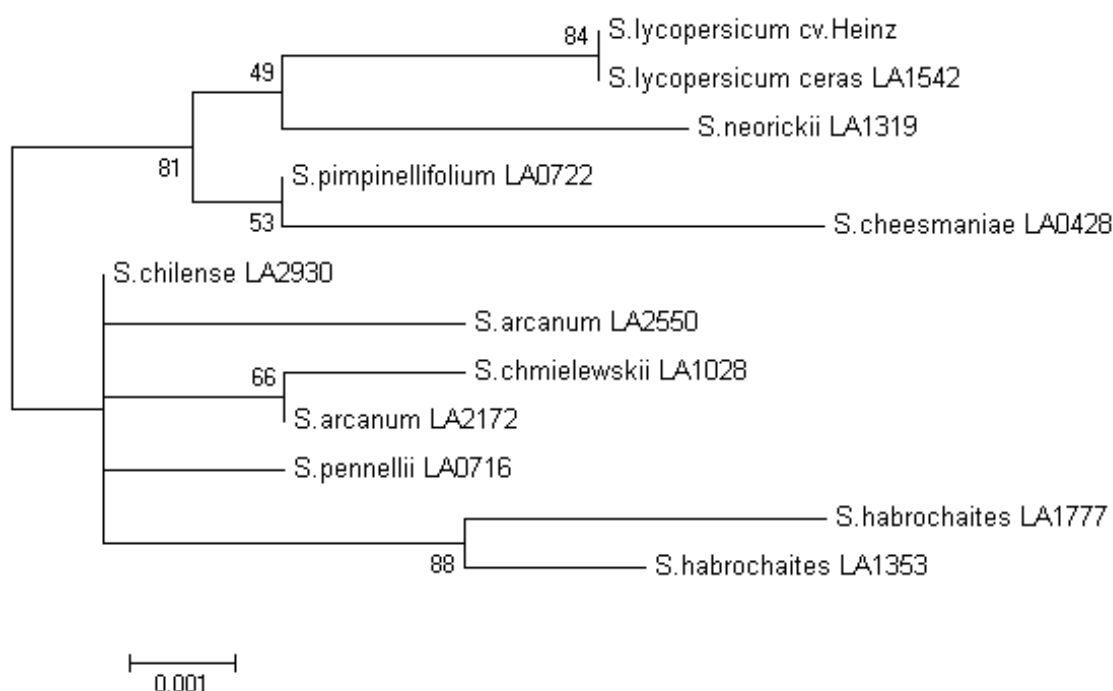
Figure 3. cry1b protein dendrogram. The evolutionary history was inferred using the Neighbor-Joining method [64]. The optimal tree with the sum of branch length = 0.03344768 is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates) are shown next to the branches [65]. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. All positions containing alignment gaps and missing data were eliminated only in pairwise sequence comparisons (Pairwise deletion option). There were a total of 583 positions in the final dataset. Phylogenetic analyses were conducted in MEGA4 [61].

Species	Accession	Latitude (degree)	Elevation (m)	Day length preference	dN	dS
S.lycopersicum v. cerasiforme	LA1542	9.9	700	Day neutral	0	0
S.cheesmaniae	LA0428	-0.63	700	Short day	0.004	0
S.arcanum	LA2172	-6.2	1000	Short day	0.004	0.018
S. arcanum	LA2550	-7.28	1700	Short day	0.004	0.015
S.habrochaites	LA1353	-7.36	2800	Short day	0.005	0.036
S.pimpinellifolium	LA0722	-8.11	50	Day neutral	0.001	0
S. habrochaites	LA1777	-9.55	3150	Short day	0.006	0.036
S.neorickii	LA1319	-13.38	2500	Day neutral	0.003	0.028
S.chmielewskii	LA1028	-13.55	2800	Day neutral	0.004	0.015
S.pennellii	LA0716	-16.21	100	Day neutral	0.004	0.025
S.chilense	LA2930	-25.4	650	Day neutral	0.003	0.025

Table 4. Estimates of Codon-based Evolutionary Divergence between wild species and *S. lycopersicum* for *CRY1b*. The number of dN and dS from analysis between sequences is shown . All results are based on the pairwise analysis of 12 sequences. Analyses were conducted using the Nei-Gojobori method in MEGA4 [66] [61]. All positions containing alignment gaps and missing data were eliminated only in pairwise sequence comparisons (Pairwise deletion option). Data are ordered by latitude value.

Table a: data grouped according to day length preference

Position	Heinz	Subst.	SIFT value	Q. value	Conserved in Arabidopsis	Conserved in Rice	DAY NEUTRAL						SHORT DAY				
							LA 1542	LA 0722	LA 1028	LA 1319	LA 0716	LA 2930	LA 0428	LA 2172	LA 2550	LA 1353	LA 1777
26	A	G	0.12	0.97	A	A			X	X	X	X		X	X	X	X
32	V	M	0.00	0.97	V	V				X							
41	E	K	0.20	0.97	E	E									X		
89	V	I	0.87	0.97	V										X		
98	V	I	0.15	0.97	V						X						
114	C	R	1.00	0.97	R	R		X	X		X	X	X	X	X	X	X
125	S	A	0.75	0.98	A	A						X					
131	A	G	1.00	0.98	A	A				X							
145	S	G	1.00	0.98	G	G				X		X	X		X	X	X
160	T	S	1.00	1.00	S											X	X
296	M	T	0.00	1.00									X				
309	L	I	0.04	1.00	L	L							X				
422	N	H	0.82	1.00	N	H			X					X			
434	A	S	0.28	1.00	S	A											X
453	Q	K	0.93	0.98	Q	Q							X				
505	D	E	0.41	0.89	D	D										X	
521	I	M	1.00	0.60	M	M		X	X	X	X	X	X	X	X	X	X

a

Table b: data grouped according to latitude value

Position	Heinz	Subst.	SIFT value	Q. value	Conserved in Arabidopsis	Conserved in Rice	9.9	-0.63	-6.2	-7.28	-7.36	-8.11	-9.55	-13.38	-13.55	-16.21	-25.4
							LA 1542	LA 0428	LA 2172	LA 2550	LA 1353	LA 0722	LA 1777	LA 1319	LA 1028	LA 0716	LA 2930
26	A	G	0.12	0.97	A	A			X	X	X		X	X	X	X	X
32	V	M	0.00	0.97	V	V								X			
41	E	K	0.20	0.97	E	E					X						
89	V	I	0.87	0.97	V						X						
98	V	I	0.15	0.97	V										X		
114	C	R	1.00	0.97	R	R			X	X	X	X	X		X	X	X
125	S	A	0.75	0.98	A	A										X	
131	A	G	1.00	0.98	A	A								X			
145	S	G	1.00	0.98	G	G				X	X		X		X	X	X
160	T	S	1.00	1.00	S							X	X				
296	M	T	0.00	1.00						X							
309	L	I	0.04	1.00	L	L				X							
422	N	H	0.82	1.00	N	H			X						X		
434	A	S	0.28	1.00	S	A							X				
453	Q	K	0.93	0.98	Q	Q			X								
505	D	E	0.41	0.89	D	D						X					
521	I	M	1.00	0.60	M	M		X	X	X	X	X	X	X	X	X	X

b

Table 5. SIFT values of substitution in cry1b. *S. lycopersicum* cv. Heinz has been used as reference sequence. SIFT values lower than 0.05 (red) indicate that the substitution affect the protein function. Values between 0.05 and 0.1 (orange) indicate possible deleterious substitutions. Values higher than 0.1 (black) indicate tolerated substitution. The prediction is based on sequence homology and the physical properties of amino acids. **a:** data grouped according to day length preference; **b:** data grouped according to latitude value.

The expression patterns of *CRY1b* are not significantly different in day neutral and short day species. The expression peak is found at ZT12 in all conditions and genotypes, with the exception of *S. arcanum,* in which it is found at ZT8 in SD and ZT16 in LD conditions (Fig. 4).
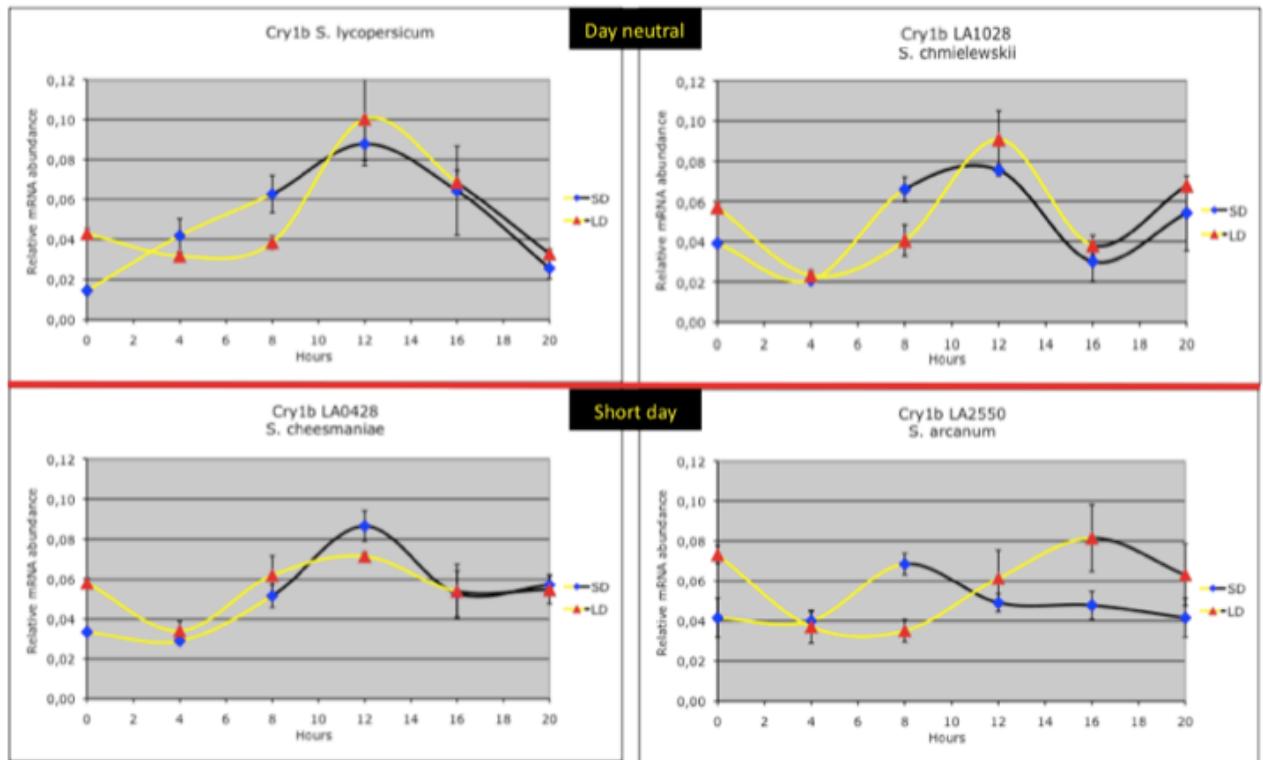
Figure 4. Diurnal oscillations of *CRY1b* transcripts analyzed by QRT-PCR in two different photoperiodic condition: short day (SD: 8 hours of ligh and 16 hours of dark) and long day (LD: 16 hours of ligh and 8 hours of dark). Four species have been analysed: two day neutral species (*S. lycopersicum* and *S.chmielewskii*) and two short day species (*S. arcanum* and *S. cheesmaniae*). Results are expressed as relative mRNA abundance, after normalization with β-actin. Yellow and black lines represent light and dark periods, respectively. Time points are measured in hours from dawn. Data shown are the average of two biological replica and 3 technical replicates, with error bars representing standard deviation.

For *CRY1b*, about 800 bp before the start codon were available and about 720 were sequenced in the four species. According to iTAG data, 496 of the available 720 bp belong to the 5' UTR region. In the remaining 224 bp 15 mutation (SNP and indels) have been identified among all the wild species respect to *S.lycopersicum*. All the promoter mutations found in *S. arcanum*, in which the peak in LD conditions is delayed respect to the other three species, are shared also by *S. chmielewskii*.

67

3.4.3 CRY2

cry2 has a high level of divergence between species, with the sum of branch length about two times respect to cry1b and more than three times respect to cry1a (Fig. 5). The number of predicted strong mutations is higher too, and they are all located close to the N-terminus, where the protein is more conserved in Arabidopsis and rice (Tab. 7). Here also there is a divergence between the two *S. arcanum* accessions.
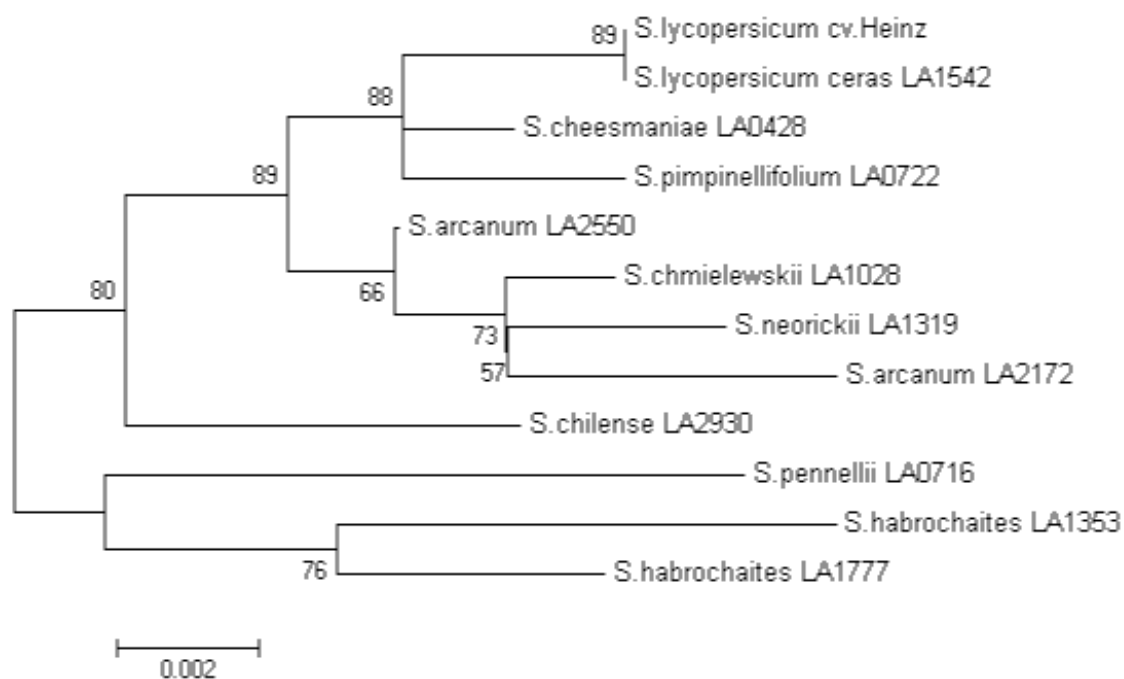

Figure 5. cry2 protein dendrogram. The evolutionary history was inferred using the Neighbor-Joining method [64]. The optimal tree with the sum of branch length = 0.05585835 is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates) are shown next to the branches [65]. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. All positions containing alignment gaps and missing data were eliminated only in pairwise sequence comparisons (Pairwise deletion option). There were a total of 647 positions in the final dataset. Phylogenetic analyses were conducted in MEGA4 [61].

Species	Accession	Latitude (degree)	Elevation (m)	Day length preference	dN	dS
S.lycopersicum v. cerasiforme	LA1542	9.9	700	Day neutral	0	0
S.cheesmaniae	LA0428	-0.63	700	Short day	0.002	0.009
S.arcanum	LA2172	-6.2	1000	Short day	0.005	0.021
S. arcanum	LA2550	-7.28	1700	Short day	0.003	0.021
S.habrochaites	LA1353	-7.36	2800	Short day	0.009	0.041
S.pimpinellifolium	LA0722	-8.11	50	Day neutral	0.003	0.012
S. habrochaites	LA1777	-9.55	3150	Short day	0.007	0.049
S.neorickii	LA1319	-13.38	2500	Day neutral	0.005	0.019
S.chmielewskii	LA1028	-13.55	2800	Day neutral	0.004	0.021
S.pennellii	LA0716	-16.21	100	Day neutral	0.008	0.046
S.chilense	LA2930	-25.4	650	Day neutral	0.005	0.034

Table 6. Estimates of Codon-based Evolutionary Divergence between wild species and *S. lycopersicum* for *CRY2*. The number of dN and dS from analysis between sequences is shown . All results are based on the pairwise analysis of 12 sequences. Analyses were conducted using the Nei-Gojobori method in MEGA4 [66] [61]. All positions containing

alignment gaps and missing data were eliminated only in pairwise sequence comparisons (Pairwise deletion option). Data are ordered by latitude value.

Position	Heinz	Subst.	SIFT value	Q. value	Conserved in Arabidopsis	Conserved in Rice	DAY NEUTRAL						SHORT DAY				
							LA 1542	LA 0722	LA 1028	LA 1319	LA 0716	LA 2930	LA 0428	LA 2172	LA 2550	LA 1353	LA 1777
2	E	E	0.86	0.67								X					
5	Y	S	0.90	0.82				X	X	X	X	X	X	X	X	X	X
32	V	L	0.01	1.00	V	V										X	X
33	L	I	0.75	1.00		L								X			
44	G	A	0.01	1.00	G	G							X				
75	V	L	0.08	1.00		V				X	X			X			
77	M	I	0.11	1.00	I	I			X	X	X	X		X	X	X	X
84	S	Y	0.01	1.00	S						X						
97	K	R	0.77	1.00	K	R					X						
121	D	E	1.00	1.00	E							X				X	X
142	D	N	0.09	1.00		D					X						
159	L	S	0.00	1.00	L											X	X
170	P	L	0.28	1.00	P				X								
186	V	I	0.48	1.00		I					X	X				X	X
205	K	R	1.00	1.00	R	R					X						
219	T	I	0.04	1.00										X			
220	E	A	0.75	1.00	E	E								X			
323	A	S	0.85	1.00													X
372	Q	K	0.51	1.00	K							X					
484	I	S	0.42	0.97							X					X	X
496	H	D	0.62	0.93				X	X	X	X	X	X	X	X	X	X
497	A	S	0.71	0.93											X		
503	V	I	0.14	0.85	I					X							
514	S	L	0.41	0.82											X		
517	I	T	0.42	0.83							X						X
546	G	C	0.11	0.47													X
546	G	S	0.81	0.47							X					X	
552	P	A	0.49	0.33								X					
571	E	G	0.38	0.45												X	
572	R	S	0.42	0.45							X						
573	I	V	0.30	0.45							X						
580	G	S	0.73	0.42								X					
596	T	S	1.00	0.42												X	

a

Position	Heinz	Subst.	SIFT value	Q. value	Conserved in Arabidopsis	Conserved in Rice	9.9	-0.63	-6.2	-7.28	-7.36	-8.11	-9.55	-13.38	-13.55	-16.21	-25.4
							LA 1542	LA 0428	LA 2172	LA 2550	LA 1353	LA 0722	LA 1777	LA 1319	LA 1028	LA 0716	LA 2930
2	E	E	0.86	0.67													X
5	Y	S	0.90	0.82				X	X	X	X	X	X	X	X	X	X
32	V	L	0.01	1.00	V	V					X		X				
33	L	I	0.75	1.00		L			X								
44	G	A	0.01	1.00	G	G		X									
75	V	L	0.08	1.00		V			X					X		X	
77	M	I	0.11	1.00	I	I			X	X	X		X	X	X	X	X
84	S	Y	0.01	1.00	S											X	
97	K	R	0.77	1.00	K	R										X	
121	D	E	1.00	1.00	E						X		X				X
142	D	N	0.09	1.00		D										X	
159	L	S	0.00	1.00	L						X		X				
170	P	L	0.28	1.00	P										X		
186	V	I	0.48	1.00		I					X		X			X	X
205	K	R	1.00	1.00	R	R										X	
219	T	I	0.04	1.00					X								
220	E	A	0.75	1.00	E	E			X								
323	A	S	0.85	1.00									X				
372	Q	K	0.51	1.00	K												X
484	I	S	0.42	0.97							X		X			X	
496	H	D	0.62	0.93				X	X	X	X	X	X	X	X	X	X
497	A	S	0.71	0.93						X							
503	V	I	0.14	0.85	I									X			
514	S	L	0.41	0.82						X							
517	I	T	0.42	0.83									X			X	
546	G	C	0.11	0.47									X				
546	G	S	0.81	0.47							X					X	
552	P	A	0.49	0.33													X
571	E	G	0.38	0.45							X						
572	R	S	0.42	0.45										X			
573	I	V	0.30	0.45										X			
580	G	S	0.73	0.42												X	
596	T	S	1.00	0.42							X						

b

Table 7. SIFT values of substitution in cry2. *S. lycopersicum* cv. Heinz has been used as reference sequence. SIFT values lower than 0.05 (red) indicate that the substitution affect the protein function. Values between 0.05 and 0.1 (orange) indicate possible deleterious substitutions. Values higher than 0.1 (black) indicate tolerated substituion. The prediction is based on sequence homology and the physical properties of amino acids. **a:** data grouped according to day length preference; **b:** data grouped according to latitude value.

