UNIVERSITÀ DEGLI STUDI DI UDINE
hic sunt futura

www.uniud.it

Corso di dottorato di ricerca in Scienze e Biotecnologie Agrarie

in convenzione con Università degli Studi di Udine

Dipartimento di Scienze Agroalimentari, Ambientali e Animali

Ciclo XXX – Coordinatore: prof. Giuseppe Firrao

TESI di DOTTORATO di RICERCA

Fine mapping of resistance genes to Sharka (PPV, *Plum pox virus*) in apricot (*Prunus armeniaca L.*), gene prediction and annotation of the region of interest

Dottorando

Gloria De Mori

Supervisore

Raffaele Testolin

Anno accademico 2017/2018

INTRODUCTION

Sharka disease

Sharka disease, caused by *Plum Pox Virus* (PPV), is considered one of the most detrimental diseases affecting many stone fruits and is among the most studied viral diseases in the world (Scholthof *et al.*, 2011).

Sharka was first reported in plum trees in Bulgaria between 1915 and 1918, at the end of the First World War, although some reports indicate that symptoms related to this virus disease had been already detected in 1910 in Macedonia (Levy L. *et al*., 2000).

However, the first article describing the viral nature of the disease was not published until 1932 when Atanosoff called this disease "Sarkaposilvite" which means "plum pox" (= Sharka).

Since then, Sharka disease has spread progressively to most European areas, around the Mediterranean basin and the Near and Middle East. Roy & Smith (1994) distinguished three zones:

- The central and eastern countries in which PPV spread relatively early and levels of disease are generally high (Bosnia-Herzegovina, Bulgaria, Croatia, Czech Republic, Hungary, Moldova, Poland, Romania, Serbia, Slovakia, Slovenia, Ukraine);
- The Mediterranean countries in which spread is recent and there is a high risk of further spread (Albania, Cuprus, Egypt, Greece, Italy, Portugal, Spain, Syria, Turkey);
- The northern and western countries in which levels of PPV are very uneven (fairly widespread in Austria, Germany, and the UK-England), very localized (Belgium, France and Luxemburg) and eradicated (Denmark, Netherlands and Switzerland).

It has also spread to South and North America and Asia (Barba *et al.*, 2011), with the exception of Australia, New Zealand, South Africa and California (USA).

The primary cause of the wide diffusion of the PPV is probably due to the illegal trade and insufficiently controlled exchanges of plant material in the global market.

Moreover, for decades there was no awareness regarding the severity of this virus disease and there were no methods of detection both reliable and suitable for large scale

application. As a result, the disease easily escaped the visual inspection because of the inefficient control methods and it spread around the world (Cambra *et al.,* 2006).

During the last decades, Sharka disease has had a significant agronomic impact and resulted in major economic losses, affecting mostly the *Prunus* genus.

The cost associated with the disease in many countries not only involves yield and quality losses and the costs of quarantine, eradication and compensatory measures, but also indirect costs related to preventative measures, inspections, diagnostics and their impact on foreign and domestic trade (Barba *et al*., 2011). It has been estimated that the costs of managing Sharka worldwide since the 1970s have exceeded 10,000 million euros (Cambra *et al*., 2006).

Plum pox virus

Plum Pox Virus (PPV) is a member of the largest and most economically important group of plant viruses: the genus *Potyvirus*. *Potyvirus* is the major genus in the *Potyviridae* family, which also includes *Rymovirus*, *Macluravirus*, *Ipomovirus*, *Bymovirus* and *Tritimovirus*.

PPV is a filamentous virus with particles 660 – 750 nm long and 12.5 – 20.0 nm in diameter (Fig. 1). Its genome consists of a positive-sense single-stranded RNA (ssRNA) of 9741 – 9795 nucleotides (Fanigliulo *et al.,* 2003; Glasa and Šubr, 2005; Glasa *et al*., 2011, 2013; James and Varga, 2005; Laín *et al*., 1989; Maejima *et al*., 2011; Maiss *et al*., 1989; Myrta *et al*., 2006; Palkovics *et al*., 1993; Teycheney *et al*., 1989; Ulubas *et al*., 2009; SharCo database, http://w3.pierroton.inra.fr:8060/).

The PPV genomic RNA carries a VPg (viral protein genome – linked) covalently bound to its 5’ end (Riechmann *et al*., 1989),and a poly(A) tail at its 3’ end (Garcia *et al.*,1994; Lain *et al.*, 1998). The genome is encapsidated by about 2,000 copies of coat protein (CP) units. Each of them are composed by a peptide of 330-332 amino acids, corresponding to a molecular weight of 36-38 kDa. However, detectable levels of another viral protein, helper component proteinase (HCPro), have been found to be associated with PPV virions (Manoussopoulos *et al.,* 2000).

This association could be related to the ability of HCPro to act as a bridge between virus particles and the stylet of aphids, which specifically transmits the virus (Urcuqui-Inchima *et al*., 2000). However, roles unrelated to aphid transmission have been suggested for interactions between HCPro and CP (Roudet-Tavert *et al*., 2002).

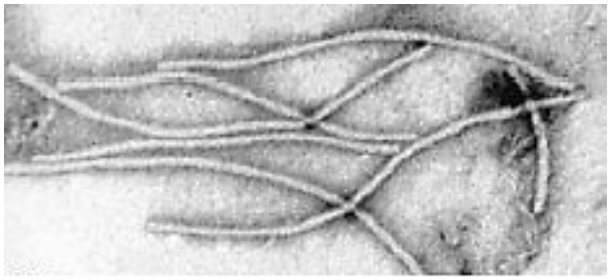


Fig. 1 – Electronic microscope view of PPV viral particles (Levy L. *et al*., 2000).

The genomic RNA encodes a long open reading frame (ORF) starting from an AUG codon (nucleotide 36). However, the results of several in vitro researches support that the genomic RNA translation begins at nucleotide 146 with the second AUG codon (Simon *at al*., 1997).

The polypeptide chain (3123 – 3143 aminoacids) is processed by three virus-encoded proteinases to produce ten mature protein products: P1, HCPro, P3, 6K1, CI, 6K2, VPg, NIapro, Nib and CP. The viral proteins are all involved in genome amplification and all of them except P3, 6K1 and 6K2 bind RNA. Movement functions are essentially controlled by proteins clustered in the N-terminal region of the polyprotein whereas the proteins forming the replication complex are contained in the C – terminal region of the polyprotein.

The general properties of these 10 proteins are as follows (Urucuqui-Inchima*et al*., 2000) (Fig. 2):

- P1 – it is a protein with proteolytic activity, which separates from the polypeptide chain for self – catalysis. For this reason, it is classified as an endopeptidase. Its precise function in viral infection has yet to be established. Non-specific RNA binding has often been attributed to the involvement of the protein in viral movement and in the symptomatology. The fusion between P1 and HCPro carries the potential

of a broad pathogenicity enhancer, which bears on suppression of host defense and on suppression of post-transcriptional gene silencing (PTGS).

- HCPro – it is a multifunctional protein: It is required for aphid transmission, it's involved in the long distance movement of the viral particle inside the plant, the viral genome amplification and the suppression of gene silencing (PTGS), and it has the ability to self-interact.
- P3 – the role of this protein is still unknown. It seems to influence the pathogenicity of some viruses. Indeed, the development of disease symptoms might be closely related to the interaction of the P3 protein and the host plant proteins.
- 6K1 – it is probably responsible for the movement of the virus from cell to cell. This peptide is normally found bound to P3, and together they appear to regulate the pathogenicity of some viruses.
- CI – this protein possesses ATPase activity and unwinds RNA duplexes. It may participate in cell-to-cell movement of the virus, being involved in the cell-to-cell passage of the viral RNA-protein complexes. The function of CI in virus replication is still largely unknown.
- NIa and NIb – NIa is composed of two domains, the N-terminal VPg domain, and the C-terminal proteinase domain. These two domains will be referred to as VPg and NIaPro. The latter is the major proteinase of potyviruses: it processes the polyprotein in *cis* and in *trans* to produce functional products. VPg domain has essential functions in viral replication and host genotype specificity. For most potyviruses, NIa is colocalized with NIb in inclusion bodies in the nucleus of infected cells. NIb protein generally forms inclusions in the nucleus of infected plants, even though it is required in the cytoplasms or in membranes associated with replication complexes during viral RNA synthesis.
- CP – it can be divided into three domains, the variable N- and C-terminal domains that are exposed on the surface of the particle and are sensitive to mild trypsin treatment, and the more conserved central or core domain required for the encapsidation of the viral RNA.

  The CP in involved in several mechanisms such as aphid transmission, cell-to-cell and systemic movement, encapsidation of the viral RNA and the regulation of viral

RNA amplification. Specifically, the core domain and the N-terminal domain combined with CI protein seem to be involved in the cell-to-cell virus spread. The terminal domains combined with HCPro and the genomic protein VPg regulate the translocation of virions through the vascular system. In the exposed N-terminal region of the CP is present the DAG motif that is highly conserved and essential for the aphid transmission. The analysis of the results obtained with CP mutants, as well as those obtained with HCPro, indicate that a strong correlation exists between aphid transmissibility and HCPro-CP interaction to form a complex indispensable for efficient transmission.

- 6K2 – this peptide possesses no established enzymatic function. It is proven that binding to membranes occurs via the central hydrophobic domain of the 6K2 peptide. The protein is associated with large vesicular compartments deriving from the endoplasmatic reticulum (ER). It has consequently been proposed that the 6K2 peptide is required for genome amplification and that it anchors the replication apparatus to ER-like membranes.

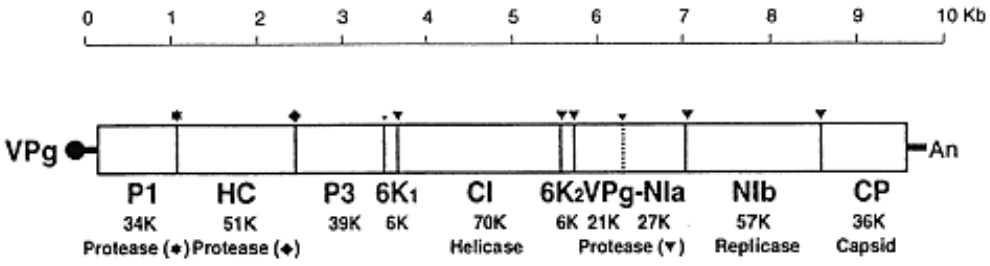


Fig. 2 – PPV genome map. RNA genome is represented below the graduated line with VPg at the 5 'end and the poly A tail at the 3' end. The only ORF is represented as a rectangle indicating the abbreviated protein names, their molecular weights (kDa) and functions. The proteolitic sites are indicated by vertical lines and a symbol above. Smaller symbols and dotted lines indicate partial or suboptimal digestion (Lopez-Moya et al., 2000).

As reported for other potyviruses (Chung *et al*., 2008), another PPV protein, P3N-PIPO, is predicted to be produced by a frameshift into a short ORF embedded within the P3 coding sequence.

Several efforts have been devoted to the study of the biological, serological and molecular variability of PPV (Fanigliulo *et al*., 2003; Glasa *et al*., 2004; James e Varga, 2005; Laín *et al*., 1989a; Maiss *et al*., 1989; Myrta *et al*., 2006; Palkovics *et al*., 1993; Sáenz *et al*., 2000; Teycheney *et al*., 1989).

These efforts have revealed that the diversity of PPV is structured into individual monophyletic ensembles of closely related isolates, which have been designated as strains.

Currently, eight strains are recognized for PPV, which may be more than for any other potyvirus:

- **PPV-D**, named Dideron (French fruit farmer in whose apricot plant was discovered the disease). It is widespread in Europe and is also responsible for most outbreaks outside of Europe. This strain is widely presented on apricots and plums and less associated with peach under natural conditions.
- **PPV-M**, named Marcus (named after the Greek peach variety of Markus). It is found mainly in southern and central European countries. This strain is efficiently aphid transmitted, causing fast epidemics, mainly in peach orchards but others orchards like plum, apricot, cherry plum and some rootstocks belonging to the genus *Prunus*.
- **PPV-C**, from the English cherry-term (due to the sour and sweet cherry trees on which it was identified). PPV isolates naturally infecting sour cherries in Moldova. In Italy it was reported a single case of PPV-C sweet cherry trees infections, in Puglia in 1992. Given its restricted natural host range, the actual epidemiological impact of this strain seems to be lower than that of the major PPV strains.
- **PPV-El Amar**, it was identified in the 1990s. It is present in some areas of apricot cultivation in Egypt.
- **PPV-Rec** (recombinant), it was found in several European countries, as well as outside Europe, mainly infecting plum and apricot trees. This strain is a homogeneous group of isolates deriving from a recombination between PPV-M and PPV-D. Given its wide distribution and prevalence, it is now considered as the third

major PPV strain. As the first reported PPV recombinant isolate originated from Serbia (Cervera *et al*., 1993), the Balkans have been suggested to be the center of origin of PPV-Rec, which then spread to other areas through the exchange of infected propagation material of tolerant plum genotypes (Glasa *et al*., 2004).

- **PPV-W**, named Winona. It was originally detected in 2003 in plum trees in Canada (James and Varga, 2005). Later, it was recorded in Latvia, Ukraine and Russia (Glasa *et al*., 2011; Mavrodieva *et al*., 2013; Sheveleva *et al*., 2012), confirming the suggestion that the origin of this strain may be found in eastern Europe. The PPV-W strain has been found in fields of plum, blackthorn, Canadian plum, cherry plum and downy cherry.
- **PPV-T** (Turkey), has been found to be widely distributed in apricots, peaches and plums in Turkey, and an occasional finding of this strain has been recorded in Albania (unpublished results of the European SharCo FP7 project). Genome characterization of this strain has revealed a recombination event affecting its 5' genomic region (Glasa & Candresse, 2005).
- **PPV-CR** (Cherry Russian). It was isolated very recently from naturally infected sour cherries in the Volga river basin (Russia). The epidemiology of this strain remains to be determined. An additional putative PPV strain (**PPV-An**) could be represented by a recently identified isolate from eastern Albania (Palmisano *et al*., 2012).

Full-length genomic sequences have been determined for PPV isolates representing each of the recognized strains, providing a clear picture of the phylogenetic relationship between strains and the PPV evolutionary history (Fig.3). PPV strains are characterized by relatively low intra-strain diversity and by comparatively high inter-strain diversity (Glasa *et al*., 2012).

Phylogenetic analysis based on the complete viral genome sequences shows that three strains (PPV-D, PPV-M and PPV-Rec) along with PPV-T create an evolutionarily related supercluster of isolates, clearly distinguished from the other. Although forming monophyletic groups, PPV-M, PPV-D, PPV-Rec, PPV-T and PPV-W are evolutionarily linked by recombination events, including an ancestral recombination affecting the 5' part of PPV-M, PPV-D and PPV-Rec strains.

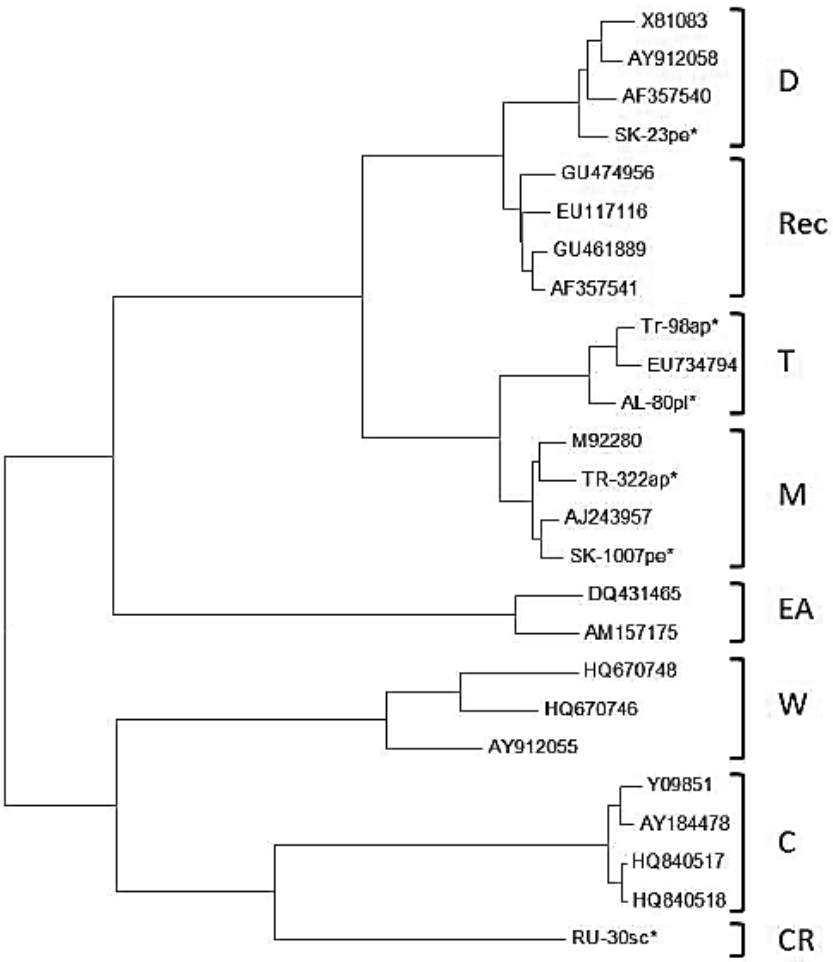


Fig. 3 – Unrooted phylogenetic tree of PPV isolates based on their complete nucleotide sequences (Glasa and Šubr, 2013).

The main pathways for PPV spread over long distances are the illegal traffic and insufficiently controlled exchanges of plant material in a global market.

In nature the virus is transmitted by grafting or by many species of aphids. The main vector species are *Myzuspersicae, Aphis spiricula* and *Hyalopteruspruni*. Other aphids have been shown to transmit at lower frequency than the main vectors: *Aphis craccivora, A. fabae, Brachycauduscardui, B. helychryi, B. persicar, Myzusvarians, Phorodonhumuli*.

Aphids transmit the virus in a non-persistent manner, this means that PPV can be transmitted from a sick plant to a healthy one through a single sting probe. Only 5-10

minutes of acquisition are needed for the aphid to become infectious, and the infectivity lasts for several hours, enabling it to spread the infection at long distance. A single probe of a viruliferous aphid is sufficient to inoculate about 26000 PPV RNA molecules in a receptor GF305 peach seedling, with a 20% chance of resulting in a systemic infection (Moreno *et al*., 2009).

The efficiency of natural transmission by aphids and the spatial pattern of spread of Sharka may differ for different PPV isolates and host cultivars: resistant and tolerant varieties require a greater density of aphid population or a longer acquisition time.

There is no confirmed evidence for seed or pollen transmission of PPV in any of its *Prunus*hostes (Pasquini and Barba, 2006).

Host Plants

Sharka disease affects plants of the genus *Prunus*, used as commercial cultivars as well as rootstocks.

The main fruit crops susceptible to this disease are: apricot (*P. armeniaca*), peach (*P. persica*), plum (*P. domestica*), almond (*P. amygdalus*), sweet cherry (*P. avium*) and sour cherry (*P. cerasus*).

In addition, several ornamental and wild *Prunus* species have been identified as natural or experimental hosts of PPV (Damstreegt *et al*., 2007, James and Thompson, 2006), for instance, *P. spinosa*, *P. laurocerasus*, *P. salicina*, *P. cerasifera*, *P. insititia*, *P. mahaleb*, *P. tomentosa*, *P. brigantina*, *P. triloba* and *P.blireiana*.

Sharka is particularly detrimental in apricots, European plums, peaches and Japanese plums because it can seriously reduce yield and fruit quality.

Numerous cultivated or weedy annual plants can be infected with PPV. A good example are plants belonging to genus *Trifolium*, *Lepidium*, *Zinnia*, *Ligustrum*, *Licium,Euonymus*. Some of these plants, such as *Arabidopsis thaliana*, can be used for biological assays and purification of the virus for experimental purposes. The natural transmission between such herbaceous plants and *Prunus* has never been demonstrated.

PPV Symptoms

The symptoms of this virus disease depend on the sensitivity and variety of the species, the climatic conditions, the vegetative state of the plant, the health condition of the host in relation to other viral infections and the virus strain.

PPV symptoms appear on leaves, shoots, bark, petals, fruits and even stones.

The damage caused by Sharka disease consist in a lower yield and a considerable deterioration of the organoleptic characteristics of the fruits, which prevents the commercialization (Fig. 4).