Similar to *CRY1a,* the expression peak is dealyed in LD conditions. No evident differences in expression are instead found in short day vs long day species (Fig 6).
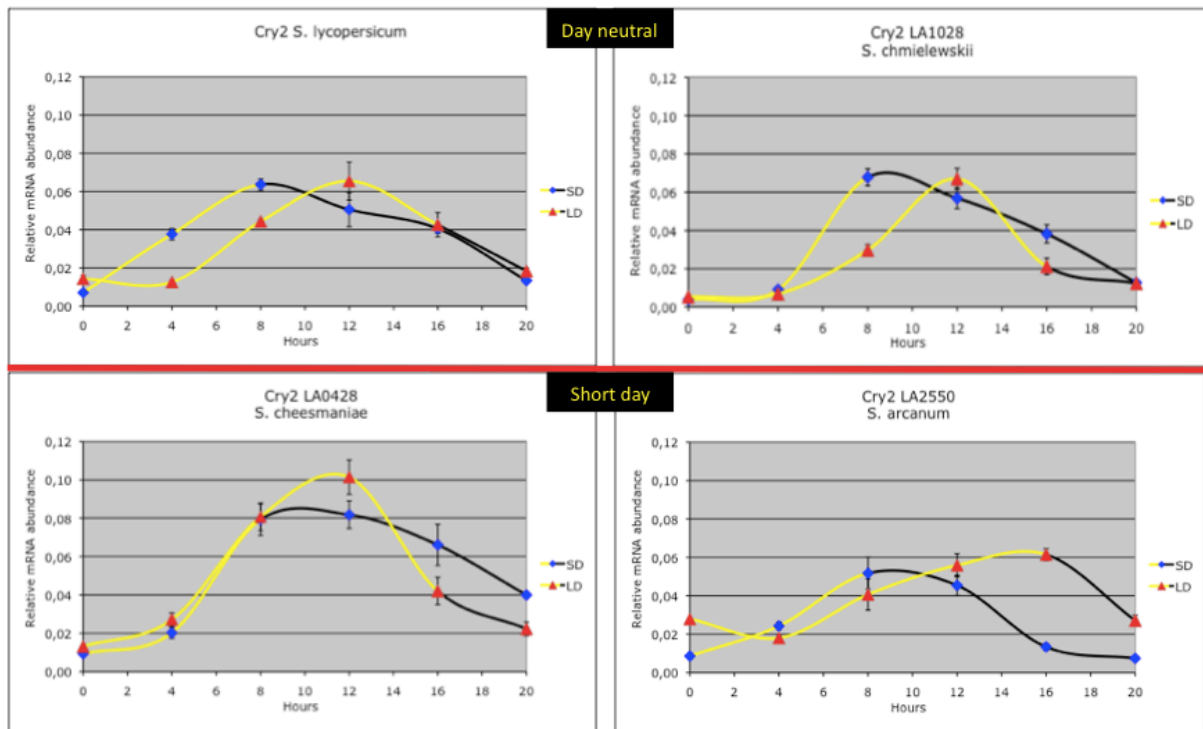

Figure 6. Diurnal oscillations of *CRY2* transcripts analyzed by QRT-PCR in two different photoperiodic condition: short day (SD: 8 hours of ligh and 16 hours of dark) and long day (LD: 16 hours of ligh and 8 hours of dark). Four species have been analysed: two day neutral species (*S. lycopersicum* and *S.chmielewskii*) and two short day species (*S. arcanum* and *S. cheesmaniae*). Results are expressed as relative mRNA abundance, after normalization with β-actin. Yellow and black lines represent light and dark periods, respectively. Time points are measured in hours from dawn. Data shown are the average of two biological replica and 3 technical replicates, with error bars representing standard deviation.

For *CRY2*, only 420 bp upstream of the start codon are available and about 340 were sequenced in the four species. According to iTAG, all this region belong to the 5' UTR region.

3.4.3 CRY3

The level of divergence between species is similar to cry1b (Fig. 7), but with a higher presence of

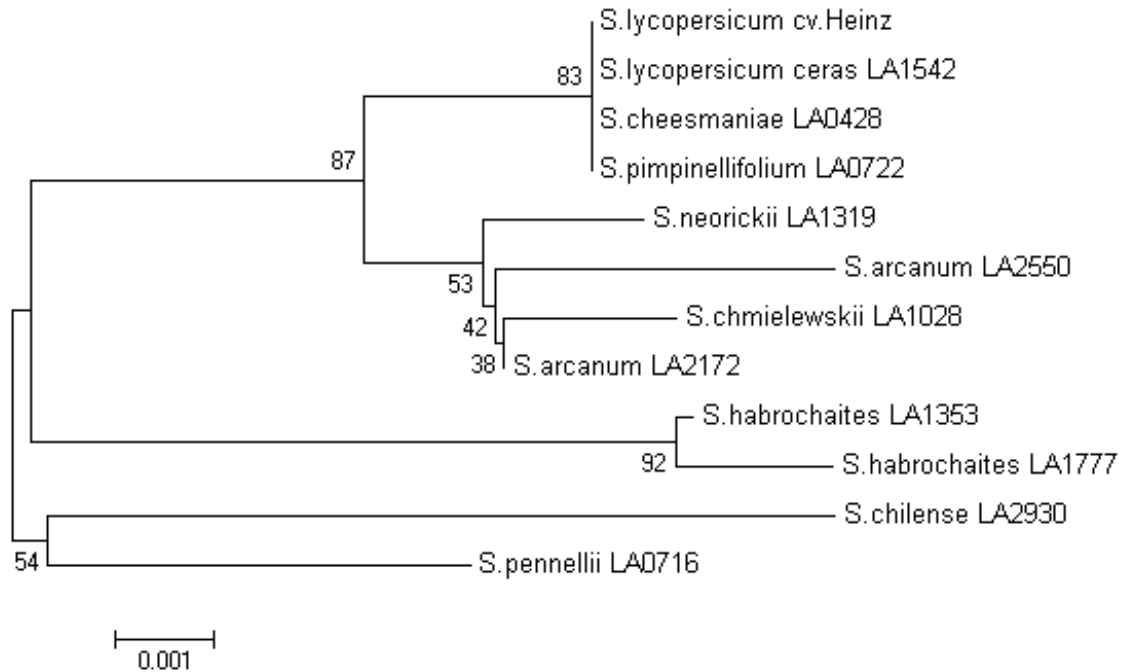predicted strong mutations, three of which are present in *S. arcanum* LA2550 (Tab. 9).

Figure 7. cry3 protein dendrogram. The evolutionary history was inferred using the Neighbor-Joining method [64]. The optimal tree with the sum of branch length = 0.03465467 is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates) are shown next to the branches [65]. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. All positions containing alignment gaps and missing data were eliminated only in pairwise sequence comparisons (Pairwise deletion option). There were a total of 577 positions in the final dataset. Phylogenetic analyses were conducted in MEGA4 [61].

Species	Accession	Latitude (degree)	Elevation (m)	Day length preference	dN	dS
S.lycopersicum v. cerasiforme	LA1542	9.9	700	Day neutral	0	0
S.cheesmaniae	LA0428	-0.63	700	Short day	0	0
S.arcanum	LA2172	-6.2	1000	Short day	0.002	0.015
S. arcanum	LA2550	-7.28	1700	Short day	0.003	0.015
S.habrochaites	LA1353	-7.36	2800	Short day	0.005	0.021
S.pimpinellifolium	LA0722	-8.11	50	Day neutral	0	0
S. habrochaites	LA1777	-9.55	3150	Short day	0.006	0.021
S.neorickii	LA1319	-13.38	2500	Day neutral	0.002	0.015
S.chmielewskii	LA1028	-13.55	2800	Day neutral	0.002	0.015
S.pennellii	LA0716	-16.21	100	Day neutral	0.005	0.015
S.chilense	LA2930	-25.4	650	Day neutral	0.006	0.037

Table 8. Estimates of Codon-based Evolutionary Divergence between wild species and *S. lycopersicum* for *CRY3*. The number of dN and dS from analysis between sequences is shown . All results are based on the pairwise analysis of 12 sequences. Analyses were conducted using the Nei-Gojobori method in MEGA4 [66] [61]. All positions containing alignment gaps and missing data were eliminated only in pairwise sequence comparisons (Pairwise deletion option). Data are ordered by latitude value.

a:

Position	Heinz	Subst.	SIFT value	Q. value	Conserved in Arabidopsis	Conserved in Rice	DAY NEUTRAL						SHORT DAY				
							LA 1542	LA 0722	LA 1028	LA 1319	LA 0716	LA 2930	LA 0428	LA 2172	LA 2550	LA 1353	LA 1777
20	H	N	0.00	0.06											X		
19-22		deletion														X	X
27	C	Y	0.22	0.06						X							
34	A	T	0.10	0.10											X		
55	M	T	0.21	0.19							X						X
155	L	F	0.03	1.00	L	L										X	X
193	M	I	0.19	1.00						X							
193	M	L	0.32	1.00	L	L					X	X				X	X
211	N	S	0.98	0.19		S										X	X
216	G	Q	0.61	0.39							X	X				X	X
240	E	Q	0.50	1.00								X					
295	S	L	0.18	0.97								X					
295	S	T	0.34	0.97	T						X						
320	L	S	0.41	1.00	L							X					
392	A	G	1.00	1.00	G	G						X					
398	A	P	0.31	1.00								X					
425	Y	F	0.35	1.00	Y	Y										X	X
440	Y	F	1.00	1.00	F	F			X	X	X	X	X	X		X	X
457	M	I	0.02	1.00	M	M			X	X			X	X			
544	A	T	0.53	0.94												X	X
C-term.		4 extra aa														X	X

b:

Position	Heinz	Subst.	SIFT value	Q. value	Conserved in Arabidopsis	Conserved in Rice	9.9	-0.63	-6.2	-7.28	-7.36	-8.11	-9.55	-13.38	-13.55	-16.21	-25.4
							LA 1542	LA 0428	LA 2172	LA 2550	LA 1353	LA 0722	LA 1777	LA 1319	LA 1028	LA 0716	LA 2930
20	H	N	0.00	0.06						X							
19-22		deletion									X		X				
27	C	Y	0.22	0.06										X			
34	A	T	0.10	0.10						X							
55	M	T	0.21	0.19									X			X	
155	L	F	0.03	1.00	L	L					X		X				
193	M	I	0.19	1.00										X			
193	M	L	0.32	1.00	L	L					X		X			X	X
211	N	S	0.98	0.19		S					X		X				
216	G	Q	0.61	0.39							X		X			X	X
240	E	Q	0.50	1.00													X
295	S	L	0.18	0.97													X
295	S	T	0.34	0.97	T											X	
320	L	S	0.41	1.00	L												X
392	A	G	1.00	1.00	G	G											X
398	A	P	0.31	1.00													X
425	Y	F	0.35	1.00	Y	Y					X		X				
440	Y	F	1.00	1.00	F	F		X	X		X		X	X	X	X	X
457	M	I	0.02	1.00	M	M		X	X					X	X		
544	A	T	0.53	0.94							X		X				
C-term.		4 extra aa									X		X				

Table 9. SIFT values of substitution in cry3. *S. lycopersicum* cv. Heinz has been used as reference sequence. SIFT values lower than 0.05 (red) indicate that the substitution affect the protein function. Values between 0.05 and 0.1 (orange) indicate possible deleterious substitutions. Values higher than 0.1 (black) indicate tolerated substituion. The prediction is based on sequence homology and the physical properties of amino acids. **a:** data grouped according to day length preference; **b:** data grouped according to latitude value.

In *S. arcanum* LA2550, the expression pattern differs strongly from the other species: in LD conditions an expression trough is observed, instead of a peak (Fig. 8). It is also noteworthy that *S. arcanum* cry3 contains 3 strong mutations, as predicted by SIFT, one of which is in a Methionine conserved in both Arabidopsis and Rice.
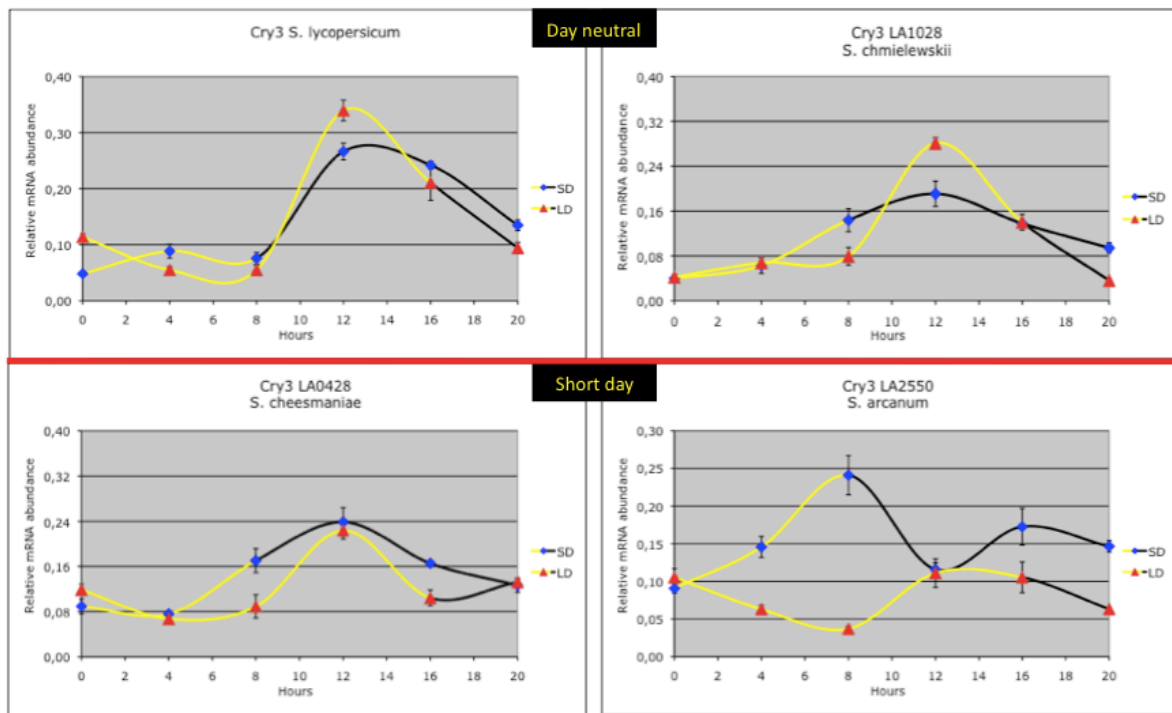
Figure 8. Diurnal oscillations of *CRY3* transcripts analyzed by QRT-PCR in two different photoperiodic condition: short day (SD: 8 hours of ligh and 16 hours of dark) and long day (LD: 16 hours of ligh and 8 hours of dark). Four species have been analysed: two day neutral species (*S. lycopersicum* and *S.chmielewskii*) and two short day species (*S. arcanum* and *S. cheesmaniae*). Results are expressed as relative mRNA abundance, after normalization with β-actin. Yellow and black lines represent light and dark periods, respectively. Time points are measured in hours from dawn. Data shown are the average of two biological replica and 3 technical replicates, with error bars representing standard deviation.

For *CRY3*, about 730 bp upstream of the start codon were available and about 650 were sequenced in the four species. No informations about 5' UTR region are provided by iTAG. Assuming that also for *CRY3* about 400 bp belong to the 5' UTR, 250 bp of putative promoter are available. 36 mutations (SNP and indels) have been identified among all the wild species with respect to *S.lycopersicum.* One of them is present only in *S. arcanum* and, according to "Jaspar Core" (http://jaspar.cgb.ki.se/cgi-bin/jaspar_db.pl), an eukaryotic transcription factor database, corresponds to a possible binding site of "squamosa", a MADS-box protein that is involved in the control of floral architecture in *Antirrhinum majus* [67] (Fig. 9).

Figure 9. Promoter motif recognized by "squamosa" MADS-box protein (boxed). A mutation in the last base is present in *S. arcanum* LA2550.

3.4.4 TGI1

Although the analyses of TGI1 were conducted only on 6 species, it's interesting to note that among them, a high degree of variability is present (Fig. 10 and Tab. 11), with a high percentage of predicted strong mutations in *S. chilense* and *S. habrochaites*.
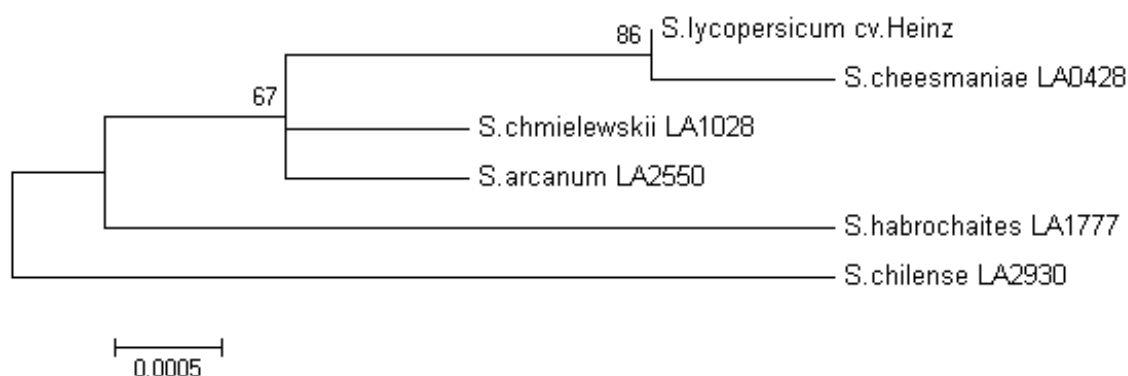

Figure 10. TGI1 protein dendrogram. The evolutionary history was inferred using the Neighbor-Joining method [64]. The optimal tree with the sum of branch length = 0.01285949 is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates) are shown next to the branches [65]. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. All positions containing alignment gaps and missing data were eliminated only in pairwise sequence comparisons (Pairwise deletion option). There were a total of 1167 positions in the final dataset. Phylogenetic analyses were conducted in MEGA4 [61].

Species	Accession	Latitude (degree)	Elevation (m)	Day length preference	dN	dS
S.cheesmaniae	LA0428	-0.63	700	Short day	0	0.001
S. arcanum	LA2550	-7.28	1700	Short day	0.001	0.008
S. habrochaites	LA1777	-9.55	3150	Short day	0.003	0.019
S.chmielewskii	LA1028	-13.55	2800	Day neutral	0.001	0.012
S.chilense	LA2930	-25.4	650	Day neutral	0.003	0.014

Table 10. Estimates of Codon-based Evolutionary Divergence between wild species and *S. lycopersicum* for *TGI1*. The number of dN and dS from analysis between sequences is shown . All results are based on the pairwise analysis of 6 sequences. Analyses were conducted using the Nei-Gojobori method in MEGA4 [66] [61]. All positions containing alignment gaps and missing data were eliminated only in pairwise sequence comparisons (Pairwise deletion option). Data are ordered by latitude value.

Position	Heinz	Subst.	SIFT value	Q. value	Conserved in Arabidopsis	Conserved in Rice	DAY NEUTRAL		SHORT DAY		
							LA1028	LA2930	LA0428	LA2550	LA1777
64	N	T	0.06	0.46							X
172	G	E	1.00	0.44	E	E	X	X		X	X
182	R	M	0.00	0.46	R	R		X			
285	P	A	0.00	0.37	P	P			X		
356	T	S	1.00	0.39	S	S		X		X	
420	G	D	0.09	0.37	G	G					X
421	G	S	0.00	0.37	G	G					X
491	L	V	0.01	0.43				X			
514	S	G	0.00	0.43	S	S		X			
835		deletion						X			
838	C	S	0.61	0.51						X	
842	T	I	0.44	0.65	T						X
848	T	A	0.28	0.65			X				
860	A	E	0.00	0.43				X			
903	T	A	1.00	0.44	A		X	X		X	X
C-terminus		1 extra aa						X			

a

Position	Heinz	Subst.	SIFT value	Q. value	Conserved in Arabidopsis	Conserved in Rice	-0.63	-7.28	-9.55	-13.55	-25.4
							LA0428	LA2550	LA1777	LA1028	LA2930
64	N	T	0.06	0.46					X		
172	G	E	1.00	0.44	E	E		X	X	X	X
182	R	M	0.00	0.46	R	R					X
285	P	A	0.00	0.37	P	P	X				
356	T	S	1.00	0.39	S	S		X			X
420	G	D	0.09	0.37	G	G			X		
421	G	S	0.00	0.37	G	G			X		
491	L	V	0.01	0.43							X
514	S	G	0.00	0.43	S	S					X
835		deletion									X
838	C	S	0.61	0.51				X			
842	T	I	0.44	0.65	T				X		
848	T	A	0.28	0.65						X	
860	A	E	0.00	0.43							X
903	T	A	1.00	0.44	A			X	X	X	X
C-terminus		1 extra aa									X

b

Table 11. SIFT values of substitution in TGI1. *S. lycopersicum* cv. Heinz has been used as reference sequence. SIFT values lower than 0.05 (red) indicate that the substitution affect the protein function. Values between 0.05 and 0.1 (orange) indicate possible deleterious substitutions. Values higher than 0.1 (black) indicate tolerated substituion. The prediction is based on sequence homology and the physical properties of amino acids. **a:** data grouped according to day length preference; **b:** data grouped according to latitude value.