The symptomsusually appear on the leaves early in the growing season and include mild light-green discoloration, chlorotic spots, bands or rings, vein clearing or yellowing and leaf deformation. Flower symptoms can occur as discoloration on petals of some cultivars. Infected fruits show chlorotic spots or lightly pigmented yellow rings or line patterns. Fruits may become deformed or irregular in shape, and may develop brown or necrotic areas under the discoloured rings. European plums and apricots may also show premature fruit drop, whereas Japanese plums and peaches show typical pale rings or spots. Sweet and sour cherry fruits generally show no or inconspicuous leaf symptoms.

Generally, the fruits of early-ripening cultivars of all susceptible species show more marked symptoms than those of late-ripening cultivars.

Fig. 4 – Sharka disease symptoms: apricot stones with plum pox virus – induced ring patterns in the fruit (Photos courtesy of M. Barba, ISPAVE, Italy, Dunez, INRA, France, and Dr. M A Cambra, DG Aragòn, Spain).

Pest Significance

Unlike fungal or bacterial plant pathogens that can be controlled chemically, there is no anti-virus treatment available to control sharka disease in orchads and the fight against the transmission vectors is inefficient. Indeed, the chemical fight against aphids doesn't prevent the infection spread in field since the virus transmission often occurs before aphids undergo the lethal effect of the aphicide (Giunchedi, 2003). Because of this, the control measures available are essentially preventive. The aim is limiting the virus spread and prevent PPV introduction in a specific area.

The most effective means of control are the following:

- Regulation regarding the importation and movement of a propagative materials and commercial propagants.
- Production of virus-free trees and selection of virus-free budwood and rootstocks.
- Early detection using surveys and subsequent removal and destruction of infected trees.
- Use of resistant cultivars and rootstocks.

Because of the magnitude of the damages caused by PPV and the high spread of the virus, the European and Mediterranean Plant Protection Organization (EPPO/OEPP) has included PPV in the list of quarantine pathogens and the federal government of the United States of America has classified the virus among the top ten most important adversities for its agriculture (Public Health Security and Bioterrorism Act of 2002).

In order to ensure protection from PPV in Italy, a specific Ministerial Decree, D.M. 28 July 2009, requires the obligation to report any suspected of Sharka infection and the eradication of infected plants to narrow the infection site. However, the eradication is not very effective, especially in regions where the disease is endemic.

For these reasons, the aims of European breeding programs are the research of resistance sources to Sharka and the development of resistant genotypes.

Genetic resources and plant breeding

Perennial fruit crops are characterized by long generation timing and large dimensions, which have limited the genetic studies development. Consequently, improvements in breeding programs have been slower than those of herbaceous species.

Plant breeding through artificial selection depends on the ability to distinguish genetic effects from those due to the environment. In this scenario, molecular markers are useful due to their potential unlimited number and their independence from environmental effects (Vogel *et al*., 1996). For that reason, linkage genetic mapping has proved to be a powerful tool for localizing and isolating genes that control both simple and complex characters. Therefore, a genetic linkage map that includes the traits associated with Sharka resistance could be a useful instrument for the marker assisted selection (MAS) in breeding programs

In European plum both quantitative and qualitative (hypersensitivity) sources of resistance have been identified. The former is found for example in the 'Stantley', 'President' and 'RuthGerstetter' varieties, while the latter was first found in the cultivar 'Jojo'.

In peach, despite the extensive screening of several varieties, no sources of PPV resistance has been found. However, several cultivars show significant differences in susceptibility to the disease. Thanks to a research (Decrooq et al., 2005), nine peach cultivars have been found to be tolerant to PPV: 'BlazePrince', 'Canadian', 'Harmony', 'Harken', 'June Price', 'Legend', 'Loring', 'Rosired 1', 'Springcrest' and 'Suncrest'. The observed tolerance is probably related to a quantitative source of resistance.

Even in some almond cultivars (P. dulcis), the resistance character has been found, at least as far as the PPV-D strain is concerned. This character, by genetic-cross, can be transferred to peach through interspecific hybridizations (Martínez-Gómez *et al*., 2004). In addition, polygenic resistance has been described in *P.davidiana* (Decrooq *et al*., 2005; Marandel *et al*., 2009a), a species closely related to peach [*Prunuspersica* (L.) Batsch.].

PPV Resistance in Apricot

FAO statistics estimated world apricot production at about 4 million tons. The main apricot growing areas are: China, Turkey, Uzbekistan, Algeria, the Mediterranean European countries and the United States of America. Italy is the fourth world producer of apricot, which is cultivated in Emilia Romagna, Campania, Basilicata and Sicily.

Sharka disease in apricot was found for the first time in Spain in 1984 (Llàcer *et al*., 1985). From that moment on, the disease has spread through all the country, severely affecting apricot crops because all native cultivars were susceptible to PPV. At the beginning, to narrow the problem and stop the virus spread, the eradication of the infected trees was attempted, with poor results.

To face this severe threat to the cultivations of apricot and the other species of Prunus, several genetic breeding programs were launched in Spain (Egea *et al*., 1999; Badenes e*t al*., 2002), France (Andergon, 1995) Italy (Bassi *et al*., 1995) and Greece (Kayiannis e*t al*., 1999) with the aim to introduce PPV resistance. Despite this, there is still little information about the genetics, sources and mechanisms of resistance to Sharka in apricot.

Several sources of resistance have been identified within apricot germoplasm. These are currently used to understand the genetic control of the disease and to select resistant cultivars. In particular, 'Bora', 'Harcot', 'Harlayne', 'Henderson', 'Lito', 'Stella', 'Sunglo', and 'Stark Early-Orange' (SEO) are considered reliable sources of resistance and bring introgressed resistant genes from wild accessions of Asian apricot.

In spite of the large body of literature available, the genetic basis of Sharka resistance in apricot is still under debate. Individual reports indicate that a single gene, (Dicenta *et al*., 2000), two genes (Moustafa *et al*., 2001) or three genes (Guillet – Bellanguer&Audergon 2001) are responsible.

This is because phenotyping for Sharka is still the major bottleneck in the breeding pipeline. Phenotypic evaluation of disease symptoms in segregating progenies is an expensive and time-consuming procedure and sometimes does not allow a reliable assignment to discrete classes of resistance/susceptibility.

This high degree of uncertainty in properly allocating an individual to a given class of resistance makes it difficult to analyze segregation in controlled crosses. For this reason there is uncertainty about the number of genes involved and therefore one, two or more genes are referred to in order to justify segregation data.

The study of the heritability of resistance in several large populations of hybrids for a long period of time, has permitted to determine that the resistance is controlled at least by a single dominant locus and that the resistant cultivars are heterozygous for the character (Karayiannis *et al*., 2008).

In order to exploit the knowledge of the genetic determinants of resistance to Sharka for marker-assisted selection (Dondini *et al*., 2011) within apricot genetic breeding programs, the detection of the genetic determinants of resistance to Sharka is a priority.

Mapping in stone fruit species is made easier by the *Prunus* Reference Map based on the F2 progeny from Texas (almond) x Earlygold (peach) cross T x E (Joobeur *et al.,* 1998; Aranzana *et al*., 2003; Dirlewanger *et al*., 2004; Howad *et al*., 2005). This reference map has allowed mapping several Prunus species like peach, plum and apricot, thanks to the strict colinearity of genomes of those species.

Several apricot genetic maps have been produced with the aim of mapping the genetic determinants of Sharka resistance in this species (Hurtado *et al*., 2002; Vilanova *et al*., 2003; Lambert *et al*., 2004*,* 2007; Dondini *et al*., 2007; Lalli *et al*., 2008; Soriano *et al*., 2008; Marandel *et al*., 2009a, b).

A first determinant was mapped on linkage group 1 (LG1) using an F1 progeny of 'Goldrich' x 'Valenciano' (Hurtado *et al*., 2002). 'Goldrich' is known to be tolerant to the pathogen while Valenciano was described as susceptible (Martinez-Gomez e*t al*., 2000; Dicenta e*t al*., 2000).

This result was confirmed by Soriano *et al.*, in 2008 who performed a quantitative trait locus (QTL) analysis using another F1 progeny, 'Goldrich' x 'Currot'.

A major QTL was also identified in LG1 by the analysis of F1 and F2 progenies of 'Stark Earli-Orange (SEO) (Lambert *et al*., 2004) and its offspring 'Lito' (Vilanova *et al*., 2003; Soriano *et al*., 2008). Minor QTL localized in LG3 and LG5 of both Polonais and SEO, have also been identified using the Polonais x SEO progeny (Lambert *et al*., 2007).

Some research claim that PPV resistance is fully expressed when multiple genes are active. This situation arises, for example, in *P. davidiana* in which there are six QTLs involved in controlling the disease after PPV infection and two QTLs that control virus movement inside the host plant.

Anyway, the comparative analysis of the different linkage maps of Prunus shows synteny between species for some QTL. Among these the genetic determinant which is able to explain a good fraction of the variability linked with PPV resistance is present in the first part of the linkage group 1 in the apricot cultivars 'Goldrich' and 'Lito'.

From a physiological point of view, Sharka resistance in apricot is based neither on an immunity mechanism (non-host resistance) nor a hypersensitive response (HR) triggered by recognition genes (Dangl & Jones, 2001, 2006). According to the literature, it is due either to the absence of initiating factors needed by the virus of its replication (Duprat e*t al*., 2002; Sicard *et al*., 2008; Marandel *et al*., 2009a, b) or to the inability of the virus to replicate rapidly, which does not require specific R genes.

A more recent gene expression analysis of *Plum pox virus* susceptibility/resistance in apricot shows that susceptibility to PPV in apricot is a complex process based on a continuous battle between the virus (PPV) and the plant, both at the pathogen resistance gene level (*allene oxide synthase*, the *S-adenosylmethioninesynthetase 2* and the *major MLPlike protein 423*) and gene silencing level. This was confirmed by transcriptomic differences at the gene expression level.

On the other hand, resistance to PPV in apricot is also a complex process that could involve MATH genes (Manuel Rubio *et al*., 2015). This result was confirmed through a genome-wide association study (Mariette *et al*., 2015) which speculated on two candidate genes for e PPV resistance in apricot: a BTB/POZ-MATH-TRAF-like protein and a MAPK dual-specificity phospatase. For the first gene, in Arabidopsis, another MATH-TRAF was demonstrated to control long-distance movement of potyviruses, that includes PPV, but in this case, the candidate gene is a coild-coiled MATH-TRAF-like protein (Cosson *et al*., 2010). For the second gene, members of this family are known from previous studies to play roles in pathogen resistance (Gupta *et al*., 1998). Castelló *et al*. (2010) demonstrated the role of a DNA-binding protein phosphatase, DBP1, in Arabidopsis infection by PPV. However, it does not belong to the same class of

phosphatases as the one found in apricot. The GWA study might indicate either a new role in viral susceptibility or resistance for a protein phosphatase distinct from DBP1 or that the gene found in apricot is tightly linked to the true (but still unknown) gene controlling resistance to Sharka, but further experiments are needed to test these hypothesis.