The expression pattern is highly conserved among the four analysed species, and presents a delay of

the expression peak, from ZT8 to ZT12, in LD conditions (Fig. 11).
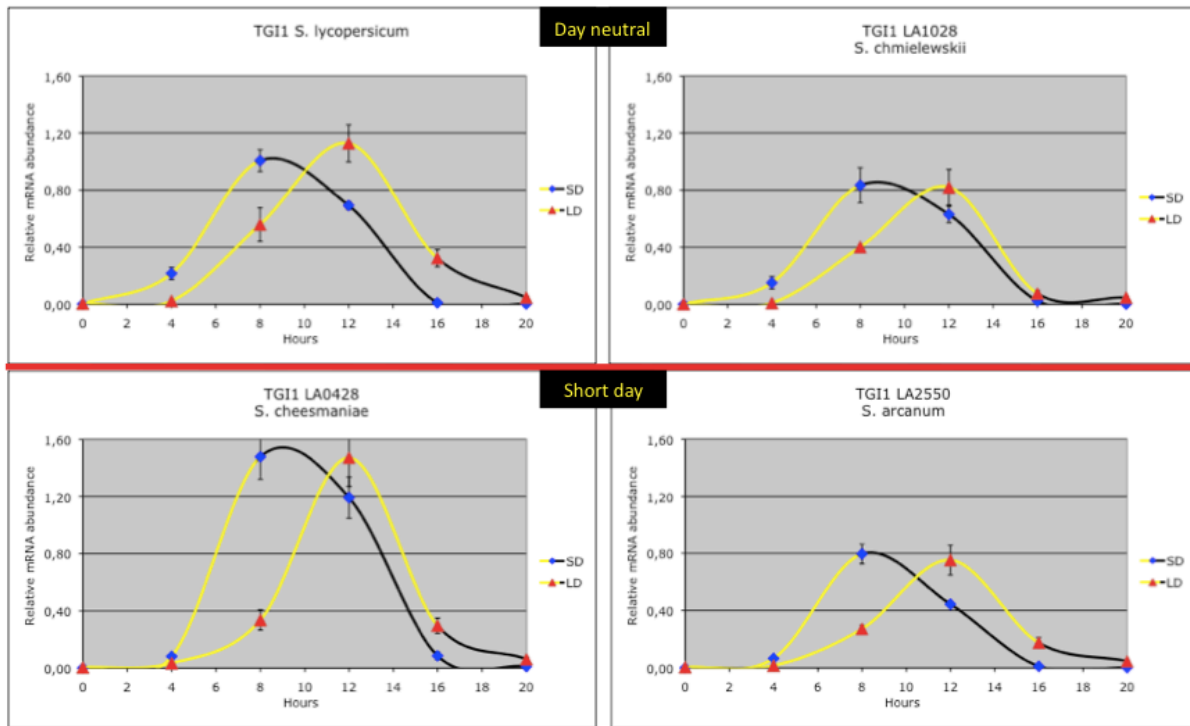
Figure 11. Diurnal oscillations of *TGI1* transcripts analyzed by QRT-PCR in two different photoperiodic condition: short day (SD: 8 hours of ligh and 16 hours of dark) and long day (LD: 16 hours of ligh and 8 hours of dark). Four species have been analysed: two day neutral species (*S. lycopersicum* and *S.chmielewskii*) and two short day species (*S. arcanum* and *S. cheesmaniae*). Results are expressed as relative mRNA abundance, after normalization with β-actin. Yellow and black lines represent light and dark periods, respectively. Time points are measured in hours from dawn. Data shown are the average of two biological replica and 3 technical replicates, with error bars representing standard deviation.

TGI1 was sequenced from cDNA and no promoter information is available, even if we do not expect alterations that can influence the expression of the gene in short day species.

3.4.5 TGI2

As for TGI1, the analyses of TG2 were conducted only on 6 species. TGI2 presents a lower degree of variability among these species, with respect to TGI1 (Fig.12 and Tab. 13) even if, like in TGI1, most predicted strong mutations are found in *S. chilense* and *S. habrochaites*.
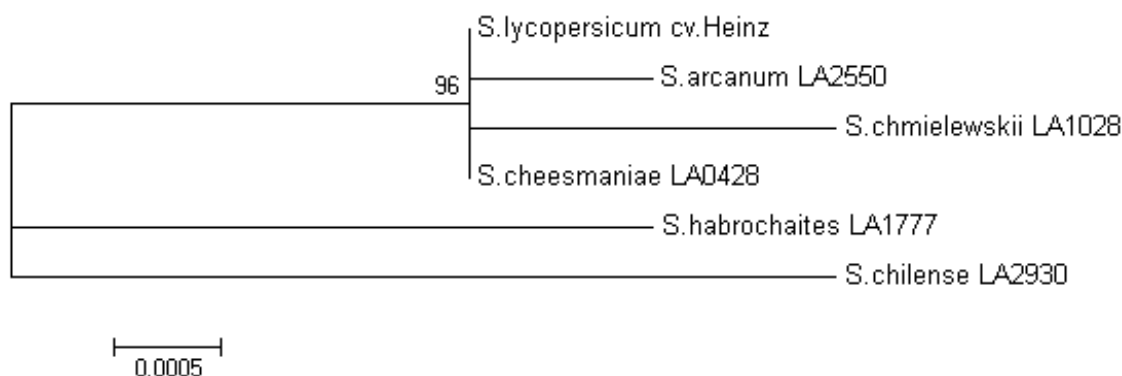
Figure 12. TGI2 protein dendrogram. The evolutionary history was inferred using the Neighbor-Joining method [64]. The optimal tree with the sum of branch length = 0.01156812 is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates) are shown next to the branches [65]. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. All positions containing alignment gaps and missing data were eliminated only in pairwise sequence comparisons (Pairwise deletion option). There were a total of 1167 positions in the final dataset. Phylogenetic analyses were conducted in MEGA4 [61].

Species	Accession	Latitude (degree)	Elevation (m)	Day length preference	dN	dS
S.cheesmaniae	LA0428	-0.63	700	Short day	0	0.002
S. arcanum	LA2550	-7.28	1700	Short day	0	0.005
S. habrochaites	LA1777	-9.55	3150	Short day	0.003	0.025
S.chmielewskii	LA1028	-13.55	2800	Day neutral	0.001	0.006
S.chilense	LA2930	-25.4	650	Day neutral	0.003	0.028

Table 12. Estimates of Codon-based Evolutionary Divergence between wild species and *S. lycopersicum* for *TGI2*. The number of dN and dS from analysis between sequences is shown . All results are based on the pairwise analysis of 6 sequences. Analyses were conducted using the Nei-Gojobori method in MEGA4 [66] [61]. All positions containing alignment gaps and missing data were eliminated only in pairwise sequence comparisons (Pairwise deletion option). Data are ordered by latitude value.

Position	Heinz	Subst.	SIFT value	Q. value	Conserved in Arabidopsis	Conserved in Rice	DAY NEUTRAL		SHORT DAY		
							LA1028	LA2930	LA0428	LA2550	LA1777
156	R	K	0.12	0.45	R		X				
174	G	D	0.26	0.41				X			
635	Q	L	0.00	0.63	Q	Q	X				
666	P	R	0.01	0.59				X			
668	Y	C	0.06	0.59							X
682	S	N	0.01	0.60						X	
834	K	R	0.19	0.60				X			
839	R	C	0.04	0.60				X			X
858	L	I	0.02	0.19				X			X
903	D	E	1.00	0.43	E	E		X			
903	D	S	0.00	0.43							X
1103	P	L	0.16	0.52				X			
1123	S	T	0.17	0.52						X	
1124	I	L	0.35	0.52						X	

a

Position	Heinz	Subst.	SIFT value	Q. value	Conserved in Arabidopsis	Conserved in Rice	-0.63	-7.28	-9.55	-13.55	-25.4
							LA0428	LA2550	LA1777	LA1028	LA2930
156	R	K	0.12	0.45	R					X	
174	G	D	0.26	0.41							X
635	Q	L	0.00	0.63	Q	Q				X	
666	P	R	0.01	0.59							X
668	Y	C	0.06	0.59					X		
682	S	N	0.01	0.60				X			
834	K	R	0.19	0.60							X
839	R	C	0.04	0.60					X		X
858	L	I	0.02	0.19					X		X
903	D	E	1.00	0.43	E	E					X
903	D	S	0.00	0.43					X		
1103	P	L	0.16	0.52							X
1123	S	T	0.17	0.52					X		
1124	I	L	0.35	0.52					X		

b

Table 13. SIFT values of substitution in TGI2. *S. lycopersicum* cv. Heinz has been used as reference sequence. SIFT values lower than 0.05 (red) indicate that the substitution affect the protein function. Values between 0.05 and 0.1 (orange) indicate possible deleterious substitutions. Values higher than 0.1 (black) indicate tolerated substituion. The prediction is based on sequence homology and the physical properties of amino acids. **a:** data grouped according to day length preference; **b:** data grouped according to latitude value.

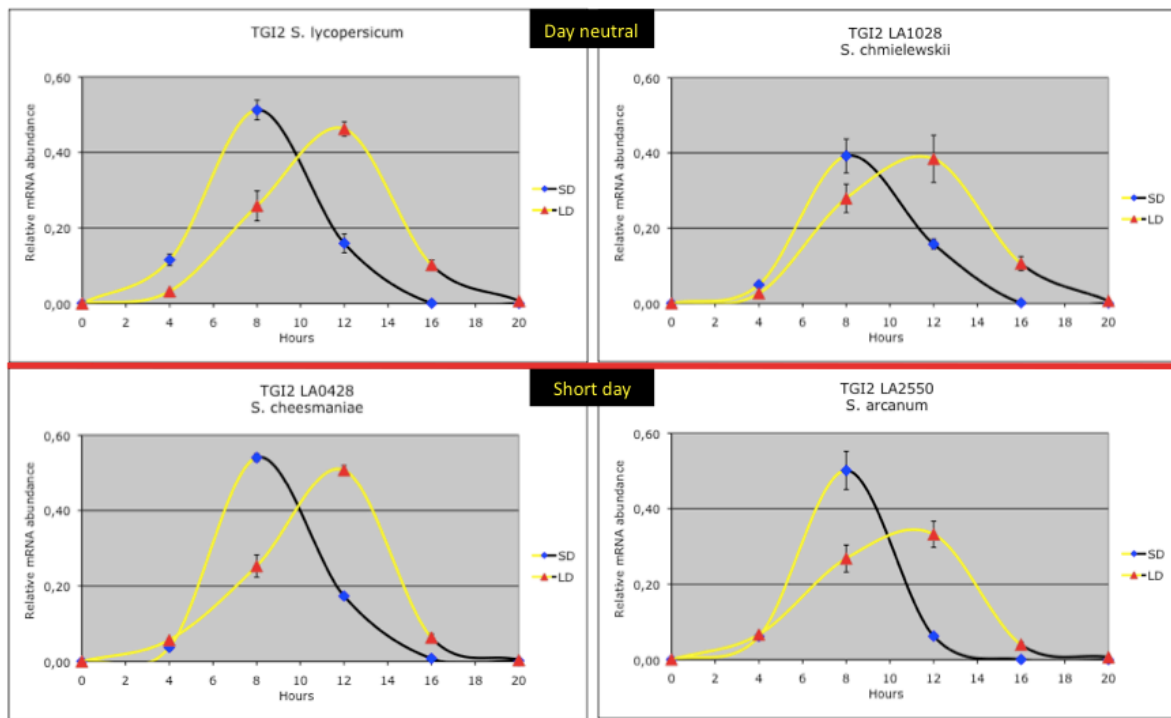Its expression pattern is very similar to *TGI1* (Fig. 13).

Figure 13. Diurnal oscillations of *TGI2* transcripts analyzed by QRT-PCR in two different photoperiodic condition: short day (SD: 8 hours of ligh and 16 hours of dark) and long day (LD: 16 hours of ligh and 8 hours of dark). Four species have been analysed: two day neutral species (*S. lycopersicum* and *S.chmielewskii*) and two short day species (*S. arcanum* and *S. cheesmaniae*). Results are expressed as relative mRNA abundance, after normalization with β-actin. Yellow and black lines represent light and dark periods, respectively. Time points are measured in hours from dawn. Data shown are the average of two biological replica and 3 technical replicates, with error bars representing standard deviation.

Like for *TGI1,* TGI2 was sequenced from cDNA and no promoter information is available.

3.4.6 TCOL1

TCOL1 has a high level of divergence between species, with a sum of branch lengths similar to cry2 (Fig. 14). The predicted strong mutations are mainly associated to latitudes close to the Equator (Tab. 15). Also for TCOL1, there is a divergence between the *S. arcanum* accessions, but this time both have strong mutations. In particular, *S. arcanum* LA2550 presents a very strong mutation in position 169 with respect to both *S. lycopersicum* and Arabidopsis (Tab. 15)
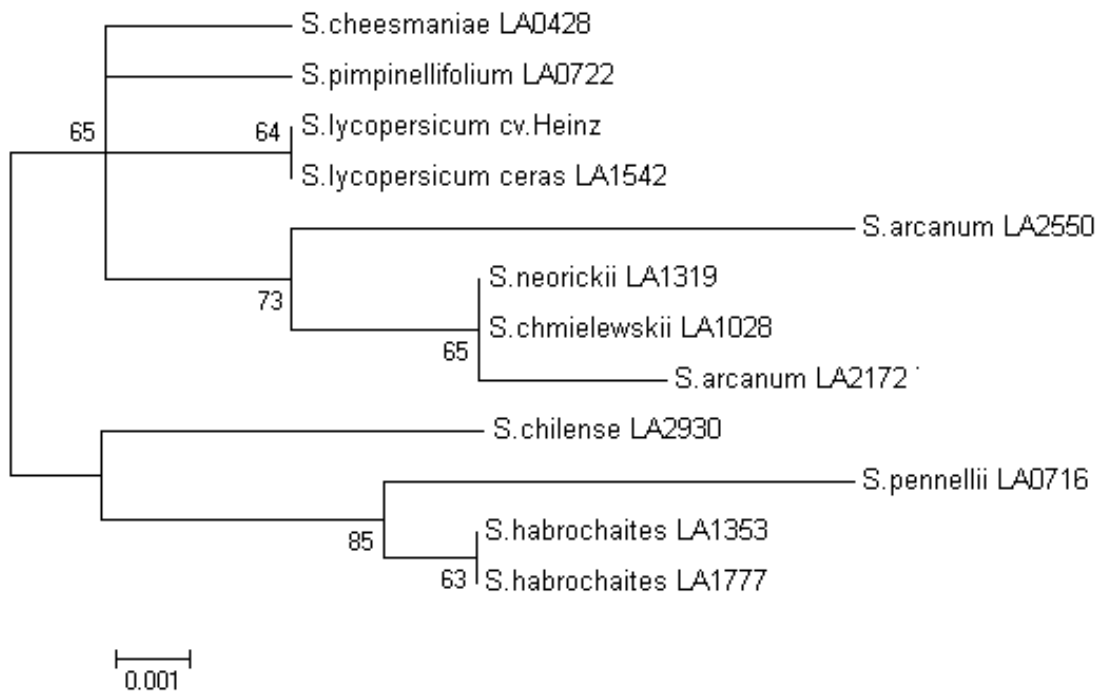
Figure 14. TCOL1 protein dendrogram. The evolutionary history was inferred using the Neighbor-Joining method [64]. The optimal tree with the sum of branch length = 0.04232120 is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates) are shown next to the branches [65]. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. All positions containing alignment gaps and missing data were eliminated only in pairwise sequence comparisons (Pairwise deletion option). There were a total of 391 positions in the final dataset. Phylogenetic analyses were conducted in MEGA4 [61].

Species	Accession	Latitude (degree)	Elevation (m)	Day length preference	dN	dS
S.lycopersicum v. cerasiforme	LA1542	9.9	700	Day neutral	0	0
S.cheesmaniae	LA0428	-0.63	700	Short day	0.002	0.008
S.arcanum	LA2172	-6.2	1000	Short day	0.004	0.019
S. arcanum	LA2550	-7.28	1700	Short day	0.006	0.028
S.habrochaites	LA1353	-7.36	2800	Short day	0.004	0.046
S.pimpinellifolium	LA0722	-8.11	50	Day neutral	0.002	0.011
S. habrochaites	LA1777	-9.55	3150	Short day	0.004	0.046
S.neorickii	LA1319	-13.38	2500	Day neutral	0.003	0.019
S.chmielewskii	LA1028	-13.55	2800	Day neutral	0.003	0.023
S.pennellii	LA0716	-16.21	100	Day neutral	0.006	0.05
S.chilense	LA2930	-25.4	650	Day neutral	0.003	0.023

Table 14. Estimates of Codon-based Evolutionary Divergence between wild species and *S. lycopersicum* for *TCOL1*. The number of dN and dS from analysis between sequences is shown . All results are based on the pairwise analysis of 12 sequences. Analyses were conducted using the Nei-Gojobori method in MEGA4 [66] [61]. All positions containing alignment gaps and missing data were eliminated only in pairwise sequence comparisons (Pairwise deletion option). Data are ordered by latitude value.

Position	Heinz	Subst.	SIFT value	Q. value	Conserved in Arabidopsis	Conserved in Rice	DAY NEUTRAL						SHORT DAY				
							LA 1542	LA 0722	LA 1028	LA 1319	LA 0716	LA 2930	LA 0428	LA 2172	LA 2550	LA 1353	LA 1777
8	N	I	0.06	0.51	N									X			
44		del										X					
46	L	F	0.01	0.96					X								
116	T	S	1.00	0.79							X	X			X	X
121	M	I	0.39	0.49											X		
169	N	I	0.00	0.89	N									X			
176	M	L	0.78	0.93	L	M				X	X				X		
194	G	E	0.01	0.93									X				
239	T	P	0.11	0.74							X					X	X
247	H	L	0.54	0.54						X	X			X	X		
251	Q	R	0.09	0.93	Q										X		
260	N	K	0.90	0.98									X				
293	N	I	0.93	1.00		N					X						
369	A	V	1.00	0.96	A						X						
380	M	S	0.00	0.84	M							X					
381-383		deletion									X						
384	N	S	0.46	0.84		S	X	X	X	X	X	X	X	X	X	X	X

a

Position	Heinz	Subst.	SIFT value	Q. value	Conserved in Arabidopsis	Conserved in Rice	9.9	-0.63	-6.2	-7.28	-7.36	-8.11	-9.55	-13.38	-13.55	-16.21	-25.4
							LA 1542	LA 0428	LA 2172	LA 2550	LA 1353	LA 0722	LA 1777	LA 1319	LA 1028	LA 0716	LA 2930
8	N	I	0.06	0.51	N					X							
44		del														X	
46	L	F	0.01	0.96								X					
116	T	S	1.00	0.79							X		X			X	X
121	M	I	0.39	0.49							X						
169	N	I	0.00	0.89	N					X							
176	M	L	0.78	0.93	L	M				X				X	X		
194	G	E	0.01	0.93					X								
239	T	P	0.11	0.74							X		X			X	
247	H	L	0.54	0.54						X	X			X	X		
251	Q	R	0.09	0.93	Q						X						
260	N	K	0.90	0.98						X							
293	N	I	0.93	1.00		N										X	
369	A	V	1.00	0.96	A											X	
380	M	S	0.00	0.84	M												X
381-383		deletion														X	
384	N	S	0.46	0.84		S	X	X	X	X	X	X	X	X	X	X	X

b

Table 15. SIFT values of substitution in TCOL1. *S. lycopersicum* cv. Heinz has been used as reference sequence. SIFT values lower than 0.05 (red) indicate that the substitution affect the protein function. Values between 0.05 and 0.1 (orange) indicate possible deleterious substitutions. Values higher than 0.1 (black) indicate tolerated substituion. The prediction is based on sequence homology and the physical properties of amino acids. **a:** data grouped according to day length preference; **b:** data grouped according to latitude value.