Genetic Map of 'Lito'

The genetic map of 'Lito', updated after its first publication (Dondini *et al.*, 2007), currently covers 532 cM with 161 markers separated by an average distance of 3.3 cM. This genetic map was obtained using a map population derived from the cross between 'Lito' and 'BO81604311'. 'Lito' is a genotype considered to be resistant to PPV and came from the cross 'SEO' (the donor of sharka resistance) x 'Early of Trynthos' (not resistant), while BO81604311 is a breeding line, that came from the cross 'San Castrese' x 'Reale di Imola', both susceptible to the disease.

The genomic region, which contains the resistance to Sharka, has been identified in the first part of the linkage group 1. In particular, 27 evenly spaced markers spanning 91.1cM, with an overall marker density of one marker every 3.4 cover the LG1. This reduces to 1.4 cM the distance between UDAp-463 and PaCITA5, which would represent the ends of the region harbouring the resistance to Sharka (Fig. 5).

The QTL analysis allowed identifying a major QTL peaking to the SSR UDAp-441, with a LOD score of 9.7, when plants were inoculated with the PPV-M strain. An adjacent QTL corresponded to the MA067 marker with a LOD of 16.1 when the progeny was inoculated with PPV-D. No QTL was detected in the susceptible parent 'BO81604311'.

The identification of a major QTL on LG1 of 'Lito' is in agreement with Vilanova *et al*., 2003, and Soriano *et al*., 2008, who suggested the presence in 'Lito' of a single dominant gene modulated by the activity of other minor genes.

Rubio *et al.*, 2007, and Karayiannis *et al*., 2008, also describe a single dominant gene inherited by SEO, Lito's mother, and a 1:1 segregation ratio is reported for the progeny.

The segregation ratios of resistant and susceptible genotypes are not easily calculated because of the complex behavior of the plant-pathogen interaction.

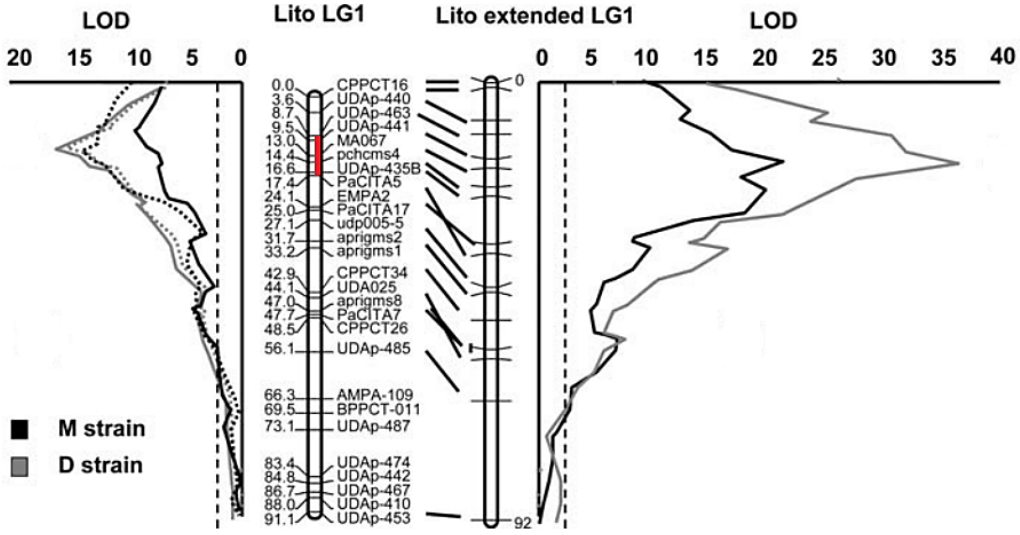


Fig. 5 - Genetic map of the linkage group 1 of cultivar 'Lito' ('Lito' from the progeny LxB; 'Lito extended' from the progeny LxB extended). The probability of the association of the markers to the "resistance trait" is indicated as LOD score (Dondini *et al*., 2010). The continuous lines indicate the QTL analysis by using the maximum level of susceptibility of the seedlings observed in 3 years of observations. The dotted lines indicate the LOD scores calculated by using the phenotypic data with recovery. Cross-hatched lines indicate the LOD threshold. The region harbouring the resistance to Sharka has been highlighted in red.

Resistant and susceptible seedlings were easily identified because of the clear-cut distinction between these two classes: resistant seedlings were always asymptomatic and negative to the ELISA test, while susceptible trees showed clear symptoms of disease and were positive to the ELISA test. However, there are also tolerant individuals that show no phenotypic symptom and are recorded as positive by the ELISA test.

Even if the tolerant and resistant seedlings are considered as one group, the segregation ratio is not 1:1, as would be expected for the segregation of a single dominant R gene heterozygous in the donor parent. A consistent bias towards an excess of susceptible individuals was always found. For this reason, the character was treated as of quantitative nature (Dondini *et al*., 2011).

The scenario was complicated by the recovery of some plants, initially classified as susceptible that became resistant or tolerant during the third year of scoring. When resistant, tolerant and such recovered seedlings were pooled all together, the ratio of this pool to susceptible plants approached 1:1.

This explains why such a large part of the phenotypic variability is accounted for by a single QTL in the LG1 while other potential QTLs, described in other studies, explain only a small part of the variability and are probably linked to other agronomic traits.

OBJECTIVE OF THIS THESIS

The national research project PRIN-VIRES started from the state of the art described above and brought together the expertise of different research centers (University of Udine, University of Bologna, University of Milan University of Bari and CNR-Institute of Plant Virology of Bari) with the aim to increase basic knowledge on the genetic control of PPV and develop tools and strategies for the control of Sharka.

In spite of the large body of literature available, the genetic basis of Sharka resistance is still under debate. This is because phenotyping for Sharka resistance is the major bottleneck in the breeding pipeline. The phenotyping protocol requires several replicates per genotype and visual inspection during two to four growing seasons, followed by ELISA and RT-PCR tests (Lommel *et al.*, 1982; Wetzel *et al*., 1991). Standardization of the resistance tests is difficult because there are several factors affecting the procedure. For instance, the response to inoculation depends on the genotype of the host, the virus strain, the time of the year when the inoculation is performed, the physiological state of the host plant and the inoculation method (Llácer *et al*. 2007).

The group I have joined is committed to the identification, isolation, and cloning of genes/QTLs for PPV resistance in apricot.

The project started from a preliminary linkage map were the resistance to PPV (strains D and M) was mapped in the linkage group 1 (LG1) of apricot using the pseudo test-cross 'Lito' x BO81604311, with 'Lito' segregating for the resistance to PPV.

Since we do not know the functions of the genes under study, the positional cloning has been exploited for the identification of resistance genes.

For this work a large population of individuals resulting from the controlled cross 'Lito' (resistant) x 'BO81604311' (susceptible) was adopted to increase the map resolution, a new set of molecular markers isolated from peach genome and scaffolds of 'Lito' sequenced at low coverage to saturate the region of resistance, and a library of BAC clones of 'Lito' to produce a minimal tiling path of the region of interest for both the resistant and the susceptible chromosomal haplotypes.

The aim was to obtain the complete sequence of the region through the Next Generation Sequencing (NGS) of BAC clones selected to produce the minimum tiling path.

The complete sequence of the region of interest was searched by aligning and ordering the BAC contigs sequences using peach as reference sequence.
Peach genome was used as reference because apricot genome has not been sequenced yet and peach was the sequence more closely related to apricot.

CHAPTER 1 - POSITIONAL CLONING OF RESISTANCE/SUSCEPTIBILITY LOCUS IN 'LITO'

INTRODUCTION

The best exploitation of different sources of resistance in breeding programs can be achieved by developing markers tightly linked to the traits of interest and using them as indicators of the presence of the trait in the progeny to be screened. Marker assisted selection (MAS) underlies this principle, and shifts from phenotype-based to genotype-based selection. Important requisites are the whole genome representation on a genetic map and as much as possible reduction of the intervals between markers.

A challenge common to all breeding programs that rely on genetic maps is the development of markers in poorly-covered regions, that is regions where several kind of markers could be under-represented. Bioinformatic tools that search for certain DNA motifs in sequenced genomes of species related to that one at the stake can easily produce large amounts of new polymorphic markers.

Typical markers that can be mined with these tools are Simple Sequence Repeats (SSRs) and Single Nucleotide Polymorphisms (SNPs). These markers have been well established in genotyping because of their ease of use and their co-dominant nature.

Positional (or map-based) cloning is a method used to discover the DNA sequences that underly a phenotypic trait, relying on its physical location along a chromosome, and without taking into account the gene function hypothesized to be responsible for the trait (Zhang *et al*., 1994).

In plants, the traditional tools used to reach the gene are segregating populations obtained by crossing the individual carrying the desired trait at heterozygous state with another homozygous for the lack of trait. Once the region of interest has been identified though a low resolution map, new, more tightly linked markers are used to reach the position where there is complete association between trait and markers. From this point, the research moves to a collection of genomic fragments, traditionally a BAC library. Inserts containing the markers selected around the QTL are kept and the cloning progresses with a chromosome walking approach until the whole region is covered with DNA sequences.

The causative factor is located within this interval, and usually the genes residing inside are tested for being responsible for the phenotype.
Following this strategy, the work to obtain the complete sequence of the region containing the PPV resistance was organized in several steps:

1. The mapping population has been extended to several hundred individuals and new molecular markers have been isolated from the region of interest, to increase the map resolution and narrow the PPV resistance region through the analysis of the recombinants.
2. A 'Lito' BAC library ((30336 clones, 10X coverage) available from a former project was screened with the molecular markers of the region to pick BAC clones which covered both the resistant and the susceptible chromosomal haplotypes.
3. Selected BAC clones were sequenced thought the Illumina NGS sequencing technology and the region of interest was assembled 'de novo'.
4. The entire assembled region was annotated, gene predicted and annotated and the candidate gene/s associated with the QTL sorted out and commented.