The expression of *TCOL1* is highly anticipated respect to *CRY* and *TGI* genes. The southernmost species (*S. chmielewskii* and *S. arcanum*) show an evident peak at ZT4 in LD conditions. In contrast, in *S. cheesmaniae* and *S. lycopersicum*, which are closer to the equator, this peak is less evident or absent (Fig. 15).
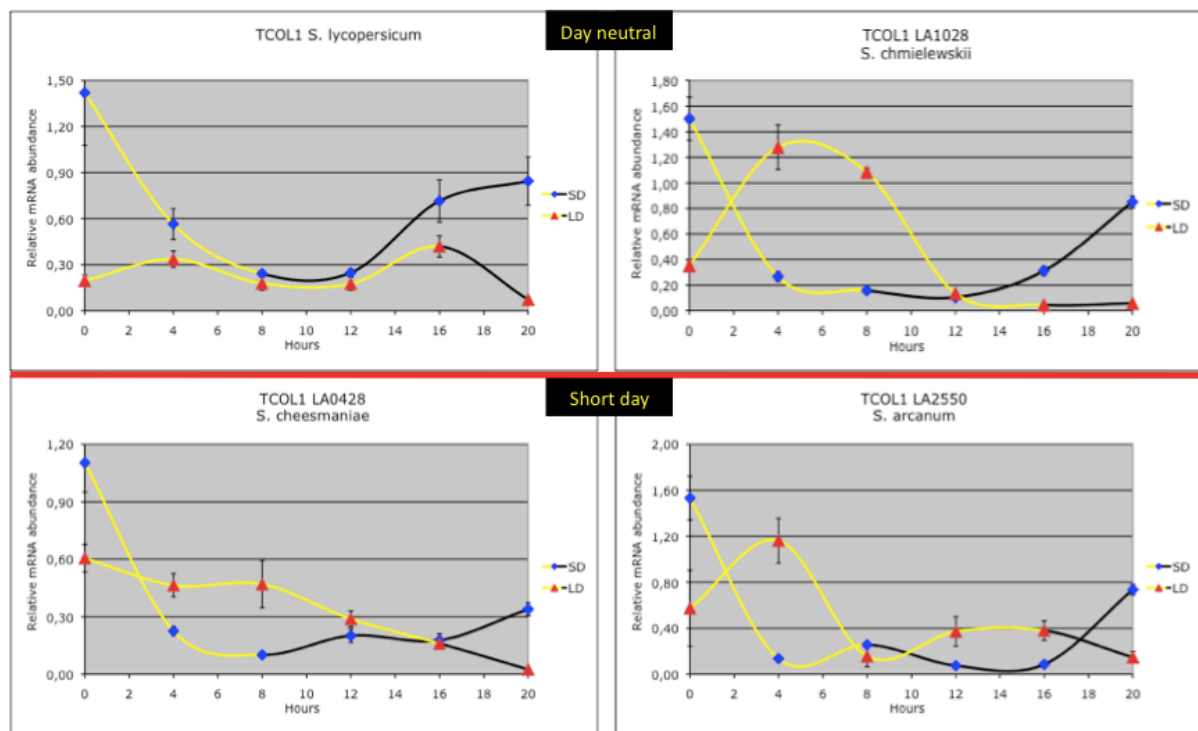

Figure 15. Diurnal oscillations of *TCOL1* transcripts analyzed by QRT-PCR in two different photoperiodic condition: short day (SD: 8 hours of ligh and 16 hours of dark) and long day (LD: 16 hours of ligh and 8 hours of dark). Four species have been analysed: two day neutral species (*S. lycopersicum* and *S.chmielewskii*) and two short day species (*S. arcanum* and *S. cheesmaniae*). Results are expressed as relative mRNA abundance, after normalization with β-actin. Yellow and black lines represent light and dark periods, respectively. Time points are measured in hours from dawn. Data shown are the average of two biological replica and 3 technical replicates, with error bars representing standard deviation.

For *TCOL1*, about 1700 bp upstream of the start codon were available and about 1600 were sequenced in the four species. As for *CRY3*, no informations about 5' UTR region are provided by iTAG. In *TCOL3,* the 5' UTR starts in position -315. Assuming the same size for the 5' UTR of TCOL1, about 1300 bp of putative promoter were analysed. 181 mutations (SNP and indels) have been identified among all the wild species respect to *S.lycopersicum.* Only 22 of them are present in *S. cheesmaniae*. For the most part, the other mutations are common to *S. arcanum* and *S. chmieleswkii*.

3.4.7 TCOL3

TCOL3 presents the highest degree of divergence between species among all the genes analysed. The sum of the branch lengths is about two times with respect to TCOL1 (Fig. 16) and the number of mutations is impressive, especially of those predicted as strong (Tab. 17).
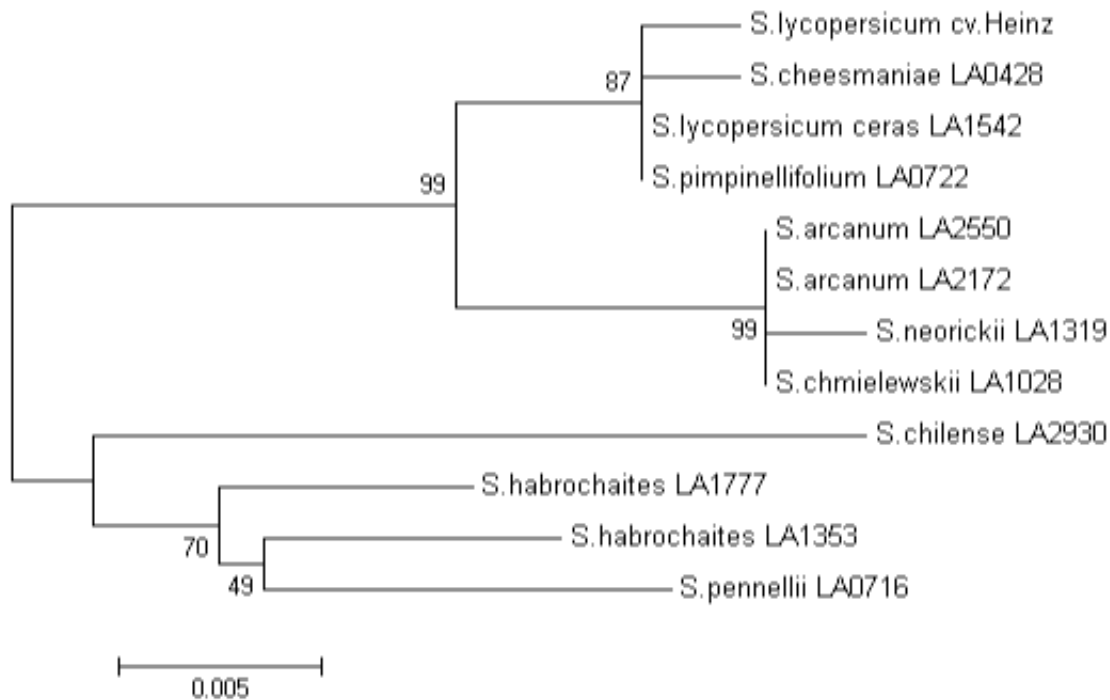
Figure 16. TCOL3 protein dendrogram. The evolutionary history was inferred using the Neighbor-Joining method [64]. The optimal tree with the sum of branch length = 0.07933115 is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates) are shown next to the branches [65]. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. All positions containing alignment gaps and missing data were eliminated only in pairwise sequence comparisons (Pairwise deletion option). There were a total of 409 positions in the final dataset. Phylogenetic analyses were conducted in MEGA4 [61].

Species	Accession	Latitude (degree)	Elevation (m)	Day length preference	dN	dS
S.lycopersicum v. cerasiforme	LA1542	9.9	700	Day neutral	0.001	0
S.cheesmaniae	LA0428	-0.63	700	Short day	0.002	0
S.arcanum	LA2172	-6.2	1000	Short day	0.006	0
S. arcanum	LA2550	-7.28	1700	Short day	0.006	0.004
S.habrochaites	LA1353	-7.36	2800	Short day	0.013	0.004
S.pimpinellifolium	LA0722	-8.11	50	Day neutral	0.001	0
S. habrochaites	LA1777	-9.55	3150	Short day	0.013	0.004
S.neorickii	LA1319	-13.38	2500	Day neutral	0.007	0
S.chmielewskii	LA1028	-13.55	2800	Day neutral	0.006	0.007
S.pennellii	LA0716	-16.21	100	Day neutral	0.014	0.007
S.chilense	LA2930	-25.4	650	Day neutral	0.017	0.019

Table 16. Estimates of Codon-based Evolutionary Divergence between wild species and *S. lycopersicum* for *TCOL3*. The number of dN and dS from analysis between sequences is shown . All results are based on the pairwise analysis of 12 sequences. Analyses were conducted using the Nei-Gojobori method in MEGA4 [66] [61]. All positions containing alignment gaps and missing data were eliminated only in pairwise sequence comparisons (Pairwise deletion option). Data are ordered by latitude value.

a: data grouped according to day length preference

Position	Heinz	Subst.	SIFT value	Q. value	Conserved in Arabidopsis	Conserved in Rice	DAY NEUTRAL						SHORT DAY				
							LA 1542	LA 0722	LA 1028	LA 1319	LA 0716	LA 2930	LA 0428	LA 2172	LA 2550	LA 1353	LA 1777
4	K	R	0.00	0.95								X					
7	S	R	0.02	0.95	S				X	X				X	X		
14	S	R	0.51	0.40							X						
20	V	M	0.04	0.95													X
46	A	T	0.05	1.00	A	A			X	X				X	X		
47	R	H	0.01	1.00		R					X					X	X
57	R	S	0.00	0.97	R	R					X						
108	T	P	1.00	0.73	T		X	X	X	X	X	X	X	X	X	X	X
114	Y	F	0.37	0.90	F											X	
124	G	S	0.60	0.44													X
129	I	N	0.02	0.47					X	X				X	X		
161-162		deletion									X	X				X	X
176	N	K	0.06	0.73	N	N						X					
182	V	F	0.06	0.84								X					
193	D	G	1.00	0.59	G	G					X					X	X
227	Q	P	0.06	0.93	Q	Q						X					
238	G	R	0.00	1.00	G								X				
256	Q	H	0.60	0.18					X	X	X	X		X	X	X	X
302	V	I	0.05	0.96	V	V					X						
314	Q	H	1.00	0.97	H											X	
327	G	V	0.10	1.00		G					X						
328	H	P	1.00	1.00	P	P					X	X					
329	P	T	0.07	1.00								X					
333	P	A	0.54	0.41								X					
334	L	P	0.00	0.40					X	X	X	X		X	X		X
383	T	M	0.02	0.92								X					
390	D	N	0.07	0.93		D					X	X				X	X
392	M	I	0.62	0.84		M					X	X				X	X
393	F	L	0.34	0.84	F	F					X	X				X	X
396	Q	H	0.03	0.85							X	X				X	X
397	L	L	1.00	0.85	L						X						
398-400		deletion									X						
399	T	A	0.59	0.82								X				X	X

b: data grouped according to latitude value

Position	Heinz	Subst.	SIFT value	Q. value	Conserved in Arabidopsis	Conserved in Rice	9.9	-0.63	-6.2	-7.28	-7.36	-8.11	-9.55	-13.38	-13.55	-16.21	-25.4
							LA 1542	LA 0428	LA 2172	LA 2550	LA 1353	LA 0722	LA 1777	LA 1319	LA 1028	LA 0716	LA 2930
4	K	R	0.00	0.95													X
7	S	R	0.02	0.95	S				X	X				X	X		
14	S	R	0.51	0.40												X	
20	V	M	0.04	0.95									X				
46	A	T	0.05	1.00	A	A			X	X				X	X		
47	R	H	0.01	1.00		R						X	X			X	
57	R	S	0.00	0.97	R	R								X			
108	T	P	1.00	0.73	T		X	X	X	X	X	X	X	X	X	X	X
114	Y	F	0.37	0.90	F						X						
124	G	S	0.60	0.44											X		
129	I	N	0.02	0.47					X	X				X	X		
161-162		deletion									X		X			X	X
176	N	K	0.06	0.73	N	N											X
182	V	F	0.06	0.84													X
193	D	G	1.00	0.59	G	G					X		X			X	
227	Q	P	0.06	0.93	Q	Q											X
238	G	R	0.00	1.00	G			X									
256	Q	H	0.60	0.18					X	X	X		X	X	X	X	X
302	V	I	0.05	0.96	V	V										X	
314	Q	H	1.00	0.97	H						X						
327	G	V	0.10	1.00		G										X	
328	H	P	1.00	1.00	P	P										X	X
329	P	T	0.07	1.00													X
333	P	A	0.54	0.41													X
334	L	P	0.00	0.40					X	X			X	X	X	X	X
383	T	M	0.02	0.92													X
390	D	N	0.07	0.93		D					X		X			X	X
392	M	I	0.62	0.84		M					X		X			X	X
393	F	L	0.34	0.84	F	F					X		X			X	X
396	Q	H	0.03	0.85							X		X			X	X
397	L	L	1.00	0.85	L											X	
398-400		deletion														X	
399	T	A	0.59	0.82							X		X				X

Table 17. SIFT values of substitution in *TCOL3*. *S. lycopersicum* cv. Heinz has been used as reference sequence. SIFT values lower than 0.05 (red) indicate that the substitution affect the protein function. Values between 0.05 and 0.1 (orange) indicate possible deleterious substitutions. Values higher than 0.1 (black) indicate tolerated substituion. The prediction is based on sequence homology and the physical properties of amino acids. **a:** data grouped according to day length preference; **b:** data grouped according to latitude value.

The expression pattern of *TCOL3* in long day conditions is similar among the four species analysed, with a peak at ZT12 to 16, which is particularly evident in the southernmost species (Fig. 17).
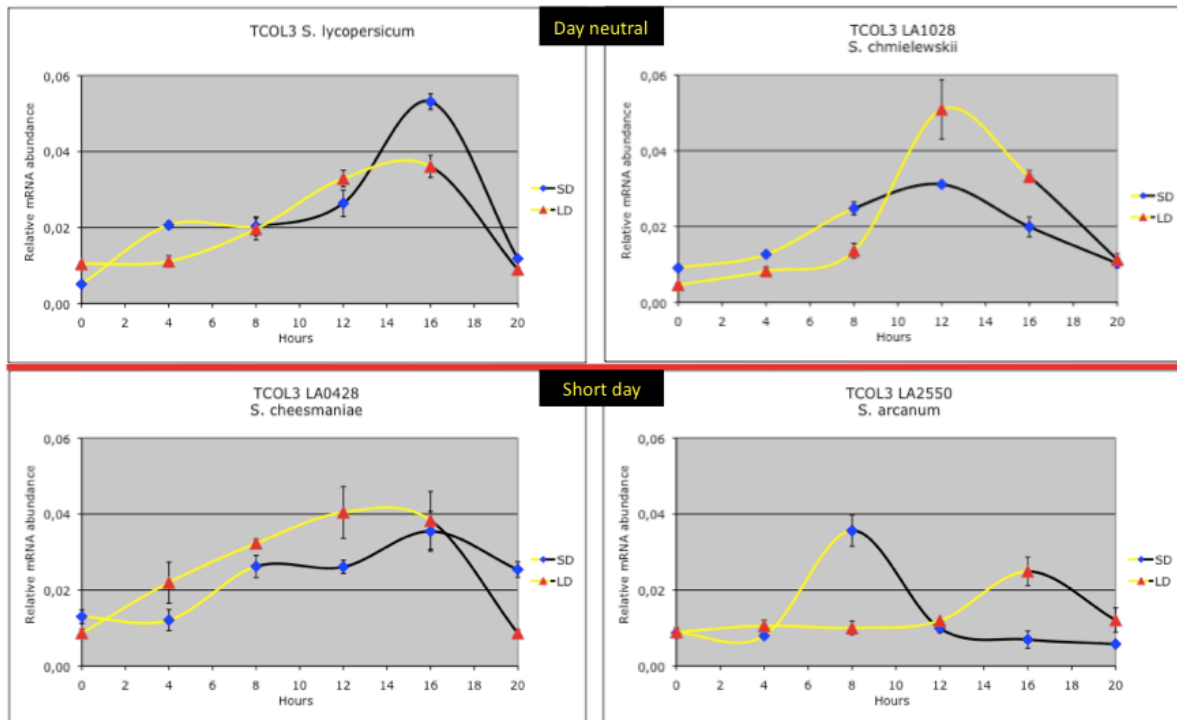

Figure 17. Diurnal oscillations of *TCOL3* transcripts analyzed by QRT-PCR in two different photoperiodic conditions: short day (SD: 8 hours of ligh and 16 hours of dark) and long day (LD: 16 hours of ligh and 8 hours of dark). Four species have been analysed: two day neutral species (*S. lycopersicum* and *S.chmielewskii*) and two short day species (*S. arcanum* and *S. cheesmaniae*). Results are expressed as relative mRNA abundance, after normalization with β-actin. Yellow and black lines represent light and dark periods, respectively. Time points are measured in hours from dawn. Data shown are the average of two biological replica and 3 technical replicates, with error bars representing standard deviation.

For *TCOL3*, only 200 bp before the start codon were available. Due to the presence of a cline in the dS values of this gene (see next paragraph), about 1500 bp upstream of the ATG were sequenced using a TAIL-PRC approach, in order to correlate dS values with mutations in the promoter. Despite the size of the sequenced region, only 26 mutations were identified in the three wild species with respect to *S. lycopersicum*. Only 3 of them are present in *S. cheesmaniae*. The other mutations are common to *S. arcanum* and *S. chmieleswkii*. 3 mutations are presents only in *S. arcanum*, but no correlation with transcription binding site motifs has been observed.

3.4.8 Gene functional divergence

dN/dS values were calculated for each gene over all species (Tab. 18). According to these values, all the genes except *TCOL3* are subjected to purifying selection, which indicate that amino acid mutations in these genes have a high probability of being deleterious. By contrast, *TCOL3* presents a high degree of diversifying selection, coupled with a slower evolution at silent positions.

Gene	dN	dN std. er.	dS	dS std. er.	dN/dS
CRY1a	0,002	0,001	0,018	0,004	0,111
CRY1b	0,003	0,001	0,023	0,004	0,130
CRY2	0,006	0,001	0,026	0,005	0,231
CRY3	0,004	0,001	0,019	0,004	0,211
TCOL1	0,004	0,001	0,026	0,006	0,154
TCOL3	**0,009**	0,002	**0,007**	0,002	**1,286**
TGI1	0,002	0,001	0,013	0,002	0,154
TGI2	0,002	0,001	0,017	0,003	0,118

Table 18. Estimates of Average Codon-based Evolutionary Divergence over all Sequence Pairs. The number of dN and dS and their ratio from averaging over all sequence pairs is shown. Results are based on the pairwise analysis of 12 sequences for *CRY* and *TCOL* genes, and of 6 sequences for *TGI* genes. Analyses were conducted using the Nei-Gojobori method in MEGA4 [66] [61]. All positions containing alignment gaps and missing data were eliminated only in pairwise sequence comparisons (Pairwise deletion option).

According to the SIFT data, a high number of strong mutations are present in TCOL3 in the different species (Tab 16 and 17). Probably these mutations are implicated in adaptation to different environments. Also, a geographical cline in the dS of the *TCOL3* gene is observed (Tab. 15 and Fig. 18). This means that the evolution at silent positions in this gene is controlled by latitude, and thus may explain its slow evolution.
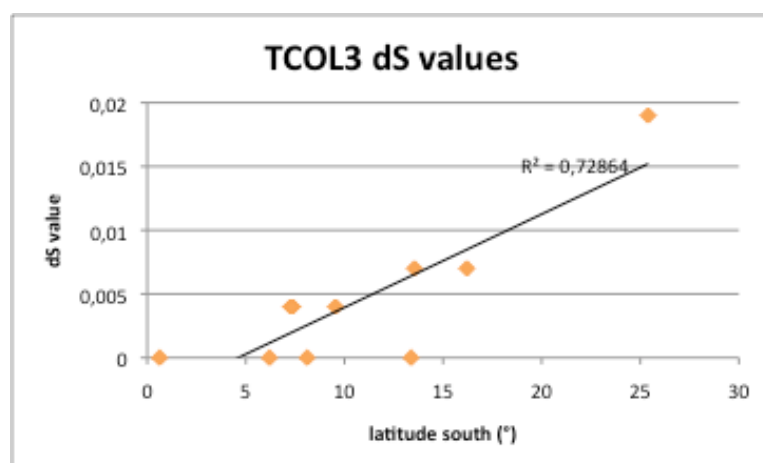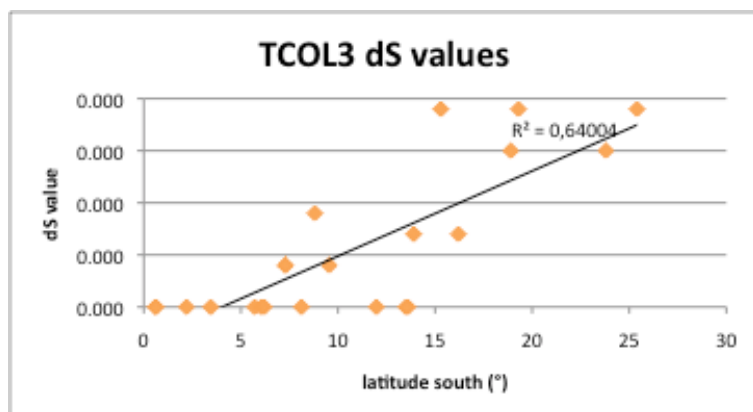

Figure 18. Cline of *TCOL3* dS values. dS values are ordered by latitude. 11 accessions are analysed respect to *S. lycopersicum.* The trend line has an R^2 value of 0.72864.

A cline describes a transition area in which a series of biocommunities display a continuous gradient of a quantitative trait [68]. In this case, the dS value shows a progressive increase (with respect to *S. lycopersicum*) with increasing distances from the Equator. Since dS measures silent mutations, this cline cannot be a consequence of a selection on the TCOL3 protein. We hypothesize that it cline is due to linkage disequilibrium with promoter mutations that are associated to adaptation to low latitudes. These mutations, in turn, are responsible for enhancing the expression peak in LD, at ZT12 to 16, which is more evident in southernmost species (Fig. 17). No such geographical correlation was evident for the other genes (Tab 1, 3, 5, 7, 9, 11, 13).