The first steps of the work, namely the extension of the mapping population, the development of new molecular markers and part of the BAC library screening, had been done before I started my PhD thesis and were part of my master thesis.
The above activities are briefly described in this chapter just to provide the reader with a general overview of the research plan and strategy adopted in the search of the candidate genes for the resistance to Sharka in apricot. Most activities were carried out in collaboration with the partners of the MIUR-PRIN project that funded this research. The scientific Institutions involved in the project were the University of Bologna, the University of Milano, the IGA (Applied Genomic Institute) Technology Services of Udine, and the CNR-Institute of Virology of Bari.

MATERIALS and METHODS

Map Population Extension

The map population derived from the cross between 'Lito' and 'BO81604311' has been extended from 118 to 359 individuals which have been kept under confinement. Plants are located partly in Castel San Pietro Terme (BO) and partly in Tebano (RA), and are managed by Astra Innovazione e Sviluppo srl.

DNA was extracted from leaf samples using the method described by Mercato et al. (1999) and here briefly illustrated:

1. Introduce 0,05 grams of plant material and a small amount of silicon carbide (carborundum) into an eppendorf tube (2 ml). Grind tissues using a mill set to 29 rotations per second for 3 minutes.
2. Resuspend shredded materials in 1 ml of washing buffer; centrifuge at 3000 rpm for 10 minutes; eliminate the supernatant and repeat the washing if the impurity content is still excessive.
3. Resuspend the samples in 0.64 ml of washing buffer (solutions compositions are described in Table 1) and add 0.5 ml of NaCl 5M, 0.1 ml of 10% N-lauryl sarcosine and 0.1 ml of 10% CTAB (Hexadecyl-trimethylammonium bromide). After shaking, incubate the samples at 60°C for 20 minutes.
4. Add to the samples an equal volume of dichloromethane:isoamyl alcohol (24:1) and mix the two phases to get a white emulsion.
5. Centrifuge for 5 min at 10000 rpm and transfer the upper liquid phase into a clean eppendorf tube. Add 5 µl RNase and incubate at 37°C for 30 min.
6. Repeat steps 3 and 4, then add 0.8 volumes of cold isopropanol for DNA precipitation and keep the samples at -20°C for 30 min. Centrifuge for 5 min at 10000 rpm, wash pellets using 80% ethanol.
7. Remove ethanol and resuspend dry pellets in 100 – 150µl of sterile water.
8. Quantify extracted DNA with a spectrophotometer.

Tab. 1 – Solutions used in the DNA extraction protocol.

Extraction buffer (sample volume)	
Washing buffer	0,65 ml
NaCl 5M	0,15 ml
N-lauryl-sarcosine 10%	0,1 ml
CTAB (Hexadecyl-trimethylammonium bromide) 10%	0,1 ml
Washing Buffer (for 1L of solution)	
Sodium acetate 100 mM (pH 5)	0,82 g
EDTA 20 mM	4 ml
Sorbitol 200 mM	3,64 g
PVP 420000 WT 2%	2 g
β-mercaptoethanol 1%	1 ml
Other solution	
RNasi 10 mg/ml	
Dichloromethane : isoamyl alcohol (24:1)	

Identification of new molecular markers in the region of the QTL

The primer sequences of markers GOL61 and pchcms4, located respectively upstream and downstream of the QTL with approximately a map distance of 5 cM, were projected onto the peach genome sequence V1.0, published in 2013 (Verde e*t al*., 2013), and the corresponding peach sequence was extracted.

SSR/SCAR markers were searched *in silico* using Sputnik modified software (Scalabrin, 2014 pers comm), which performed an automatic annotation of targeted core repeat sequences from the peach genome. A tentative Lito assembly was also available, although obtained with a limited coverage of NGS (Next Generation Sequencing) reads. Therefore the molecular markers identified 'in silico' on the peach genome were mapped to the Lito scaffold assemblies to verify synteny and polymorphism.

SNPs were called according to the SNP calling protocol of the IGA (Applied Genomics Institute) Technology service of Udine. The parameters set were as follows: minimum number of reads 6, minimum frequency of less frequent allele 30%.

These markers were validated and analyzed through a Sequenom platform on the 'Lito' x 'BO81604311' population. This platform performs a multiplexing genotyping using a primer extension chemistry and analyzing SNP alleles with a MALDI-TOF MS (Matrix-

Assisted Laser Desorption/Ionization Time-OF-Flight mass) spectrometry. Molecular markers were amplified on 'Lito' and 'BO81604311' and a small panel of individuals of the progeny, and those segregating in a suitable pattern were mapped in the extended population.

Genotyping and Phenotyping the extended population

Individuals of the extended population were genotyped with markers of the LG1 and the association group was rebuilt increasing the accuracy of the recombination frequency and map distances. This work was carried out by the colleagues of the University of Bologna. As a result, recombinants for the QTL region were detected. They were propagated by grafting onto GF 305 rootstock and phenotyped by the virology group of CNR of Bari according to the methodology described in Dondini *et al*., 2011.

Plant visual inspections were carried out weekly during each growth cycle and a score of 0 in absence of leaf symptoms, 1 in presence of very mild symptoms (particularly on the basal part of the plant), 2 in presence of symptoms on several leaves and throughout the plant were assigned (Fig. 6).

According to the evaluation system reported in the literature (Kegler *et al*., 1998; Dicenta *et al*., 2000) seedlings have been classified as: resistant (negative to the ELISA test and free of symptoms), tolerant (positive to the ELISA test but without symptoms) and susceptible (plants where PPV spread both on the rootstock and scion).

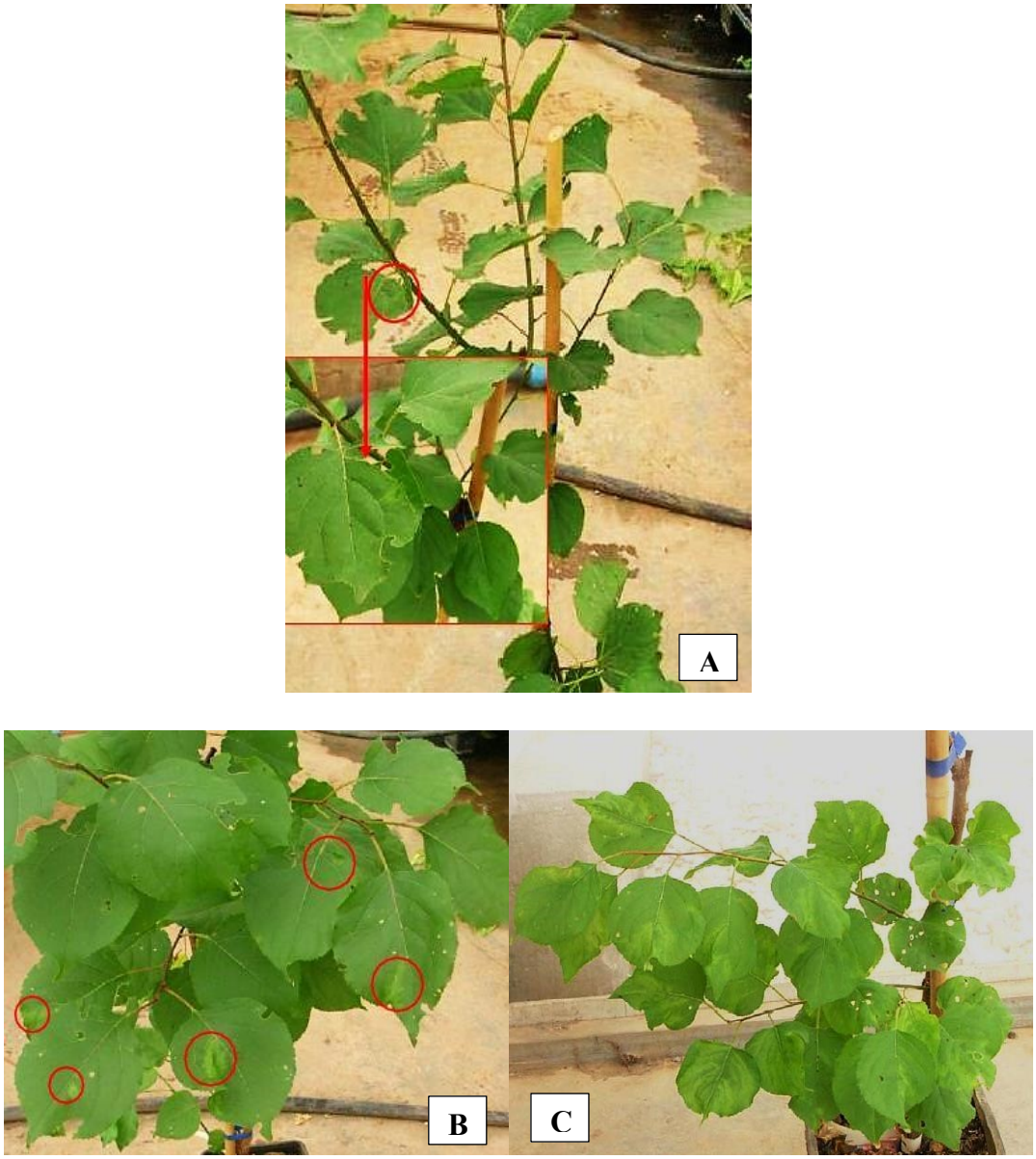


Fig. 6 – Class 1 symptoms (A), Class 2 symptoms (B,C) - photos made from the colleagues of the CNR- Institute of Virology of Bari.

BAC Library Screening

A wide-insert BAC library of 'Lito' has been commissioned to the Lucigen company, Middeleton WI, USA. The BAC library obtained through random DNA shearing included 30.336 clones with a declared size of 110/130 kb (Fig. 7) and 10X coverage.

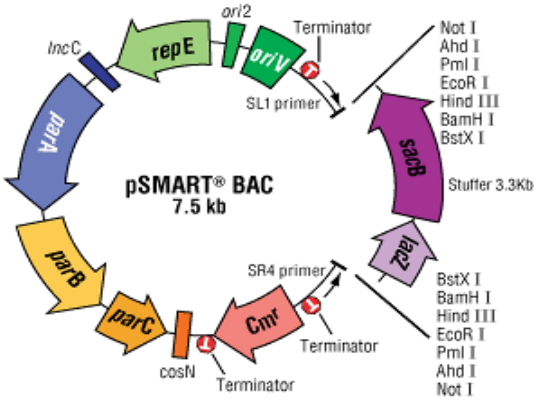


Fig.7 - pSMART BAC Vector used by Lucigen company to produce a wide-insert BAC library of 'Lito': ori2, repE, IncC - origin of replication (single copy); oriV - inducible origin of replication; par A,B,C- partition genes; Cmr - chloramphenicol resistance gene; cosN - lambda packaging signal; T – CloneSmart transcription terminators; sacB, sucrase gene; lacZ, alpha peptide portion of the beta galactosidase gene.

The BAC library was replicated at the Applied Genomics Institute (IGA) of Udine.

The screening of the BAC library was carried out through a 3-dimensional pooling strategy. Samples from each plate were pooled into 79 plate pools, samples of each row of all plates were pooled into 16 row pools and samples of each column of all plates were pooled into 24 row pools (fig 8 and 9). In such a way, a total number of 119 pools containing the whole BAC library were produced.

This 3-dimensional system allows to screen the whole set of BACs and to isolate the clones that bring in their inserts the sequence of the marker used for the screening, performing initially only 119 PCR amplifications instead of 30,336.

The conflicts resulting from the combinations of the three dimensions were solved considering all possible combinations of rows, columns and plates that gave a positive signal to the PCR and testing all resulting BACs (fig. 10).

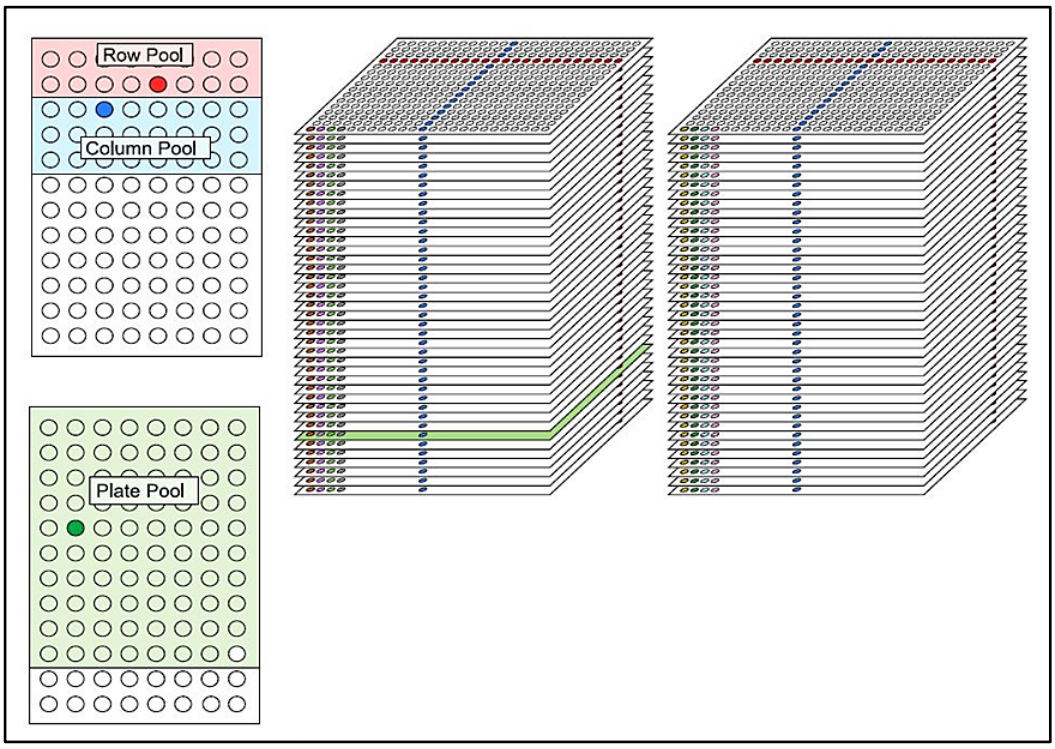


Fig. 8 – Three-dimensional pooling working scheme.

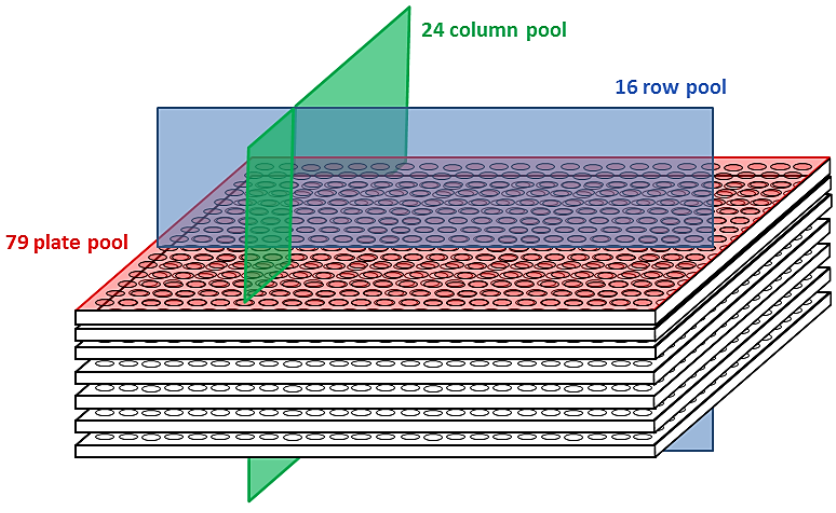


Fig. 9 – 3D - pooling diagram used to screen the BAC library of 'Lito'.

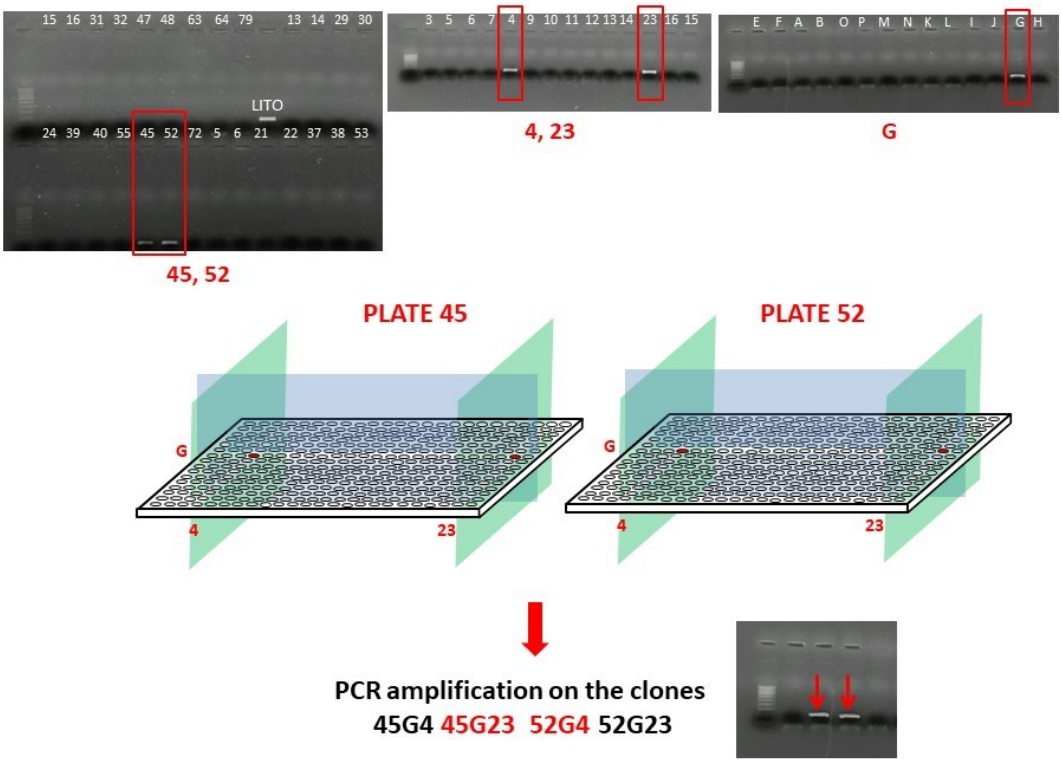


Fig. 10 – Example of BAC library screening. On the top, gel electrophoresis of the PCR amplifications of some plate pools, columns pools and row pools with specific primers for the marker S1-6798SCAR. The two plate pools (45 and 52) that provide amplifications (highlighted in red) will be the subject of subsequent analysis aiming the resolution of the conflicts resulting from the combinations of the three dimensions and the identification of the clones 45G23 and 52G4. The positive control, generated by Lito's DNA, ensures the success of the reaction.

PCRs were performed using the Master Mix (5PRIME) with the new markers primers of the QTL region. The reaction mixture and the PCR protocol are described in tables 2 and 3.

The next step is electrophoresis, a technique that separates DNA fragments based on their molecular weight. This technique allows displaying rows, columns and plates positive to the PCR. The run was performed on 1% agarose gel (tab. 4).

Being Lito heterozygous for the markers tested, BAC clones were assigned to the corresponding resistant or susceptible haplotype. This, was done according to the alleles of the marker analysed carried by 'Lito' in coupling or repulsion with the resistance.
The colleagues of the University of Bologna carried out this work in largest part.

Tab. 2 – PCR reaction mixture for BAC library screening.

Solution	Sample volume (µl)
Water	13,6
Buffer 10X	2
Primer F	0,5
Primer R	0,5
DNTPs	0,2
Taq polymerase	0,2
Total volume	20

Tab. 3 – Thermocycling conditions for PCR.

	PCR AMPLIFICATION PROTOCOL		
	Step	T (°C)	Time
	Initial Denaturation	94	2'
× 35	Annealing	58	10"
	Extension	68	30"
	Denaturation	94	20"
	Final Extension	68	5'
	Hold	10	∞

Tab. 4 – Electrophoresis solution composition.

AGAROSE GEL ELECTROPHORESIS	
TBE 5X (1 L)	
Tris (0,89 M)	108 g
Boric acid (0,89 M)	55 gr
EDTA 0,02M (pH 8,3)	1903 gr

DNA extraction from positive BACs

Single clones were grown first on solid medium LB + agar at 37 °C for 12 h and then on liquid culture using Multitron:

- Pre-inoculation in TB + Chl (50 mg / ml) at 37° C for 12h under stirring at 320 rpm.
- Inoculation in TB + Chl (50 mg / ml) at 37° C for 20h under stirring at 320 rpm.

Inoculation was carried out automatically using "Biomek fx" (Beckman Coulter), a robot capable of transferring 5 µl pre-inoculum on the plates. Culture aliquots were added with glycerol and stored at -80°C.

Selected BAC clones were grown on four plates of 394 wells. Each plate contained four replicas of the BACs.

Mini preps were performed on the grown cultures using an alkaline lysis protocol. Bacterial cultures were centrifuged and the liquid medium was discarded.