To confirm the cline data, we extended the analysis of dS to a higher number of accessions, selected to cover intermediate latitudes. 12 new accessions were selected, for a total of 24 (Tab. 19).

Species	Accession	Latitude (degree)	Elevation (m)	Day length preference	dN	dS
S.lycopersicum v. cerasiforme	LA1542	9.9	700	Day neutral	0.001	0,000
S.galapagense	LA0317	-0.6	15	Short day	0.003	0,000
S.cheesmaniae	LA0428	-0.63	700	Short day	0.002	0,000
S.habrochaites	LA1223	-2.2	2200	Short day	0.016	0,000
S.neorickii	LA2727	-3.45	2300	Day neutral	0.006	0,000
S.pimpinellifolium	LA2184	-5.71	450	Day neutral	0.002	0,000
S.lycopersicum v. cerasiforme	LA2845	-6.05	900	Day neutral	0	0,000
S.arcanum	LA2172	-6.2	1000	Short day	0.006	0,000
S.arcanum	LA2550	-7.28	1700	Short day	0.006	0.004
S.habrochaites	LA1353	-7.3	2800	Short day	0.013	0.004
S.pimpinellifolium	LA0722	-8.12	50	Day neutral	0.001	0,000
S.huaylasense	LA1982	-8.81	1400	?	0.029	0.009
S.habrochaites	LA1777	-9.55	3150	Short day	0.013	0.004
S.lycopersicum v. cerasiforme	LA1268	-11.98	900	Day neutral	0.001	0,000
S.lycopersicum v. cerasiforme	LA1320	-13.5	1900	Day neutral	0	0,000
S.neorickii	LA1319	-13.6	1800	Day neutral	0.007	0,000
S.chmielewskii	LA1028	-13.9	2800	Day neutral	0.006	0.007
S.chilense	LA1930	-15.3	700	Day neutral	0.022	0.019
S.pennellii	LA0716	-16.2	100	Day neutral	0.014	0.007
S.chilense	LA0458	-18.9	850	Day neutral	0.018	0.015
S.peruvianum	LA4125	-19.3	2550	Day neutral	0.019	0.019
S.chilense	LA2880	-23.8	2500	Day neutral	0.017	0.015
S.chilense	LA2930	-25.4	650	Day neutral	0.017	0.019

Table 19. Estimates of Codon-based Evolutionary Divergence between wild species and *S. lycopersicum* for *TCOL3*. The number of dN and dS from analysis between sequences is shown . All results are based on the pairwise analysis of 24 sequences. Analyses were conducted using the Nei-Gojobori method in MEGA4 [66] [61]. All positions containing alignment gaps and missing data were eliminated only in pairwise sequence comparisons (Pairwise deletion option). Data are ordered by latitude value.

The cline has been confirmed, with a stronger separation between northern and southern species (Fig 19).

Figure 19. Cline of *TCOL3* dS values. dS values are ordered by latitude. 23 accessions are analysed respect to *S. lycopersicum.* The trend line has an R^2 value of 0.64004.

In order to verify the hypothesis of linkage disequilibrium, we sequenced 1.5 Kb 5' to the start codon of *TCOL3*. According to cDNA sequence data, the 5' UTR is located from -315 to -92 bp with respect to the start codon. Only one mutation putatively connected with dS values has been identified in position -1111 (Fig. 20). Using "Jaspar Core" (http://jaspar.cgb.ki.se/cgi-bin/jaspar_db.pl), a eukaryotic transcription factor database, we identified a binding motif for the AP2 MBD-like family, which is altered by this mutation. In *Zea mays*, this motif is recognized by ABI4, a protein involved in the regulation of ABA- and sugar-regulated pathways [69].

Figure 20. Correlation between dS values and the mutation in the promoter site identified as an AP2 MBD-like binding motif (boxed).

The coding sequence data were used to identify mutations putatively connected with short day behaviour. All of them have been subjected to SIFT prediction (Tab. 2, 4, 6, 8, 10, 12, 14 ,16 and 17) but no candidate short day-specific mutation has been identified so far. *S. arcanum* is the short day species that presents the higher number of strong mutations and that shows variable degrees of alteration in the expression pattern of *CRY* and *TCOL* genes.

The *TCOL* genes are the most interesting, especially *TCOL3*, for their degree of geographical variation both in sequence and expression patterns: for *TCOL1* a geographical variation is observed only in the expression pattern, while *TCOL3* presents geographical variation in both expression patterns and primary sequence. This suggest that *TCOL3*, and perhaps also *TCOL1*, are involved in the adaptation to different latitudes. However, their expression differs: *TCOL1* expression seems to ne promoted by dark periods while *TCOL3* by light periods. A possible model is that the TCOL1/TCOL3 ratio is controlling flowering induction, or other photoperiodic responses. Further analyses have to be conducted in order to elucidate this phenomenon.

References:

1.   An, H., et al., *CONSTANS acts in the phloem to regulate a systemic signal that induces photoperiodic flowering of Arabidopsis.* Development, 2004. **131**(15): p. 3615-26.

2.   Putterill, J., et al., *The CONSTANS gene of Arabidopsis promotes flowering and encodes a protein showing similarities to zinc finger transcription factors.* Cell, 1995. **80**(6): p. 847-57.

3.   Suarez-Lopez, P., et al., *CONSTANS mediates between the circadian clock and the control of flowering in Arabidopsis.* Nature, 2001. **410**(6832): p. 1116-20.

4.   Yanovsky, M.J. and S.A. Kay, *Molecular basis of seasonal time measurement in Arabidopsis.* Nature, 2002. **419**(6904): p. 308-12.

5.   Imaizumi, T., et al., *FKF1 is essential for photoperiodic-specific light signalling in Arabidopsis.* Nature, 2003. **426**(6964): p. 302-6.

6.   Valverde, F., et al., *Photoreceptor regulation of CONSTANS protein in photoperiodic flowering.* Science, 2004. **303**(5660): p. 1003-6.

7.   Simon, R., M.I. Igeno, and G. Coupland, *Activation of floral meristem identity genes in Arabidopsis.* Nature, 1996. **384**(6604): p. 59-62.

8.   Martinez-Garcia, J.F., A. Virgos-Soler, and S. Prat, *Control of photoperiod-regulated tuberization in potato by the Arabidopsis flowering-time gene CONSTANS.* Proc Natl Acad Sci U S A, 2002. **99**(23): p. 15211-6.

9.   Borden, K.L., *RING domains: master builders of molecular scaffolds?* J Mol Biol, 2000. **295**(5): p. 1103-12.

10.   Strayer, C., et al., *Cloning of the Arabidopsis clock gene TOC1, an autoregulatory response regulator homolog.* Science, 2000. **289**(5480): p. 768-71.

11.   Griffiths, S., et al., *The evolution of CONSTANS-like gene families in barley, rice, and Arabidopsis.* Plant Physiol, 2003. **131**(4): p. 1855-67.

12.   Robson, F., et al., *Functional importance of conserved domains in the flowering-time gene*

CONSTANS demonstrated by analysis of mutant alleles and transgenic plants. Plant J, 2001. **28**(6): p. 619-31.

13.   Yano, M., et al., *Hd1, a major photoperiod sensitivity quantitative trait locus in rice, is closely related to the Arabidopsis flowering time gene CONSTANS.* Plant Cell, 2000. **12**(12): p. 2473-2484.

14.   Hayama, R. and G. Coupland, *The molecular basis of diversity in the photoperiodic flowering responses of Arabidopsis and rice.* Plant Physiol, 2004. **135**(2): p. 677-84.

15.   Kardailsky, I., et al., *Activation tagging of the floral inducer FT.* Science, 1999. **286**(5446): p. 1962-5.

16.   Kobayashi, Y., et al., *A pair of related genes with antagonistic roles in mediating flowering signals.* Science, 1999. **286**(5446): p. 1960-2.

17.   Kojima, S., et al., *Hd3a, a rice ortholog of the Arabidopsis FT gene, promotes transition to flowering downstream of Hd1 under short-day conditions.* Plant Cell Physiol, 2002. **43**(10): p. 1096-105.

18.   Samach, A. and G. Coupland, *Time measurement and the control of flowering in plants.* Bioessays, 2000. **22**(1): p. 38-47.

19.   Takada, S. and K. Goto, *Terminal flower2, an Arabidopsis homolog of heterochromatin protein1, counteracts the activation of flowering locus T by constans in the vascular tissues of leaves to regulate flowering time.* Plant Cell, 2003. **15**(12): p. 2856-65.

20.   Hayama, R., et al., *Adaptation of photoperiodic control pathways produces short-day flowering in rice.* Nature, 2003. **422**(6933): p. 719-22.

21.   Cheng, X.F. and Z.Y. Wang, *Overexpression of COL9, a CONSTANS-LIKE gene, delays flowering by reducing expression of CO and FT in Arabidopsis thaliana.* Plant J, 2005. **43**(5): p. 758-68.

22.   Hepworth, S.R., et al., *Antagonistic regulation of flowering-time gene SOC1 by CONSTANS and FLC via separate promoter motifs.* Embo J, 2002. **21**(16): p. 4327-37.

23.   Ben-Naim, O., et al., *The CCAAT binding factor can mediate interactions between CONSTANS-like proteins and DNA.* Plant J, 2006. **46**(3): p. 462-76.

24.   Lifschitz, E., et al., *The tomato FT ortholog triggers systemic signals that regulate growth and flowering and substitute for diverse environmental stimuli.* Proc Natl Acad Sci U S A, 2006. **103**(16): p. 6398-403.

25.   Sancar, A., *Structure and function of DNA photolyase and cryptochrome blue-light photoreceptors.* Chem Rev, 2003. **103**(6): p. 2203-37.

26.   Lin, C., et al., *Association of flavin adenine dinucleotide with the Arabidopsis blue light receptor CRY1.* Science, 1995. **269**(5226): p. 968-70.

27.   Malhotra, K., et al., *Putative blue-light photoreceptors from Arabidopsis thaliana and Sinapis alba with a high degree of sequence homology to DNA photolyase contain the two photolyase cofactors but lack DNA repair activity.* Biochemistry, 1995. **34**(20): p. 6892-9.

28.   Ahmad, M. and A.R. Cashmore, *HY4 gene of A. thaliana encodes a protein with characteristics of a blue-light photoreceptor.* Nature, 1993. **366**(6451): p. 162-6.

29.   Koornneef, M., E. Rolff, and C.J.P. Spruit, *Genetic control of light-inhibited hypocotyl elongation in Arabidopsis thaliana (L.).* Heynh. Z. Pflanzenphysiol. Bd., 1980. **100**: p. 147Ð160.

30.   Deng, X.W. and P.H. Quail, *Signalling in light-controlled development.* Semin. Cell Dev. Biol., 1999. **10**: p. 121Ð129.

31.   Ahmad, M., C. Lin, and A.R. Cashmore, *Mutations throughout an Arabidopsis blue-light photoreceptor impair blue-light-responsive anthocyanin accumulation and inhibition of hypocotyl elongation.* Plant J, 1995. **8**(5): p. 653-8.

32.   Lin, C., M. Ahmad, and A.R. Cashmore, *Arabidopsis cryptochrome 1 is a soluble protein mediating blue light-depend- ent regulation of plant growth and development.* Plant J., 1996. **10**: p. 893Ð 902.

33.   Lin, C., et al., *Enhancement of blue-light sensitivity of Arabidopsis seedlings by a blue light*

receptor cryptochrome 2. Proc Natl Acad Sci U S A, 1998. **95**(5): p. 2686-90.

34. Bagnall, D.J., R.W. King, and R.P. Hangarter, *Blue-light promotion of flowering is absent in hy4 mutants of Arabidopsis.* Planta, 1996. **200**: p. 278Ð280.

35. Guo, H., et al., *Regulation of flowering time by Arabidopsis photoreceptors.* Science, 1998. **279**(5355): p. 1360-3.

36. Koornneef, M., C.J. Hanhart, and J.H. van der Veen, *A genetic and physiological analysis of late flowering mutants in Arabidopsis thaliana.* Mol Gen Genet, 1991. **229**(1): p. 57-66.

37. Brudler, R., et al., *Identification of a new cryptochrome class. Structure, function, and evolution.* Mol Cell, 2003. **11**(1): p. 59-67.

38. Kleine, T., P. Lockhart, and A. Batschauer, *An Arabidopsis protein closely related to Synechocystis cryptochrome is targeted to organelles.* Plant J., 2003. **35**: p. 93Ð103.

39. Pokorny, R., et al., *Recognition and repair of UV lesions in loop structures of duplex DNA by DASH-type cryptochrome.* Proc Natl Acad Sci U S A, 2008. **105**(52): p. 21023-7.

40. Casal, J.J. and M.A. Mazzellam, *Conditional synergism between cryptochrome 1 and phytochrome B is shown by the analysis of phyA, phyB, and hy4 simple, double, and triple mutants in Arabidopsis.* Plant Physiol., 1998. **118**: p. 19Ð25.

41. Somers, D.E., P.F. Devlin, and S.A. Kay, *Phytochromes and cryptochromes in the entrainment of the Arabidopsis circadian clock.* Science, 1998a. **282**(5393): p. 1488-90.

42. Neff, M.M. and J. Chory, *Genetic interactions between phytochrome A, phytochrome B, and cryptochrome 1 during Arabidopsis development.* Plant Physiol., 1998. **118**: p. 27Ð35.

43. Mockler, T.C., et al., *Antagonistic actions of Arabidopsis cryptochromes and phytochrome B in the regulation of floral induction.* Development, 1999. **126**(10): p. 2073-82.

44. Hennig, L., et al., *Functional interaction of cryptochrome 1 and phytochrome D.* Plant J, 1999. **20**(3): p. 289-94.

45. El-Din El-Assal, S., et al., *A QTL for flowering time in Arabidopsis reveals a novel allele of CRY2.* Nat Genet, 2001. **29**(4): p. 435-40.

46.   El-Din El-Assal, S., et al., *The role of cryptochrome 2 in flowering in Arabidopsis.* Plant Physiol, 2003. **133**(4): p. 1504-16.

47.   Ninu, L., et al., *Cryptochrome 1 controls tomato development in response to blue light.* Plant J, 1999. **18**(5): p. 551-556.

48.   Perrotta, G., et al., *Tomato contains homologues of Arabidopsis cryptochromes 1 and 2.* Plant Mol Biol, 2000. **42**(5): p. 765-73.

49.   Weller, J.L., et al., *Genetic dissection of blue-light sensing in tomato using mutants deficient in cryptochrome 1 and phytochromes A, B1 and B2.* Plant J., 2001. **25**: p. 427Ð440.

50.   Giliberto, L., et al., *Manipulation of the blue light photoreceptor cryptochrome 2 in tomato affects vegetative development, flowering time, and fruit antioxidant content.* Plant Physiol, 2005. **137**(1): p. 199-208.

51.   Yanovsky, M.J. and S.A. Kay, *Living by the calendar: how plants know when to flower.* Nat Rev Mol Cell Biol, 2003. **4**(4): p. 265-75.

52.   Searle, I. and G. Coupland, *Induction of flowering by seasonal changes in photoperiod.* Embo J, 2004. **23**(6): p. 1217-22.

53.   Fowler, S., et al., *GIGANTEA: a circadian clock-controlled gene that regulates photoperiodic flowering in Arabidopsis and encodes a protein with several possible membrane-spanning domains.* Embo J, 1999. **18**(17): p. 4679-88.

54.   Park, D.H., et al., *Control of circadian rhythms and photoperiodic flowering by the Arabidopsis GIGANTEA gene.* Science, 1999. **285**(5433): p. 1579-82.

55.   Hayama, R., T. Izawa, and K. Shimamoto, *Isolation of rice genes possibly involved in the photoperiodic control of flowering by a fluorescent differential display method.* Plant Cell Physiol, 2002. **43**(5): p. 494-504.

56.   Somers, D.E., et al., *The short-period mutant, toc1-1, alters circadian clock regulation of multiple outputs throughout development in Arabidopsis thaliana.* Development, 1998b. **125**(3): p. 485-94.

57.   Mizoguchi, T., et al., *LHY and CCA1 are partially redundant genes required to maintain circadian rhythms in Arabidopsis.* Dev Cell, 2002. **2**(5): p. 629-41.

58.   Huq, E., J.M. Tepperman, and P.H. Quail, *GIGANTEA is a nuclear protein involved in phytochrome signaling in Arabidopsis.* Proc Natl Acad Sci U S A, 2000. **97**(17): p. 9789-94.

59.   Mizoguchi, T., et al., *Distinct roles of GIGANTEA in promoting flowering and regulating circadian rhythms in Arabidopsis.* Plant Cell, 2005. **17**(8): p. 2255-70.

60.   Jimenez-Gomez, J.M., et al., *Quantitative genetic analysis of flowering time in tomato.* Genome, 2007. **50**(3): p. 303-15.

61.   Tamura, K., et al., *MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0.* Mol Biol Evol, 2007. **24**(8): p. 1596-9.

62.   Peralta, I.E. and D.M. Spooner, *Granule-bound starch synthase (GBSSI) gene phylogeny of wild tomatoes (Solanum L. section Lycopersicon [Mill.] Wettst. subsection Lycopersicon).* Am J Bot, 2001. **88**: p. 1888Ð1902.

63.   Darwin, S.C., S. Knapp, and I.E. Peralta, *Taxonomy of tomatoes in the Galapagos Islands: native and introduced species of Solanum section Lycopersicon (Solanaceae).* Syst Biodiver, 2003. **1**: p. 29Ð53.

64.   Saitou, N. and M. Nei, *The neighbor-joining method: a new method for reconstructing phylogenetic trees.* Mol Biol Evol, 1987. **4**(4): p. 406-25.

65.   Felsenstein, J., *Confidence limits on phylogenies: An approach using the bootstrap.* Evolution, 1985. **39**: p. 783-791.

66.   Nei, M. and T. Gojobori, *Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions.* Mol Biol Evol, 1986. **3**(5): p. 418-26.

67.   Egea-Cortines, M., H. Saedler, and H. Sommer, *Ternary complex formation between the MADS-box proteins SQUAMOSA, DEFICIENS and GLOBOSA is involved in the control of floral architecture in Antirrhinum majus.* Embo J, 1999. **18**(19): p. 5370-9.

68.   Huxley, J., *Clines: an auxiliary method in taxonomy.* Bijdragen tot de Dierkunde (Leiden),

1938d. **27**: p. 491Ð520.

69.   Niu, X., T. Helentjaris, and N.J. Bate, *Maize ABI4 binds coupling element1 in abscisic acid and sugar response genes.* Plant Cell, 2002. **14**(10): p. 2565-75.

CHAPTER 4
TOOLS FOR FUNCTIONAL ANALYSIS OF GENES IN TOMATO SPECIES

4.1 Introduction

   4.1.1 Functional genomic and reverse genetic

   4.1.2 The EUSOL Project

   4.1.3 EUSOL WP 5.3: tools for functional analysis of genes in tomato and potato

   4.1.4 PTGS

   4.1.5 VIGS

   4.1.6 RNAi by ihpRNA

   4.1.7 Silencing vectors and Gateway technology

   4.1.8 TILLING

4.2 Aim of the work

4.3 Material and methods

   4.3.1 VIGS

   4.3.2 RNAi

   4.3.3 TILLING

4.4 Results and discussion

   4.4.1 VIGS

   4.4.2 RNAi

   4.4.3 TILLING

4.5 Preliminary characterization of 35S::*CRY1a* lines

4.1 Introduction

4.1.1 Functional genomics and reverse genetics

Functional genomics is a field of molecular biology whose objective is the use of the vast wealth of data produced by genomic projects (e.g. genome sequencing projects) to describe gene (and protein) functions and interactions. Reverse genetics is an approach of functional genomics whose objective is to discover the function of a gene by analysing the phenotypic effects of specific alterations of its DNA sequence. This approach advances in the opposite direction of so-called forward genetics, which seeks to discover the genetic basis of a phenotype or trait. Forward genetics has been helpful in the understanding of many biological processes and is an excellent strategy for identifying genes that function in a particular process.

Reverse genetics instead seeks to find what phenotypes arise as a result of perturbation of the function of particular genes. Using various techniques, gene function is altered and the effect on the development or behaviour of the organism is analyzed. Reverse genetics is useful, for example, in the functional investigation of traits controlled by gene families, something not very easy with forward genetics approaches [1]. In the past years, several reverse genetic approaches for plants have been developed [2] [3] [4] [5]. This, and the availability of complete sequenced genomes for many plant species [6] [7] [8], including tomato [9] has boosted reverse genetics approaches.

Tomato is an excellent model plant for genomic research of Solanaceous plants, as well as for studying the development, ripening, and metabolism of fruit. In 2003, the International Solanaceae Project (SOL, www.sgn.cornell.edu) was initiated by members from more than 30 countries, and the tomato genome sequencing project is now close to being complete with a pre-release of the genome in December 2009 (see chapter 2).

4.1.2 The EUSOL Project (http://www.eu-sol.net/)

Eu-Sol is a research project funded by the European Commission. It is focussed on improving the quality of the *Solanaceae* family, in particular potatoes and tomatoes. The project involves 56 partners from 15 different countries. The general aim of the project is to improve nutritional value, quality traits (e.g. taste, aroma, texture), productive traits (e.g. yield and resistance to biotic and abiotic stresses) and storage.

To achieve the above, there are a number of secondary aims EU-SOL has to fulfil:

1. To elucidate biochemical-molecular processes within tomato and potato that affect interesting agronomical traits;

2. To identify genes that affect these processes

3. To take advantage of the largely unexplored biodiversity in the tomato and potato by searching in wild relatives, alleles of these genes that could lead to improvements in the quality.

4. To test these genes in elite cultivars and verify if they confer the desired quality.

Module 5 of the project is concerned with developing structural and functional genomics technologies, such as genome sequence, microarrays, and tools for reverse genetics.

4.1.4 PTGS

RNA silencing is a gene regulatory mechanism that limits the transcript level by either suppressing transcription (transcriptional gene silencing, TGS) or by activating a sequence-specific RNA degradation process (post-transcriptional gene silencing, PTGS) [10]. This phenomenon was discovered by Fire et al. [11], who demonstrated the biochemical nature of inducers in gene silencing by introducing purified dsRNA directly into the body of *Caenorhabditis elegans*. The silencing a functional gene by exogenous application of dsRNA was termed RNA interference (RNAi). dsRNA is degraded into approximately 21 nucleotide RNAs, known as 'small interfering RNAs' (siRNAs), by the enzyme Dicer. These siRNAs then provide specificity to the

endonuclease-containing, RNA-induced silencing complex (RISC), which targets homologous host mRNA for degradation and the symptoms in the infected plant would reflect the loss of the function in the encoded protein [12-14] (Fig. 1).
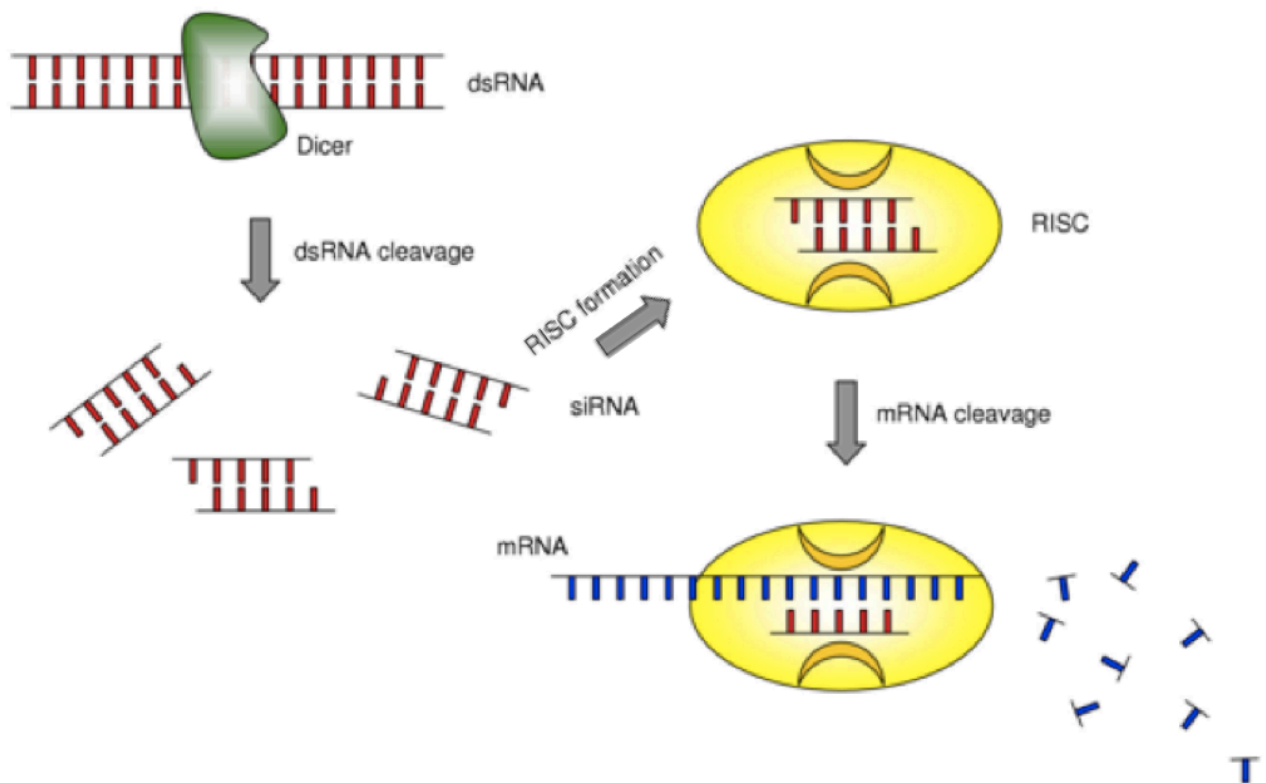
Figure 1. Mechanism of RNA interference (RNAi). The appearance of double stranded (ds) RNA within a cell (e.g. as a consequence of viral infection) triggers a complex response, which includes among other phenomena (e.g. interferon production and its consequences) a cascade of molecular events known as RNAi. During RNAi, the cellular enzyme Dicer binds to the dsRNA and cleaves it into short pieces of ~ 20 nucleotide pairs in length known as small interfering RNA (siRNA). These RNA pairs bind to the cellular enzyme called RNA-induced silencing complex (RISC) that uses one strand of the siRNA to bind to single stranded RNA molecules (i.e. mRNA) of complementary sequence. The nuclease activity of RISC then degrades the mRNA, thus silencing expression of the viral gene. [15]

Post-transcriptional gene silencing can be induced in plants by viral vectors harbouring specific gene sequences [16] [17] [18]. Several techniques exist by which this silencing can be obtained:

1. Virus-induced gene silencing (VIGS) [19] [20]. This technique has the advantage of being relatively fast and applicable also to plant species recalcitrant to transformation.

2. The use of inverted repeat transgenes producing hairpin transcripts (hpRNA) [21].

100

4.1.5 VIGS

In early studies, VIGS was studied in *Nicotiana benthamiana*, a wild tobacco species that is highly susceptible to virus infection. The first RNA virus successfully used as a silencing vector was Tobacco mosaic virus (TMV), for the silencing of the phytoene desaturase (PDS) gene [19]. A more stable vector is based on Potato virus X (PVX) [22], but it has a more limited host range than TMV, with only three plant families having members that are susceptible to PVX infection compared with nine families for TMV [23]. However, both TMV and PVX-based vectors cause disease symptoms on inoculated plants, sometimes making the interpretation of phenotypes difficult [24]. Furthermore, these viruses are excluded from the meristems of their hosts, which precludes effective silencing of genes in those tissues [25] [24]. These limitations (meristem exclusion and host range) were overcome with the development of VIGS vectors based on Tobacco rattle virus (TRV) [26] [27] [24], that is able to spread more vigorously throughout the entire plant, including meristem tissue, even if the overall symptoms of infection are mild compared with other viruses. The use of a duplicated 35S promoter and a ribozyme at the C-terminus permit to obtain a more efficient production of viral RNA, as well as a number of amino acid changes in the viral sequence itself [27].

4.1.6 RNAi by ihpRNA

To gain a stable and heritable gene silencing, transformation with hpRNA-expressing transgenes is necessary. hpRNA transgenes proved to be very effective for a wide range of target genes in various plant species [28]. Using intron-containing constructs (ihpRNA) containing sense/anti-sense arms ranging from 98 to 853 nt gave efficient silencing in a wide range of plant species, with generally 90-100% of independent transgenic plants showing silencing. The degree of silencing with these constructs was much greater than that obtained using either co-suppression or anti-sense constructs. Moreover, variable levels of gene silencing can be achieved in different transgenic lines using the same ihpRNA construct [29], allowing selection of lines with different degrees of silencing.

Furthermore, the expression of ihpRNAs from inducible promoters can control the growth stages and plant organs in which the gene silencing have to be activated [30] [31].

4.1.7 Silencing vectors and Gateway technology

VIGS and RNAi now have become be high-throughput methodologies, with the use of Invitrogen's Gateway recombination-based cloning technology [32] to replace previously time-limiting conventional cloning steps. This feature is implemented both in pHELLSGATE ihpRNA vectors [33] [34] and in TRV-based VIGS vector [26].

4.1.8 TILLING

Among several "reverse genetic" approaches developed, T-DNA technology is used with large success for the insertional mutagenesis in *A. thaliana*, and more recently in rice (http://signal.edu/cgi-bin/tdnaexpress – ahttp://signal.salk.edu/cgi-bin/RiceGE). However, this technique requires a vast population to have good chance to identify gene insertion s in specific genes with good probability. Second, it results in gene inactivation, which causes deleterious loss-of-function phenotypes. And finally, it is based on transgenosis, a fact that prevents the use of its products due to existing legislation.

In recent years, TILLING (Targeting Induced Local Lesions In Genomes) [35] [36], has appeared as a new emerging technology that doesn't rely on genetic transformation techniques and allows systematic functional genomic studies in any plant. TILLING is a non-transgenic reverse genetic strategy that utilises chemical mutagenesis for inducing point mutations and sensitive molecular screening to identify them [37]. This strategy generates allelic series of the targeted genes, including hypo- and hypermorphs as well as knockouts, which makes it possible to dissect the function of a protein as well as to investigate the role of essential genes that are otherwise not likely to be recovered in genetic screens based on insertional mutagenesis.

Until now, the TILLING technique has been validated by its successful application in plants: *Arabidopsis thaliana* [38] [39], wheat (*Triticum aestivum* L.) [40], barley (*Hordeum vulgare* L.) [41], rice (*Oryza sativa* L.) [42] [43], maize (*Zea mais* L.) [44], soybean (*Glycine max* L.) [45], pea (*Pisum sativum* L.) [46], sorghum (*Sorghum vulgare* Pers.) [47], melon (*Cucumis melo* L.) [48], tomato (*Solanum lycopersicon* L.) [49] [50], *Medicago truncatula* [51]. TILLING has been applied on animals too: zebrafish (*Danio rerio* H.) [52], *Drosophila melanogaster* [53] and *Caenorhabditis elegans* [54].

TILLING applies advances in molecular biology investigations to the identification of genetic variation at the level of a single base pair. TILLING consists of three main steps (Fig. 2):

1. Development of a mutagenized population.

2. Development of TILLING platform: DNA preparation and pooling.

3. Mutation discovery by molecular screening.

To obtain mutagenized plant populations, point mutations are induced by treating seeds with a chemical mutagen. The ideal mutagen for TILLING induces, with a high frequency in the genome, single nucleotide substitutions. The most common mutagen used to develop a TILLING population is ethylmethansulfonate (EMS) that generates a high density of point mutations, including nonsense and missense mutations [55] [56].
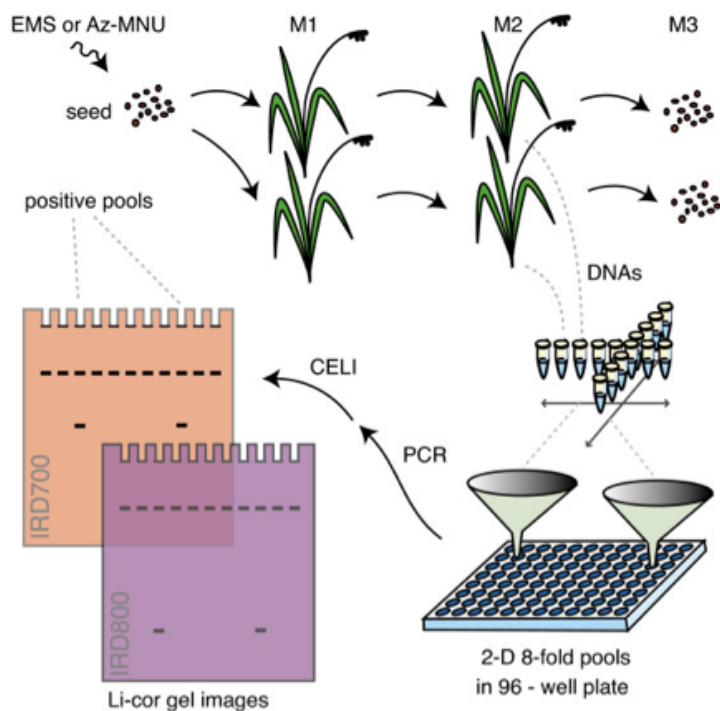
Figure 2. TILLING strategy for rice. After seed mutagenesis, chimaeric M1 plants are allowed to self-pollinate and a single M2 plant is grown to provide DNA for mutation discovery and seed for banking. DNAs are pooled eightfold and arrayed in a two-dimensional format on 96-well plates. After PCR amplification of target genes, heteroduplexes are formed upon heating and annealing, and then digested using crude celery juice extract containing the CEL I nuclease. Cut strands are separated by denaturing polyacrylamide gel electrophoresis, and visualized by fluorescence detection using a Li-Cor DNA analyzer. The presence of cut products in two pools identifies the individual harboring the polymorphism. [42]

4.2 Aim of the work

In order to carry out functional characterization of CRY genes, we applied the above reverse genetic techniques. This work is performed in the frame of EUSOL workpackage 5.3 (tools for functional analysis of genes in tomato and potato) and it consisted in the generation and comparison of RNAi, VIGS and TILLING lines for *CRY1a, CRY1b* and *CRY2*. The functions of these genes have been described in chapter 3. The project has both technical and scientific objectives, including the comparison of the three reverse genetic techniques, in order to identify the most fast and efficient reverse genetics method, and the characterization of the function of the three *CRY* genes in tomato.

4.3 Materials and methods

4.3.1 Gateway Technology for cloning in TRV and pHELLSGATE12

The same silencing fragments were cloned in TRV VIGS-vector [27] and pHELLSGATE12 hpRNA- vector, using the Gateway technology from INVITROGEN. The method is described in the Gateway manual: http://tools.invitrogen.com/content/sfs/manuals/gatewayman.pdf For silencing of the *CRY* genes, a region in the 3' UTR was chosen presenting minimal sequence homology between the three genes, in order to avoid cross-silencing. The silencing fragments were produced with PCR amplification from cDNA using specific attB primer pairs:

CRY1a:

forward: GGGGACAAGTTTGTACAAAAAAGCAGGCTTCGGACCAATATGTTGGTGA

reverse: GGGGACCACTTTGTACAAGAAAGCTGGGTCTAATTAGTTCCAGGACTTC

fragment size: 512 bp.

CRY1b:

forward: GGGGACAAGTTTGTACAAAAAAGCAGGCTAATATCGATGTAATGCAAGAA

reverse: GGGGACCACTTTGTACAAGAAAGCTGGGTCACACACAGGTTAAGAAAG;

fragment size: 207 bp.

CRY2:

forward: GGGGACAAGTTTGTACAAAAAAGCAGGCTGTGAATACTTCACATG

reverse: GGGGACCACTTTGTACAAGAAAGCTGGGTGAGGGAATCATCGCA

fragment size: 553 bp.

After the first Gateway cloning step (BP reaction), the entry clone is used in the LR reaction in combination with the destination vector: TRV2 for VIGS and pHELLSGATE12 for hpRNA

4.3.2 VIGS

The VIGS protocol was carried out according to [57], with some modification:

A. tumefaciens growth

Single colonies of *A. tumefaciens* strain GV3101 containing each TRV derivative were screened via PCR specific primer pairs in order to verify the presence of the silencing vectors:

TRVRNA1 for: GAAAATATTGCTGCGCCTAACG

TRVRNA1 rev: ACCTGCCACGGTTCGAAGTA

TRVRNA2 up1 TACCAAGGCGAACACTGG

TRVRNA2 dw1 TAACGTCATGCATTACATG

A 1 ml culture of *A. tumefaciens* strain GV3101 containing each TRV derivative was grown 8-10 h at 28°C in Luria-Bertani (LB) medium containing 50 μg ml-1 kanamycin and 100 μg ml-1 rifampicin. Cultures containing *A. tumefaciens* with pTRV2 were transferred into 40 ml of LB with antibiotics and the same for A. tumefaciens pTRV1 strain. Then they were shaken overnight at 28°C. *A. tumefaciens* cells were pelleted, washed, re-suspended in infiltration buffer to a final OD600 of 2.0 (pTRV1) or 2.0 (pTRV2), and shaken slowly for 3 h or overnight at room temperature, prior to infiltration.

Infiltration buffer: (10 mM MgCl2, 10 mM MES (pH 5.6), 200 μM acetosyringone)

Infiltration

Just prior to infiltrating the suspension, Silwet L-77 was added to the suspension to a final concentration of 0.04%. Plants were infected with both TRV1 and TRV2 when the first pair of leaves had emerged, usually at 2-3 weeks after sowing. The plant was completely submerged into the Agrobacterium suspension by turning it upside down into a plastic plate placed in a vacuum dessicator, and a vacuum was pulled for 1 min and 30 sec to a maximum of 29,5 Hg and slowly

released. Plants were placed in the dark overnight, then in photoperiodic conditions of 16 hours oh light and 8 hours of dark, with constant temperature of 18°C.

Molecular screening

In order to verify the silencing efficiency of the *CRY* genes, RNA isolation was performed (according to [58]) four weeks after infection. Real-time RT-PCR was used to evaluate transcript abundance (see protocol in chapter 3) with the following primers:

CRY1a:

cry1 rt up2: GTCGTGAGCTTGATCGGATTG

cry1 rt dw2: CGCCGGACATATTCCCC

CRY1b:

cry1b rt up2: ACCGGTACCCTACCTGATGGTT

cry1b rt dw2: AAGCCATCGTCGAACATATTCC

CRY2:

cry2 rt up3: GAGCGCCTCGATAACCCA

cry2 rt dw3: GCCCTCCGGATCATAATTGA

4.3.3 RNAi

Plant transformation

Cotyledons of *S. lycopersicum* cv. Moneymaker have been used for transformation with *Agrobacterium tumefaciens* as vector, according to [59].

In order to isolate independent transformation events, only one regenerant has been selected from each cocultivated tomato explant. Putative regenerants were checked by PCR using two transgene-

specific primer pairs, one for the first Gateway box, with the fragment cloned in sense orientation, and one for the second Gateway box, with the fragment cloned in antisense orientation:

I box:

PH12-4UP: ATTCCATTGCCCAGCTAT

PH12-4DW: AATTCCTCGAGACCACTT

II box:

PH12-2UP1: GATTGATTACAGTTGGGAA

PH12-2DW1: AGGATCTGAGCTACACATG

In vivo culture and climatic conditions

PCR-positive, rooted tomato plantlets were adapted in greenhouse in pots (diameter: 25 cm) in a Vigorplant TN soil. Photoperiod was set at 16 hours of light and 8 hours of dark, with temperature set at 25°C during the light period and at 20°C during the dark period. During fruit production, watering was reduced.

Seed harvesting

Seeds are harvested from mature fruit (10-15 days after breaker stage) and treated with 100 mM HCL for 20 minutes, then washed extensively under tap water and dried overnight on filter paper. They are stored at room temperature with silica gel in sealed tubes.

Blue light screening

Germination and plantlet growth was performed in Magenta boxes (SIGMA), placed in growth chambers with 16 hours of blue light and 8 hours of dark, with temperature set at 25°C during the light period and at 20°C during the dark period.

Blue light source: Osram 67 lamp with Lee Dark Blue N° 119 filter.

Two different light conditions have been tested:

High irradiance (15 μE m^{-2} s^{-1}) by closing Magenta with its normal lid;

Low irradiance (2 μE m^{-2} s^{-1}) by adding one layer of Lee Neutral Density N° 211 filter.

Hypocotyl length was measured after 7 days.