After -20°C storage for one hour, bacterial cells were resuspended in 50 mM Tris-HCl (pH=8) and 10mM EDTA supplemented with 100 µg/ml of Rnase. Lysis of bacterial cells was performed for 10 min under mild stirring in 0,2 M NaOH 1% SDS and blocked using 3 M Potassium acetate solution (pH=5,5 - 4°C). Crude lysate was incubated for 10 minutes in wet ice. Sedimentation of cell debris was obtained by centrifugation.

Plasmid DNA in the supernatant was precipitated with isopropanol, rinsed in 70% ethanol and resuspended in water.

DNA quantification was performed using both a Nanodrop ND-1000 spectrophotometer and a Qubit® 2.0 Fluorometer.

BAC ends Sanger sequencing

Nine µl of plasmid DNA (200-300 ng / µl) were used as a template for sequencing (tab. 5). The sequencing PCR reaction contained 1X Sequencing Buffer (Applied Biosystem), BigDye® Terminator v.3.1 (Applied Biodystems) and the specific primers of pSMART BAC vector:

SL1: 5'– CAGTCCAGTTACGCTGGAGTC–3';

SR4: 5'–TTGACCATGTTGGTATGATTT–3';

Sequencing PCR conditions were 96°C for 10'', 50 °C for 5'', 60 °C for 4', for 99 cycles. The produced sequences were aligned against the peach genome reference by BLAST (http://services.appliedgenomics.org/blast/prunus/).

Tab. 5 – Sequencing PCR reaction Mix.

Solution	Sample volume (µl)
Water	1,226
Buffer 5X	2,47
Primer F/R (100 µM)	0,044
Big Dye	0,26

Paired – end libraries and mate – pair libraries

Paired – end libraries were obtained starting from the DNA derived from the union of 8 replicas of each BAC clone extracted with miniprep protocol and a concentration of 350 ng of DNA in 100 µl.

BAC clones were extracted in two different times. The first 34 BAC clones were used to construct paired-end libreries following Nextera DNA Sample preparation protocol and run on an Illumina HiSeq2000.

The second group of BAC clones was used to construct paired – end libraries following ThruPLEX® DNA-seq Quick Protocol, Dual Indexes and run on an Illumina MiSeq.

Mate-pair libraries of resistant and susceptible BAC clone pools were prepared starting from a concentration of 12 ng/µl and 19 ng/µl respectively.

Mate-pair libraries were constructed using Nextera Mate Pair Gel-Free Sample Preparation protocol and run on an Illumina Hiseq2000.

NGS sequencing and *de novo* assembly of BAC clones

Selected BAC clones were sequenced in pool with HiSeq Illumina technology and MiSeq Illumina technology using a tagging system of individual clones.

Mate pairs of resistant and susceptible BAC clone pools were sequenced with HiSeq Illumina technology.
The Institute of Applied Genomics (IGA) of Udine was charged for sequencing.
Both pair-end and mate-pair raw reads were searched for possible contaminants with bbduk2 (https://github.com/BioInfoTools/BBMap/blob/master/sh/bbduk2.sh).
Reads were then cleaned masking possible residuals of adapter sequences using cutadapt (https://cutadapt.readthedocs.io/en/stable/), trimmed by quality and filtered by possible contaminants using ERNE-FILTER (erne-soruceforge.net).
Reads of each BAC clone were assembled separately with CLC Genomics Workbench v3 using a *de novo* paired – end assembly algorithm.

***De novo* assembly of the region containing the QTL for PPV resistance**

The sequences of each BAC clones were aligned against Peach genome (http://www.rosaceae.org/species/prunus_persica/genome_v1.0) with the aim to order the sequences of each BAC contigs and reconstruct the entire region for the two haplotypes (resistant and susceptible). Peach genome was used as reference because apricot genome assembly was not available yet and peach genome is the sequence more closely related to apricot among those available. BLASTn and GEvo comparative sequence alignment tool (https://genomevolution.org/Coge/Gevo.pl) were used to identify shared regions between apricot and peach genome.
The peach genome served initially as a good-guideline but the contig order was not solved in apricot regions lacking collinearity with the peach genome.
For this reason, in a second time, the order of the BAC contigs was solved aligning one BAC clone against each other using Dotter tool (http://www.sanger.ac.uk/science/tools/seqtools), regardless of the peach genome. Dotter tool is part of SeqTools package and is a graphical dot-matrix program for detailed comparison of two sequences.
Overlapped sequences have been assembled with iAssembler-v1.3.2 software (http://bioinfo.bti.cornell.edu/tool/iAssembler/), setting the minimum overlap length at 100 and the minimum percent identify at 99 for sequence clustering and assembly.

The mate pair reads of the resistant and susceptible BAC pools were aligned respectively against the assembled sequences using BWA. This allowed to check the order and direction (forward or reverse) of the assembled supercontigs for the two haplotypic regions. The alignments were visualized using Tablet (https://ics.hutton.ac.uk/tablet/).

RESULTS

Identification of new molecular markers in the region of the QTL

From scaffolds of 'Lito' sequenced at low coverage, 39 SNPs were identified. Primers used for the single base primer extension analysis are reported in table 1 of the supplementary materials. SNPs useful to saturate the region under study were mapped on the LG1.

The analysis of peach genome V.1 sequence between the two markers, GOL61 and pchcms4, flanking the QTL region allowed the identification of 17 new molecular markers (SCAR/SSR).

The list of all new molecular markers of the region concerned by QTL is reported in tables 7 and 8. These data were provided by the colleagues of the University of Bologna.

Tab. 7 – List of new SSR/SCAR molecular markers identified in the region of interest and their position along reference peach genome V.1.

SSR/SCAR	Position
S1-6798SCAR	6,798,000
S1-6835SCAR	6,835,500
S1_6994SCAR	6,994,315
S1-7045SSR	7,045,894
S1-7164SSR	7,164,204
S1-7186SCAR	7,186,317
S1-7218SCAR	7,217,957
S1-7284SSR	7,284,822
S1-7361SSR	7,361,871
S1-7418SSR	7,418,382
S1-7484SSR	7,484,404
S1-7518SCAR	7,518,000
S1-7700SSR	7,700,369
S1-7745SSR	7,745,510
S1-7982SSR	7,982,484
S1-8060SCAR	8,060,402
S1-8109SSR	8,109,407

Tab. 8 – List of new SNP molecular markers identified in the region of interest and their position along reference peach genome V.1.

SNP	Position
s1_5511078	5,511,078
s1_5540944	5,540,944
s1_5586095	5,586,095
s1_5616242	5,616,242
s1_5683686	5,683,686
s1_5801422	5,801,422
s1_5820311	5,820,311
s1_5864975	5,864,975
s1_5878914	5,878,914
s1_6127634	6,127,634
s1_6180800	6,180,800
s1_6223532	6,223,532
s1_6280616	6,280,616
s1_6345556	6,345,556
s1_6422355	6,422,355
s1_6537106	6,537,106
s1_6698823	6,698,823
s1_6761324	6,761,324
s1_6828246	6,828,246
s1_6965213	6,965,213
s1_7077554	7,077,554
s1_7112235	7,112,235
s1_7217828	7,217,828
s1_7241764	7,241,764
s1_7267206	7,267,206
s1_7316247	7,316,247
s1_7442314	7,442,314
s1_7473604	7,473,604
s1_7505322	7,505,322
s1_7526901	7,526,901
s1_7555803	7,555,803
s1_7579835	7,579,835
s1_7786880	7,786,880
s1_7805039	7,805,039
s1_7960013	7,960,013
s1_7983920	7,983,920
s1_8042406	8,042,406
s1_8105025	8,105,025
s1_8132703	8,132,703

New genetic map of the linkage group 1 of 'Lito'

A new genetic map was constructed for the linkage group 1 using the individuals of the extended population (Fig. 11). This map includes 48 molecular markers and covers 90.8 cM length. The average distance between molecular markers is 2.67 cM which is reduced to 1,04 cM in the genomic region dealing with the resistance. QTL analysis using an 'interval mapping' approach confirmed the presence of a major QTL around the marker MA067, as previously reported in Dondini *et al.* (2011). These data were provided by the colleagues of the University of Bologna.

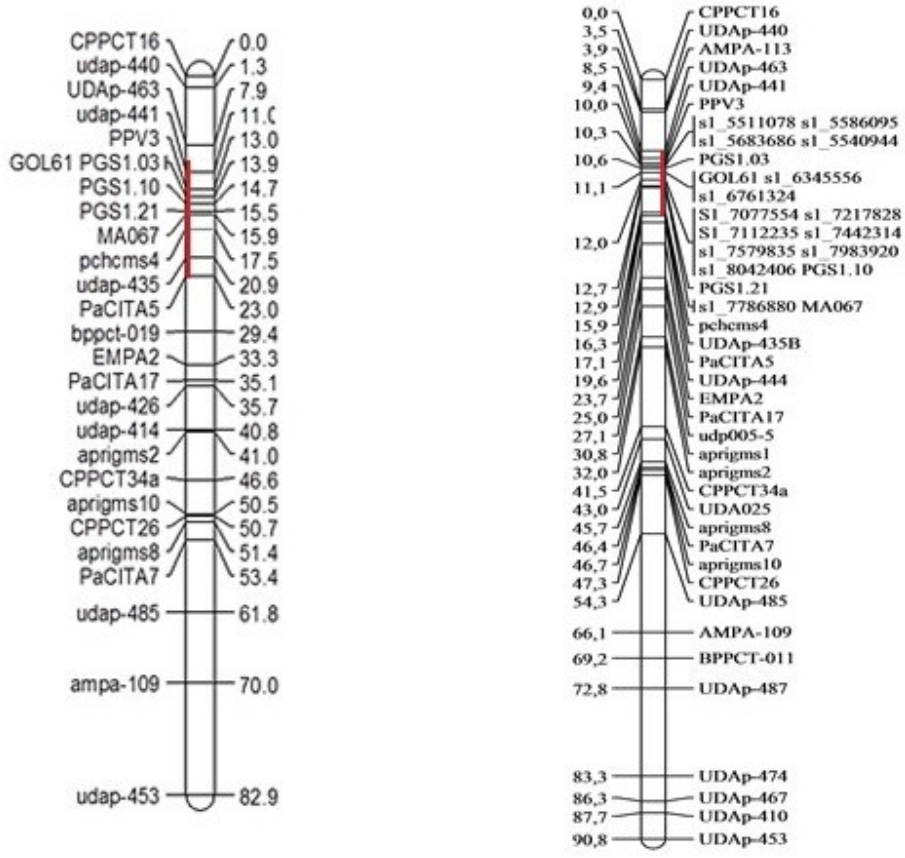


Fig. 11 – On the left the old LG1 linkage map of 'Lito' (Dondini *et* al., 2011), on the right the new LG1 linkage map obtained using extended population 'Lito' x 'BO81604311' with new markers mapped to saturate the region harbouring the resistance to Sharka (in red).