4.3.4 TILLING

M82 EMS population

A population of *S. lycopersicum* cv. M82 was mutagenized by EMS at the Hebrew University of Jerusalem by Prof. Dani Zamir [60]. M82 EMS collection was obtained by treating seed by 0.5% of EMS, corresponding to LD15. After treating seeds by EMS, M1 and M2 generations were produced and, at the end, a total of 6000 M2 families were phenotyped and organized in 15 "Classes" and 48 "Subclasses" in the SGN database (http://zamir.sgn.cornell.edu/mutants/). A TILLING platform was generated from M3 4759 families, for each family, DNAs were extracted from six independent plants. The platform is organized in 7 bulk plates that derive from 50 independent DNA plates. Each bulk plate contains 96 pools of 8 samples, for a total of 768 samples [61].
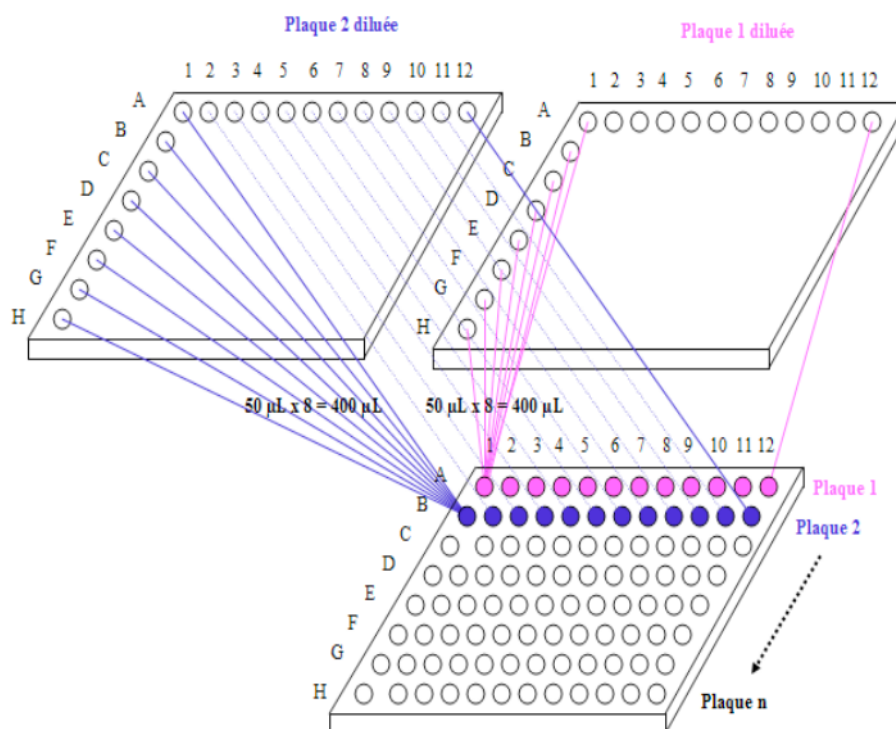
Figure 3. TILLING platform construction. Bulk plates are generated from the pooling of individual plates.

TILLING analysis

Molecular analyses were carried out according to [36], with some modifications according to [46]:

Nested PCR

PCR amplification was based on nested-PCR.

The first PCR amplification was a standard PCR reaction using target-specific primers and pooled DNA from bulk plates:

25 ng of DNA,

1x PCR Buffer,

0.2 mM of dNTPs,

0.4 μM of forward standard primer,

0.4 μM of reverse standard primer,

1 U of Taq polimerase.

Thermocycler conditions:

94°C x 2 min

30 cycles: 94 °C x 15 sec, Ta °C x 20 sec, 72 °C x 1 min.

72°C x 5 min

Ta (annealing temperature) is primer pair-specific.

Amplification products were separated in 2% TAE agarose gel with 0,5 μg/ml of ethidium bromide, using 100 bp DNA ladder (Biolabs) as molecular weight marker. The second PCR amplification was a standard PCR reaction using one microliter of the first PCR and target-specific inner primers carrying a universal labelled at the 5'end with infra-red dyes IRD700 and IRD800 (LI-COR®, Lincoln, NE, USA):

1 ul of the first PCR

1x PCR Buffer,

0.2 mM di dNTPs,

0.4 μM forward IRD700 inner primer,

0.4 μM reverse IRD800 inner primer,

1 U of Taq polimerase.

Thermocycler conditions:

94°C x 2 min

35 cycles: 94 °C x 15 sec, Ta °C x 20 sec, 72 °C x 1 min.

72°C x 5 min

Ta (annealing temperature) is primer pair specific.

Amplification products were separated in 2% TAE agarose gel with 0,5 μg/ml of ethidium bromide, using 100 bp DNA ladder (Biolabs) as molecular weight marker. Gel imaging was performed with "Quantity One Quantitation Software" v. 4.2.

Heteroduplex formation

After the second PCR, samples were subjected to denaturation and renaturation steps, in order to produce mismatch (heteroduplex), in the polymorphic sites between mutated and wild type DNA:

94 °C x 10 min followed by slow temperature decreasing from 94 °C to 8 °C (-0.1 °C/sec)

Endonuclease digestion

Endonuclease ENDO1 [62] recognizes mismatch sites present in the amplicons and produces a cleavage. This cleavage causes, after denaturation, to the formation of four fragments, two of which are labeled with fluorochromes: IRD700 for the "forward" fragment and IRD800 for the "reverse" fragment.

Digestion was performed with:

60 ng of PCR product,

3 μl of digestion buffer 10x,

3 μl of ENDO1 endonuclease (diluted 1:5000 with elution buffer),

final volume: 30 μl.

Incubation: 45 °C x 20 min followed by addition of 5 μl of EDTA 0.15 M and ice storage.

Digestion buffer 10X: 0.1 M Hepes pH 7.5; 0.1 M $MgSO_4$; 0.02% Triton X-100; 0.1 M Kcl.

Elution buffer: 50 mM Tris HCl pH8; 100mM KCl; 10 μM $ZnCl_2$;0.01% Triton X-100; 50% glycerol; 20 μg/ml acetylated BSA)

Purification

Digestion products were purified on Sephadex G50 columns (Sephadex[TM] G-50 Superfine, Amersham Biosciences) using Millipore 96 wells filtration plate (MultiScreen-HV, MAHVN4550).

Purified samples were dried at 65 °C for 1 hour and resuspended in 5 ul of formamide loading dye.

Formamide loading dye:

94% Formamide,

10mM EDTA pH8,

0.05% Xilene cyanol,

0,05% Bromophenol blue.

Denaturing polyacrylamide gel preparation

25 cm polyacrylamide gel for "Li-Cor 4300 fragment analyser" were prepared with:

25 ml of 6,5 % KB plus Gel Matrix (LICOR),

280 μl of APS 10%

28 μl TEMED.

<u>Electrophoretic run and fragment analysis</u>

Before loading samples in the polyacrylamide gel, they were denaturated at 93 °C x 3 min. Electrophoretic runs were performed in a "Li-Cor 4300 fragment analyser" using TBE 1X buffer and the following run conditions: 1500V, 40W, 40mA, and 45°C.

<u>Sequencing</u>

Positive samples were subjected to Sanger sequencing in order to identify the exact mutation site and the type of mutation.

<u>*In vivo* culture and growth conditions, seed harvesting and blue light screening</u>

see paragraph *4.3.3.*

4.4 Results and discussion

4.4.1 VIGS

The gene silencing based on plants agro-infection with TRV (bipartite positive sense RNA virus) has been tested using the phytoene desaturase (*PDS*) gene, in order to optimize the method with a visible phenotype. When *PDS* is silenced by VIGS, inhibition of carotenoid biosynthesis occurs and leaves are characterized by photo-bleached regions. The silencing efficiency can be correlated with the ratio of photo-bleached/green area of the leaf. We adopted the vacuum infiltration method with tomato cv. Moneymaker plantlets at the stage of two true leaves. Four weeks after infiltration, plants infected with TRV/*PDS* show a strong photo-bleached phenotype (Fig. 4).

Figure 4. Photo-bleached phenotype in tomato plantlets two weeks after infiltration with TRV/*PDS*.

Furthermore, we assessed the efficiency of the method in fruits of the tomato MicroTom cultivar, in which the efficient inhibition of carotenoid biosynthesis produces yellow fruits, instead of red. The carpopodium of 1-2 day-old fruits was infected with a needless syringe and we obtained efficient silencing (Fig. 5).

Figure 5. Inhibition of lycopene accumulation occurs in fruits infected with TRV/*PDS*.

The method was then tested with TRV/*CRY1a*, TRV/*CRY1b* and TRV/*CRY2* vectors. Inconsistent results were obtained with wild type plants, with only a weak increase in internode length and no alteration in leaf pigmentation (Fig. 6), while the expected phenotypes are longer internodes and less pigmented leaves (see below).
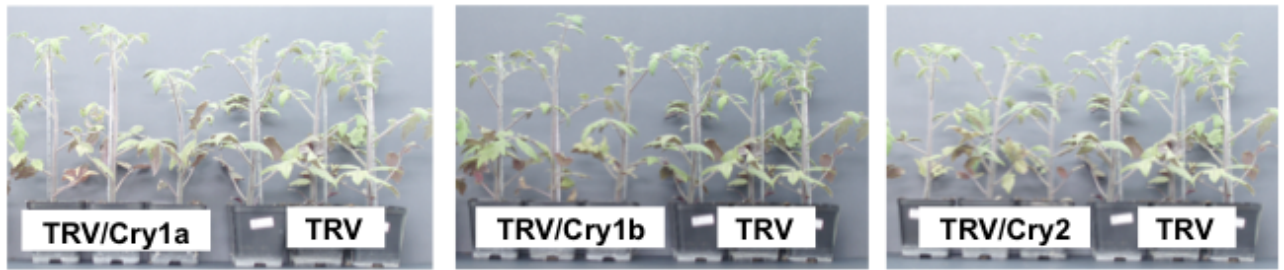
Figure 6. Phenotypes of plants length infected with TRV/*CRY1a*, TRV/*CRY1b* and TRV/*CRY2*, compared with those infected with empty TRV.

.

To test the functionality of our TRV/*CRY* constructs, we infected transgenic plants overexpressing *CRY2* [63] with TRV/*CRY2*. The internode shortening induced by *CRY2* overexpression was reversed (Fig 7), in agreement with previous literature on the silencing of transgenes [63] [20].
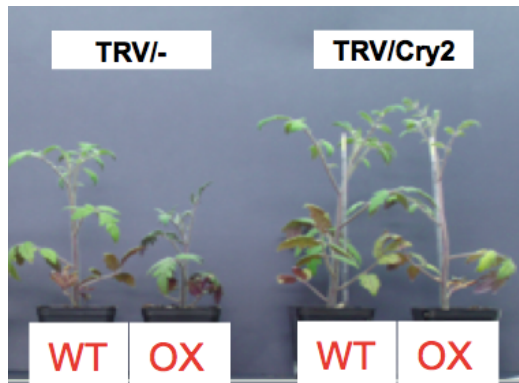
Figure 7. VIGS silencing mediated by TRV/*CRY2* vector reverts the short internode phenotype of *35S::CRY2* palnts.

The expression levels of the endogenous *CRY* genes were measured in leaf tissues collected from TRV/*CRY*-infected plants, four weeks after the infection, when photo-bleached phenotypes in TRV/*PDS*-infected plants became evident. TRV/- plants have been used as control. The data indicate a 30 to 50% reduction in the mRNA levels of the three genes (Fig 8).
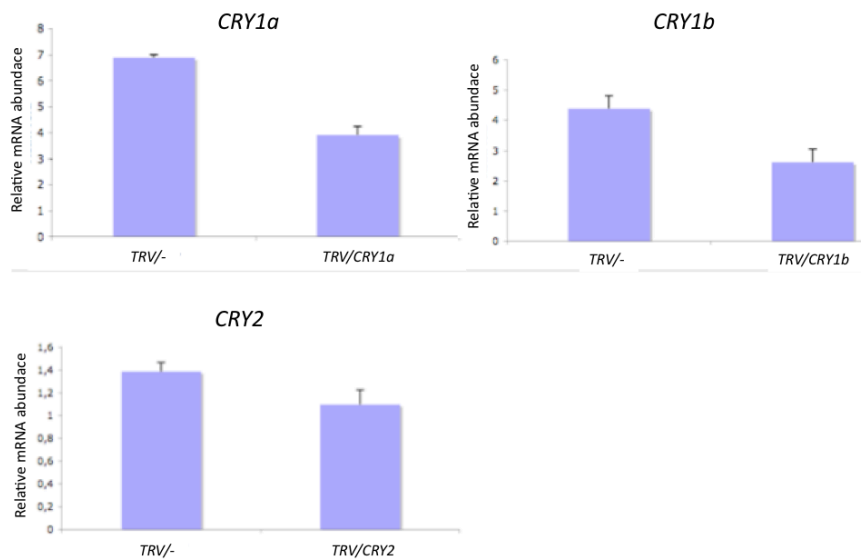
Figure 8. Relative mRNA abundance, after normalization with β-actin, of *CRY1a*, *CRY1b* and *CRY2* genes in plants infected respectively with TRV/*CRY1a*, TRV/*CRY1b*, TRV/*CRY2*. TRV/- has been used as control. Data shown are the average of two biological and three technical replicas, with error bars representing standard deviation.

Our main conclusion is that VIGS shows best results in the silencing of genes with a localized phenotype, like pigment accumulation, while its applicability decreases in the silencing of genes with systemic effects, like Cryptochromes.

4.4.2 RNAi

RNAi lines have been obtained for the three candidate genes. pHg-12_*CRY1b* and pHg-12_*CRY2* T1 lines have been subjected to screening under blue light (Fig 9). T1 indicates the progeny of transformed plants (T0). Three different RNAi lines have been analysed for each gene, both in high and low irradiance. Wild type, *cry1a* mutant and overexpressors of the three genes have been used as controls.
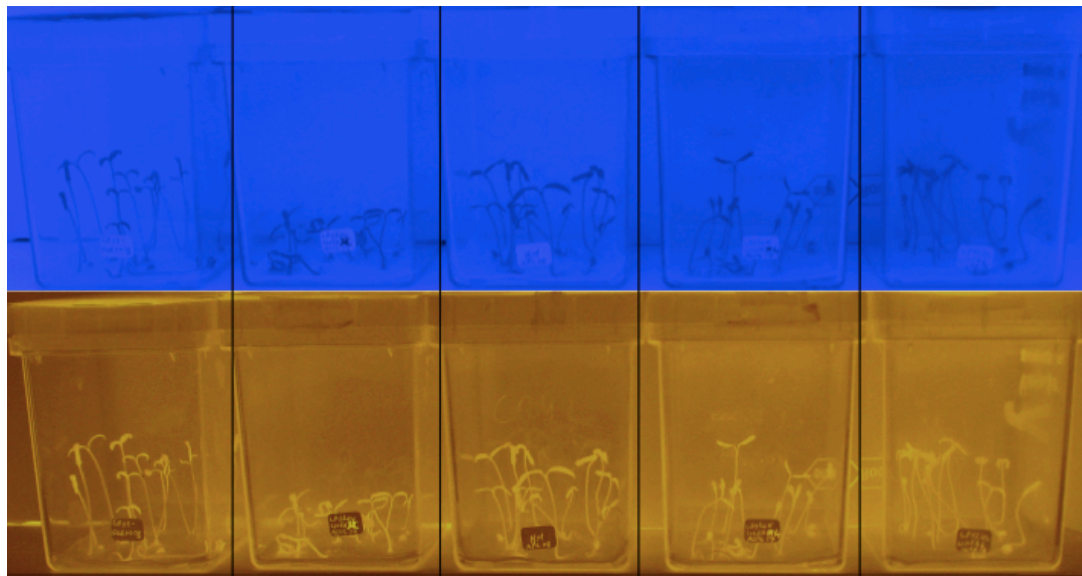
Figure 9. Seedlings growth in Magenta box under blue light. Original colour (upper) and false colour (lower) image after 7 days of growth.

In high irradiance conditions, for both *CRY2* and *CRY1b*, two lines showed strong and reliable phenotypes (lines 6 and 18 for *CRY2* and lines 11 and 3 for *CRY1b*), while the third showed an intermediate phenotype (Fig. 10).
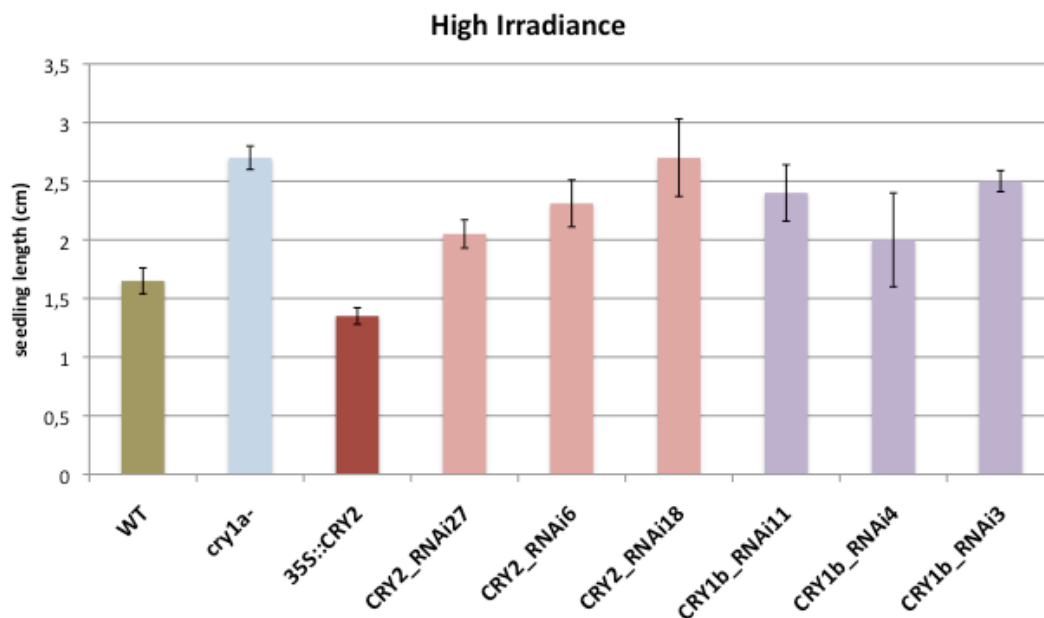

Figure 10. Seedling length after a treatment of 7 days with blue light (high irradiance, 15 μE m^{-2} s^{-1}). Data shown are the average of 12 biological replicas, with error bars representing standard deviation.

In low irradiance, only pHg-12_*CRY2* lines 6 and 8 showed a strong and reliable phenotype while all the pHg-12_*CRY1b* lines showed an intermediate phenotype (Fig. 11).
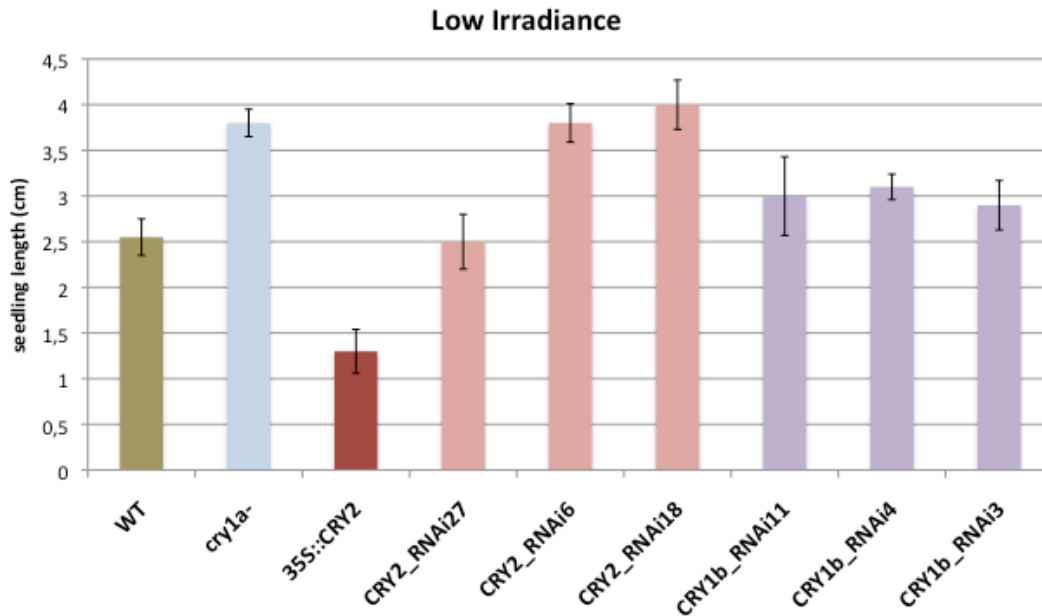
Figure 11. Seedling length after a treatment of 7 days with blue light (low irradiance, 2 μE m^{-2} s^{-1}). Data shown are the average of 12 biological replicas, with error bars representing standard deviation.

Our conclusions are that RNAi shows strong, reliable phenotypes already in the T1 generation. *CRY2* seems to control both high and low irradiance blue light responses while *CRY1b* is mainly involved in high irradiance responses.

4.4.2 TILLING

4759 M3 families of the M82 EMS TILLING platform were screened and mutations were detected in the three *CRY* genes using the mismatch-specific endonuclease ENDO1 as previously described [46]. Individual mutant lines were identified following a pool deconvolution step and Li-Cor visualization. Then, mutated nucleotides were identified by sequencing.

A total of 14 mutations were identified: 6 on *CRY1a*, 3 on *CRY1b*, and 5 on *CRY2*. The putative impact of the missense mutations on the function of the tilled genes was analysed by SIFT (see chapter 3) (Tab. 1).

Gene	nt mutation	aa mutation	aa position	SIFT prediction
CRY1A	A to G	S to G	198/679	tolertaed 0,59
	C to T	P to S	224/679	tolertaed 0,65
	G to A	G to R	288/679	AFFECT PROTEIN FUNCTION 0,0
	C to T	L to F	309/679	tolertaed 0,06
	G to A	D to N	320/679	tolertaed 0,21
	C to T	R to C	404/679	tolertaed 0,1

Gene	nt mutation	AA mutation	aa position	SIFT prediction
CRY1B	C to T	S to F	152/583	AFFECT PROTEIN FUNCTION 0,0
	C to T	H to Y	248/583	AFFECT PROTEIN FUNCTION 0,01
	G to A	G to S	276/583	tolertaed 0,74

Gene	nt mutation	AA mutation	aa position	SIFT prediction
CRY2	T to A	F to I	148/635	AFFECT PROTEIN FUNCTION 0,0
	G to A	E to K	196/635	tolertaed 0,2
	G to A	W to *stop*	318/635	
	G to A	V to I	338/635	tolertaed 0,13
	A to G	M to V	375/635	AFFECT PROTEIN FUNCTION 0,05

Table 1. List of mutations identified in the three selected genes, with their position and the SIFT value.

Phenotyping of these mutant lines is necessary to understand if the single nucleotide mutation affects protein function. Multiplication and genotyping of mutant lines were necessary before starting the phenotyping and is actually in progress. Many lines are still heterozygous or have a low germination rate. In spite of this, a first screening of a *cry1b-* mutant and a *cry2-* mutant, show hypocotyl elongation under blue light (Fig 12).
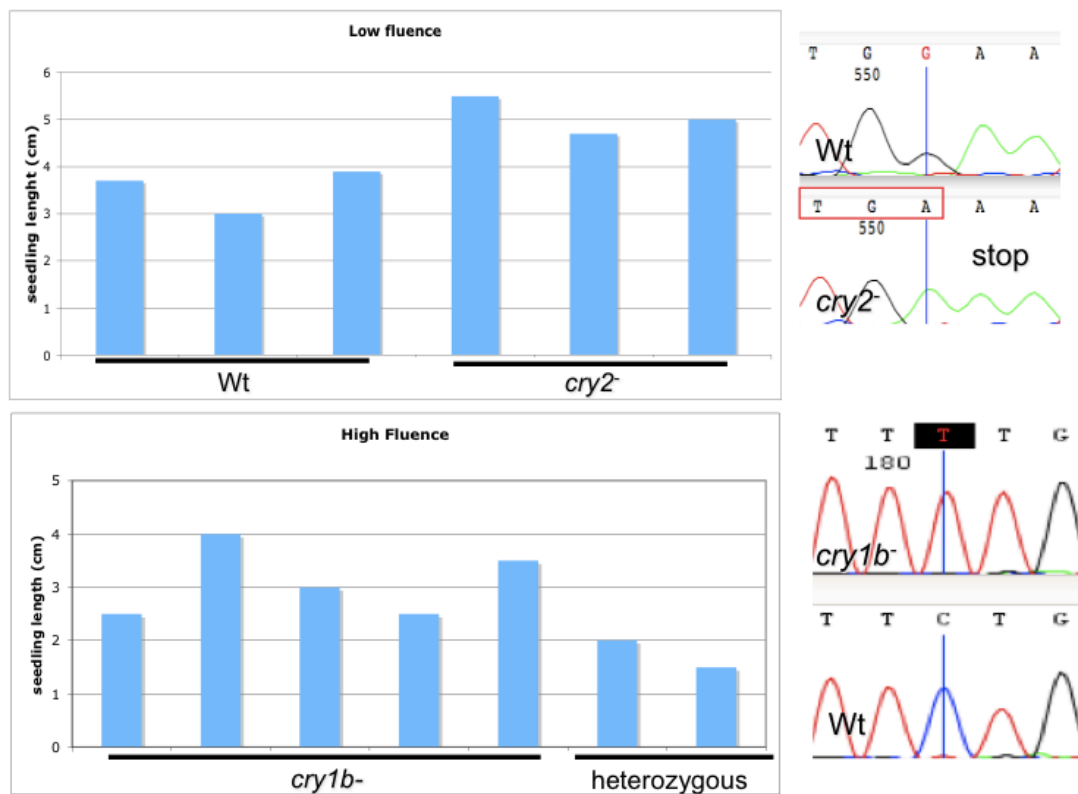
Figure 12. Seedling length after a treatment of 7 days with blue light (low irradiance, 2 µE m^{-2} s^{-1}; high irradiance 15 µE m^{-2} s^{-1}). Measures of individual seedlings are shown.

Thus, in both mutants, both predicted protein function and phenotype are affected. Until now, no *cry1b-* and *cry2-* mutants have been described in tomato and these preliminary results, resulting both from RNAi and TILLING, open interesting perspectives for the functional characterization of *CRY1b* and *CRY2*. Both genes seem to be functional. On the technical side, we conclude that VIGS is a fast and high throughput method, but is unreliable for developmental phenotypes, while RNAi is slower and has a medium throughput, but shows robust phenotypes already in T1. TILLING is slower than RNAi but has the same throughput, with the advantage that allelic series are obtained and it is a non-GMO method.

4.5 Preliminary characterization of 35S::CRY1a lines

In order to produce a complete set of functional data for cryptochrome genes, overexpressor lines of *CRY1a* have been obtained. Overexpressors of *CRY2* were previously described [63]. A complete characterization of 35S::*CRY1a* is actually in progress in our laboratory. However, data on the flowering time of three overexpressor lines have been obtained and reported in this thesis, as complementary data on the importance of cryptochromes in the regulation of flowering. We measured the number of days from germination to the anthesis of the first flower, in green house conditions under long days, as described in paragraph *4.3.3*. Data are shown in figure 13.
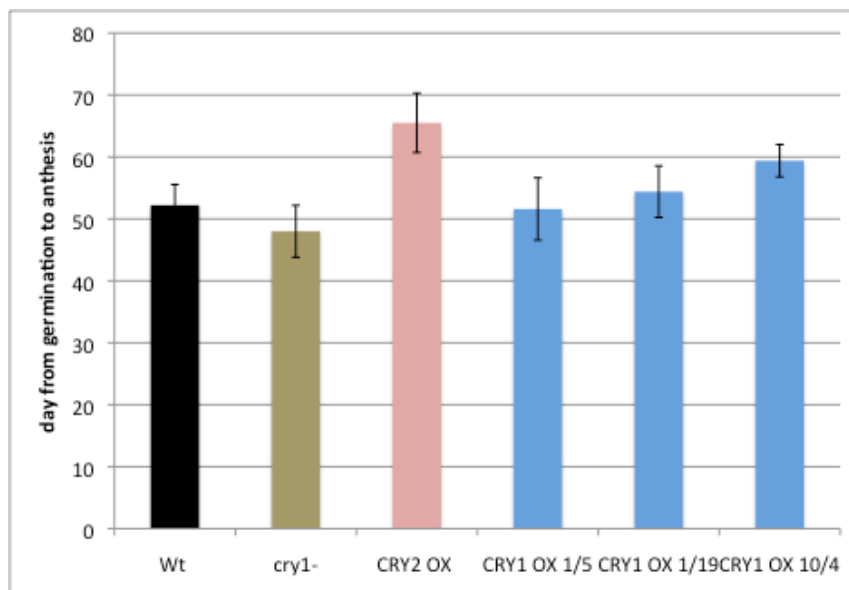
Figure 12. Number of days from seed germination to the anthesis of the first flower. Three 35S::<u>*CRY1a*</u> lines are analysed, with wild type, cry1a- mutant and overexpressor of *CRY2* as controls. Data shown are the average of three biological replicas, with error bars representing standard deviation.

The data indicate that, in one of the lines, the overexpression of *CRY1a* induce a delay in flowering, in agreement with the phenotype of the mutant *cry1a-*, in which flowering is slightly anticipated with respect to the wild type. However, the effect of *CRY1a* in retarding flowering, seems to be less strong than that of *CRY2*.

References:

1.   Ahringer, J., *ed. Reverse genetics, WormBook, ed. The C. elegans Research Community, WormBook, doi/10.1895/wormbook.1.47.1*, ed. V. Ambros. 2006.

2.   An, G., et al., *Reverse genetic approaches for functional genomics of rice.* Plant Mol Biol, 2005. **59**(1): p. 111-23.

3.   Burch-Smith, T.M., et al., *Applications and advantages of virus-induced gene silencing for gene function studies in plants.* Plant J, 2004. **39**(5): p. 734-46.

4.   Kusaba, M., *RNA interference in crop plants.* Curr Opin Biotechnol, 2004. **15**(2): p. 139-43.

5.   Henikoff, S. and L. Comai, *Single-nucleotide mutations for plant functional genomics.* Annu Rev Plant Biol, 2003. **54**: p. 375-401.

6.   The_Arabidopsis_Genome_Initiative, *Analysis of the genome sequence of the flowering plant Arabidopsis thaliana.* Nature, 2000. **408**(6814): p. 796-815.

7.   International_Rice_Genome_Sequencing_Project, *The map-based sequence of the rice genome.* Nature, 2005. **436**(7052): p. 793-800.

8.   Jaillon, O., et al., *The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla.* Nature, 2007. **449**(7161): p. 463-7.

9.   Mueller, L.A., et al., *The SOL Genomics Network: a comparative resource for Solanaceae biology and beyond.* Plant Physiol, 2005. **138**(3): p. 1310-7.

10.   Agrawal, N., et al., *RNA interference: biology, mechanism, and applications.* Microbiol Mol Biol Rev, 2003. **67**(4): p. 657-85.

11.   Fire, A., et al., *Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans.* Nature, 1998. **391**(6669): p. 806-11.

12.   Hannon, G.J., *RNA interference.* Nature, 2002. **418**(6894): p. 244-51.

13.   Pickford, A.S. and C. Cogoni, *RNA-mediated gene silencing.* Cell Mol Life Sci, 2003. **60**(5): p. 871-82.

14.   Tijsterman, M., R.F. Ketting, and R.H. Plasterk, *The genetics of RNA silencing*. Annu Rev Genet, 2002. **36**: p. 489-519.

15.   Mocellin, S. and M. Provenzano, *RNA interference: learning gene knock-down from cell physiology*. J Transl Med, 2004. **2**(1): p. 39.

16.   Covey, S.N., et al., *Plants combat infection by gene silencing*. Nature, 1997. **385**: p. 781Ð782.

17.   Ratcliff, F., B.D. Harrison, and D.C. Baulcombe, *A similarity between viral defense and gene silencing in plants*. Science, 1997. **276**: p. 1558Ð 1560.

18.   Al-Kaff, N.S., et al., *Transcriptional and posttranscriptional plant gene silencing in response to a pathogen*. Science, 1998. **279**(5359): p. 2113-5.

19.   Kumagai, M.H., et al., *Cytoplasmic inhibition of carotenoid biosynthesis with virus-derived RNA*. Proc Natl Acad Sci U S A, 1995. **92**(5): p. 1679-83.

20.   Baulcombe, D.C., *Fast forward genetics based on virus-induced gene silencing*. Curr Opin Plant Biol, 1999. **2**(2): p. 109-13.

21.   Smith, N.A., et al., *Total silencing by intron-spliced hairpin RNAs*. Nature, 2000. **407**(6802): p. 319-20.

22.   Ruiz, M.T., O. Voinnet, and D.C. Baulcombe, *Initiation and maintenance of virus-induced gene silencing*. Plant Cell, 1998. **10**(6): p. 937-46.

23.   Brunt, A.A., et al., *Plant viruses online: descriptions and lists from the VIDE data base. Version: 20th August 1996*. 1996.

24.   Ratcliff, F., A.M. Martin-Hernandez, and D.C. Baulcombe, *Technical Advance. Tobacco rattle virus as a vector for analysis of gene function by silencing*. Plant J, 2001. **25**(2): p. 237-45.

25.   Hull, R., *MatthewsÕ plant virology, 4th edn*. 2002, Academic, New York.

26.   Liu, Y., M. Schiff and S.P. Dinesh-Kumar, *Virus-induced gene silencing in tomato*. Plant J, 2002. **31**(6): p. 777-86.

27.	Liu, Y., et al., *Tobacco Rar1, EDS1 and NPR1/NIM1 like genes are required for N-mediated resistance to tobacco mosaic virus.* Plant J, 2002. **30**(4): p. 415-29.

28.	Waterhouse, P.M. and C.A. Helliwell, *Exploring plant genomes by RNA-induced gene silencing.* Nat Rev Genet, 2003. **4**(1): p. 29-38.

29.	Wesley, S.V., et al., *Construct design for efficient, effective and high-throughput gene silencing in plants.* Plant J, 2001. **27**(6): p. 581-90.

30.	Chen, S., et al., *Temporal and spatial control of gene silencing in transgenic plants by inducible expression of double-stranded RNA.* Plant J, 2003. **36**(5): p. 731-40.

31.	Guo, H.S., et al., *A chemical-regulated inducible RNAi system in plants.* Plant J, 2003. **34**(3): p. 383-92.

32.	Hartley, J.L., G.F. Temple, and M.A. Brasch, *DNA cloning using in vitro site-specific recombination.* Genome Res, 2000. **10**(11): p. 1788-95.

33.	Helliwell, C. and P. Waterhouse, *Constructs and methods for high-throughput gene silencing in plants.* Methods, 2003. **30**(4): p. 289-95.

34.	Helliwell, C., et al., *High-throughput vectors for efficient gene silencing in plants.* Funct Plant Biol, 2002. **29**: p. 217 Đ 1225.

35.	McCallum, C.M., et al., *Targeting induced local lesions IN genomes (TILLING) for plant functional genomics.* Plant Physiol, 2000. **123**(2): p. 439-42.

36.	Colbert, T., et al., *High-throughput screening for induced point mutations.* Plant Physiol, 2001. **126**(2): p. 480-4.

37.	Till, B.J., et al., *High-throughput TILLING for functional genomics.* Methods Mol Biol, 2003. **236**: p. 205-20.

38.	Till, B.J., et al., *High-throughput TILLING for Arabidopsis.* Methods Mol Biol, 2006. **323**: p. 127-35.

39.	Martin, B., et al., *A high-density collection of EMS-induced mutations for TILLING in Landsberg erecta genetic background of Arabidopsis.* BMC Plant Biol, 2009. **9**: p. 147.

40.   Slade, A.J., et al., *A reverse genetic, nontransgenic approach to wheat crop improvement by TILLING.* Nat Biotechnol, 2005. **23**(1): p. 75-81.

41.   Caldwell, D.G., et al., *A structured mutant population for forward and reverse genetics in Barley (Hordeum vulgare L.).* Plant J, 2004. **40**(1): p. 143-50.

42.   Till, B.J., et al., *Discovery of chemically induced mutations in rice by TILLING.* BMC Plant Biol, 2007. **7**: p. 19.

43.   Wu, J.L., et al., *Chemical- and irradiation-induced mutants of indica rice IR64 for forward and reverse genetics.* Plant Mol Biol, 2005. **59**(1): p. 85-97.

44.   Till, B.J., et al., *Discovery of induced point mutations in maize genes by TILLING.* BMC Plant Biol, 2004. **4**: p. 12.

45.   Cooper, J.L., et al., *TILLING to detect induced mutations in soybean.* BMC Plant Biol, 2008. **8**: p. 9.

46.   Dalmais, M., et al., *UTILLdb, a Pisum sativum in silico forward and reverse genetics tool.* Genome Biol, 2008. **9**(2): p. R43.

47.   Xin, Z., et al., *Applying genotyping (TILLING) and phenotyping analyses to elucidate gene function in a chemically induced sorghum mutant population.* BMC Plant Biol, 2008. **8**: p. 103.

48.   Boualem, A., et al., *A conserved mutation in an ethylene biosynthesis enzyme leads to andromonoecy in melons.* Science, 2008. **321**(5890): p. 836-8.

49.   Rigola, D., et al., *High-throughput detection of induced mutations and natural variation using KeyPoint technology.* PLoS One, 2009. **4**(3): p. e4761.

50.   Gady, A.L., et al., *Implementation of two high through-put techniques in a novel application: detecting point mutations in large EMS mutated plant populations.* Plant Methods, 2009. **5**: p. 13.

51.   Le Signor, C., et al., *Optimizing TILLING populations for reverse genetics in Medicago truncatula.* Plant Biotechnol J, 2009. **7**(5): p. 430-41.

52.   Moens, C.B., et al., *Reverse genetics in zebrafish by TILLING*. Brief Funct Genomic Proteomic, 2008. **7**(6): p. 454-9.

53.   Winkler, S., et al., *Target-selected mutant screen by TILLING in Drosophila*. Genome Res, 2005. **15**(5): p. 718-23.

54.   Gilchrist, E.J., et al., *TILLING is an effective reverse genetics technique for Caenorhabditis elegans*. BMC Genomics, 2006. **7**: p. 262.

55.   Greene, E.A., et al., *Spectrum of chemically induced mutations from a large-scale reverse-genetic screen in Arabidopsis*. Genetics, 2003. **164**(2): p. 731-40.

56.   Koornneef, M., *Classical mutagenesis in higher plants*, in *Molecular Plant Biology*, C. Gilmartin P. M. and Bowler, Editor. 2002, Oxford University Press: Oxford, GB. p. 1-10.

57.   Ekengren, S.K., et al., *Two MAPK cascades, NPR1, and TGA transcription factors play a role in Pto-mediated disease resistance in tomato*. Plant J, 2003. **36**(6): p. 905-17.

58.   Lopez-Gomez, R. and M.A. Gomez-Lim, *A Method for extracting intact RNA from fruits rich in polysaccharides using ripe mango mesocarp*. HortScience, 1992. **27**: p. 440-442.

59.   van Roekel, J.S.C., et al., *Factors influencing transformation frequency of tomato (Lycopersicon esculentum)*. Plant Cell Rep, 1993. **12**: p. 644-647.

60.   Menda, N., et al., *In silico screening of a saturated mutation library of tomato*. Plant J, 2004. **38**(5): p. 861-72.

61.   Piron, F., et al., *Engineering tomato plants resistant to potyviruses using the TILLING approach*. submitted.

62.   Triques, K., et al., *Characterization of Arabidopsis thaliana mismatch specific endonucleases: application to mutation discovery by TILLING in pea*. Plant J, 2007. **51**(6): p. 1116-25.

63.   Giliberto, L., et al., *Manipulation of the blue light photoreceptor cryptochrome 2 in tomato affects vegetative development, flowering time, and fruit antioxidant content*. Plant Physiol, 2005. **137**(1): p. 199-208.

ABBREVIATIONS

ABA, abscisic acid

AFLP , Amplified fragment length polymorphism

BAC , bacterial artificial chromosome

cDNA, complementary DNA

dN, numbers of non-synonymous substitutions per non-synonymous site

DNA, deoxyribonucleic acid

DNP, day neutral plant

dNTPs, deoxynucleotide triphosphates

dS, numbers of synonymous substitutions per synonymous site

dsRNA, double-stranded RNA

EMS, ethylmethane sulfonate

EST, expressed sequence tagess

FISH, fluorescent in-situ hybridization

GA, Gibberellins

GBSSI , Granule-bound starch synthase

GMO , genetically modified organism

hpRNA, hairpin RNA

ihpRNA, intron-containing hpRNA

IL, introgression line

iTAG, International Tomato Annotation Group

IU, international unit

LD, long day

LD15, lethal dose 15%

LDP, long day plant

mRNA, messenger RNA

PCR , polymerase chain reaction

pHg-12, pHellsgate 12

PTGS, post-transcriptional gene silencing

pTRV, plasmid TRV

PVX, Potato virus X

QTL, quantitative trati locus

RFLP, restriction fragment length polymorphism

RISC, RNA-induced silencing complex

RNA, ribonucleic acid

RNAi, RNA interference

SAM, shoot apical meristem

SD, short day

SDP, short day plant

SGN , SOL genomic network

siRNA, small interfering RNA

SNP, single nucleotide polymorphism

SOL, Solanaceae Genome Project

TAIL-PCR, thermal asymmetric interlaced PCR

Taq, DNA polymerase from Thermus acquaticus

TGRC, Tomato Genetics Resource Centre

TILLING, Targeting Induced Local Lesions In Genomes

TMV, Tobacco mosaic virus

TRV, Tobacco rattle virus

UTR , untranslated region

VIGS, virus induced gene silencing

WGS, whole genome shotgun sequencing

ZT , Zeitgeber time

<u>Genes</u>

ABI4, ABA-INSENSITIVE 4

AP1, APETALA1

AP2, Activator protein 2 alpha

CAB, CHLOROPHYLL a/b BINDING PROTEIN

CCA, CIRCADIAN CLOCK ASSOCIATED

CDF1, CYCLING DOF FACTOR1

CO, CONSTANS

COL, CONSTANS-like

CRY, Cryptochrome

FA, FALSIFLORA

FKF1, FLAVIN-BINDING, KELCH REPEAT, F-BOX

FT, FLOWERING LOCUS T

GI, GIGANTEA

HD1, Heading date 1

HD3a, Heading date 3a

LFY, LEAFY

LHY, LATE ELONGATED HYPOCOTYL

MBD, Methyl-CpG binding domain

PDS, Phytoene desaturase

PHY, Phytochrome

S, COMPOUND INFLORESCENCE

SFT, SINGLE FLOWER TRUSS

TCOL, tomato CONSTANS-like

TGI, tomato GIGANTEA

TOC, TIMING OF CAB EXPRESSION

UF, UNIFLORA

ACKNOWLEDGMENTS

APPENDIX: pictures and geographical localization of tomato wild species

S. neorickii (DN)

S. chmielewskii (DN)

S. chilense (DN)

S. pennellii (DN)

Equatore

S5°

W85°

W75°

S15°

Tropico del Capricorno

S25°

MapLink/Tele Atlas
of State Geographer
2010 DMapas
Europa Technologies