Recombinants phenotyping

Genotyping of individuals with all markers of the region allowed the identification of 18 recombinants in the region of interest. Recombinants were phenotyped by the Virology Group of the CNR of Bari, partner of the project.

The analysis of the offsprings that recombined within the region of interest provided a complex landscape, whose interpretation requires a genetic hypothesis a little different compared to the initial one, based on a single QTL that operates in heterozygosity with a dominant R allele.

We considered resistant the individuals with the score 0 at the phenotypical evaluation following artificial infection, while individuals with score 1 oe 2 were considered susceptible (table 12). The resistant individuals, namely E156, E104 and E124, would restrict the region of resistance in the linkage map from the marker S1_7077554 at the top to the markers PGS1-24 at the bottom, with both markers included. Unexpectedly, there are individuals with the same resistant haplotypic region, such as E029, 17, and E191 that resulted being susceptible when challenged with the PPV virus.

The conflicting results for these latter recombinants should postulate the presence of a second locus that control the resistance to PPV in apricot and would suggest the hypothesis that the QTL/gene at the stake should be condition necessary and not sufficient for the genotypes deploy resistance. This hypothesis has been already speculated by authors who conducted a meta-analysis of segregation of PPV resistance in many apricot crosses (Marandel *et al*., 2009) and the Decrocq's group suggested the presence of a second locus 1-2 Mb below the first one and called the two loci *GWAS-PPB1a* and *GWAS-PPV1b* respectively with epistasis between the two loci (Mariette *et al*. 2016). Alternatively the second locus would reside in another chromosome as it has been already speculated (Lambert *et al*. 2007; Pilarova *et al*. 2010; Mariette *et al.* 2016). These conclusions refer only to 'Lito' and the other resistant cultivars studied up to now, that are genetically related to each other (Zhebentyayeva *et al.* 2008). It is possible that other QTLs of resistance exist in the world apricot germplasm (Mariette *et al.* 2016).

There is one final point that would deserve attention. Several susceptible individuals, scored initially as susceptible, recovered and the third year did no longer show symthoms. The recovery is a little studied phenomenon and would likely require a different control of such a delayed ‘tolerance’ to sharka. Because of this, in this work the phenomenon was disregarded but it might merit reconsideration.

Marker	segregation	E156	E019	31	E181	E189	E220	E257	77	E095	E182	63	89	99	E024	E104	27	E214	E191
CPPCT16	<a**b**xaa>	ab	ab	ab	ab	ab	ab	ab	ab	ab	ab	aa	aa	aa	aa	ab	aa	aa	aa
AMPA-113	<a**b**xaa>	ab	ab	ab	ab	ab	ab	ab	ab	ab	ab	aa	aa	aa	aa	ab	aa	aa	aa
UDAp-440	<a**b**xaa>	ab	ab	ab	ab	ab	ab	ab	ab	ab	ab	aa	aa	aa	aa	ab	aa	aa	aa
UDAp-463	<a**b**xaa>	ab	ab	ab	ab	ab	ab	ab	ab	ab	ab	aa	aa	aa	aa	aa	aa	aa	aa
UDAp-441	<a**b**xaa>	ab	ab	ab	ab	ab	ab	ab	ab	ab	ab	aa	aa	aa	aa	aa	aa	aa	aa
s1_5511078	<a**b**xaa>	ab	ab	ab	ab	ab	ab	ab	ab	ab	ab	aa	aa	aa	aa	aa	aa	aa	aa
s1_5540944	<a**b**xaa>	ab	ab	ab	ab	ab	ab	ab	ab	ab	ab	aa	aa	aa	aa	aa	aa	aa	aa
s1_5586095	<a**b**xaa>	ab	ab	ab	ab	ab	ab	ab	ab	ab	ab	aa	aa	aa	aa	aa	aa	aa	aa
s1_5683686	<a**b**xaa>	ab	ab	ab	ab	ab	ab	ab	ab	ab	ab	aa	aa	aa	aa	aa	aa	aa	ab
PPV3	<a**b**xac>	ab	ab	ab	ab	ab	ab	ab	aa	aa	aa	aa	aa	aa	aa	aa	aa	aa	ab
s1_6345556	<a**b**xaa>	ab	ab	ab	ab	ab	ab	ab	aa	aa	aa	aa	aa	aa	aa	aa	aa	aa	ab
GOL61	<a**b**xaa>	ab	ab	ab	ab	ab	ab	aa	aa	aa	aa	aa	aa	aa	aa	aa	aa	ab	ab
s1_6761324	<a**b**xaa>	ab	ab	ab	ab	ab	ab	aa	aa	aa	aa	aa	aa	aa	aa	aa	aa	ab	ab
PGS1.03	<a**b**xac>	ab	ab	ab	ab	ab	ab	aa	aa	aa	aa	aa	aa	aa	aa	aa	aa	ab	ab
S1_7077554	<a**b**xaa>	ab	ab	ab	ab	ab	ab	aa	aa	aa	aa	aa	aa	aa	aa	ab	ab	ab	ab
S1_7112235	<a**b**xaa>	ab	ab	ab	ab	ab	ab	aa	aa	aa	aa	aa	aa	aa	aa	ab	ab	ab	ab
s1_7217828	<a**b**xaa>	ab	ab	ab	ab	ab	ab	aa	aa	aa	aa	aa	aa	aa	aa	ab	ab	ab	ab
PGS1.10	<a**b**xaa>	ab	ab	ab	ab	ab	aa	aa	aa	aa	aa	aa	aa	aa	aa	ab	ab	ab	ab
s1_7983920	<a**b**xaa>	ab	ab	ab	aa	aa	aa	aa	aa	aa	aa	aa	aa	aa	aa	ab	ab	ab	ab
s1_8042406	<a**b**xaa>	ab	ab	ab	aa	aa	aa	aa	aa	aa	aa	aa	aa	aa	aa	ab	ab	ab	ab
s1_7786880	<a**b**xaa>	ab	ab	aa	aa	aa	aa	aa	aa	aa	aa	aa	aa	aa	aa	ab	ab	ab	ab
PGS1-21	<a**b**xac>	ab	ab	aa	aa	aa	aa	aa	aa	aa	aa	aa	aa	aa	aa	ab	ab	ab	ab
PPB	<a**b**xaa>	ab	ab	aa	aa	aa	aa	aa	aa	aa	aa	aa	aa	aa	ab	ab	ab	ab	ab
MA067	<a**b**xaa>	ab	ab	aa	aa	aa	aa	aa	aa	aa	aa	aa	aa	aa	ab	ab	ab	ab	ab
PGS1-24	<a**b**xac>	ab	ab	aa	aa	aa	aa	aa	aa	aa	aa	aa	aa	aa	ab	ab	ab	ab	ab
pchcms4	<a**b**xac>	aa	aa	aa	aa	aa	aa	aa	aa	aa	aa	ab	ab	ab	ab	ab	ab	ab	ab
UDAp-435B	<a**b**xaa>	aa	aa	aa	aa	aa	aa	aa	aa	aa	ab	ab	ab	ab	ab	ab	ab	ab	ab
PaCITA5	<a**b**xaa>	aa	aa	aa	aa	aa	aa	aa	aa	aa	ab	ab	ab	ab	ab	ab	ab	ab	ab
Phenotypic class		R	S	S	S	S	S	S	S	S	S	S	S	S	S	R	S	R	S
Phenotypic score		0	2	1	2	2	2	2	2	2	2	2	2	2	1	0	2	0	2

Fig. 12 - Recombinants of the 'putative' Sharka resistance region. Molecular marker are shown on the first colomn of the table, recombinants on the third row. Background red color means that the genotype carry the allele associated to the resistance in 'Lito'. Background green color identify markers carrying the allele associated to the susceptibility in the 'Lito' haplotype.

BAC library screening

The BAC genomic library of 'Lito' was screened with all markers reported in table 2 (supplementary materials) and primers developed on the BAC ends as well (tab. 3, supplementary materials).
This screening allowed to identify about a hundred of positive clones. Among these, 56 BAC clones covering the region of interest between 6,2 Mbp and 8,3 Mbp of the LG1 (the coordinates are relative to Peach genome V.1) were selected.
Since 'Lito' is heterozygous, BAC clones were divided into the two haplotypes, resistant and susceptible. This, was done according to the alleles carried by the markers used for the screening. In particular, 26 clones that covered the region of the resistant chromosome and 30 clones that covered the region of the susceptible one were selected.

Sequencing and de novo assembly of 'Lito' BAC clones

Fifty-six BAC clones were sequenced and paired-end Hiseq and MiSeq Illumina reads were first checked for contaminants and trimmed by quality, then assembled into sequence contigs.
Genomic *E.coli* contamination level was high for the second round of BAC clone presumably due to an error during the extraction protocol, but the quantity of reads produced was very high so it was still possible to assembly the BAC clones. In the case of the clones 5J1 and 41A1 the contamination level was very high, and these two BAC clones were discarded. Statistics about paired-end Illumina reads for each BAC clones, the number of good bases, the assembled bases and the number of contig sequences for each BAC clones are reported in table 9.

Tab. 9 – Paired-end Illumina reads statistics and the number of contig assembled for each BAC clones.

CLONE	Raw reads	Quality trimmed reads	Contaminated reads	Good bases	Bases assembled	Number of sequences
14B18	511,534	484,274	7,782	96,334,269	125,825	9
16K16	452,116	85,348	2,918	17,532,696	105,063	10
19F18	461,642	728,412	12,378	150,376,102	120,911	8
28P4	488,352	145,688	3,422	29,979,028	121,170	6
30M18	542,254	888,458	12,730	180,321,874	160,188	8
33B17	463,450	321,366	4,034	65,731,579	113,049	9
35I18	521,224	738,052	17,312	139,111,197	80,618	7
36C4	789,402	279,816	3,082	57,921,809	113,661	23
36E17	472,168	826,748	8,800	165,468,177	103,952	9
37M10	535,994	822,562	7,556	162,001,312	134,623	19
39E10	575,374	1,440,264	13,324	279,857,715	86,353	11
3L8	501,390	1,015,572	4,518	200,205,894	108,958	2
40A13	503,302	801,468	11,766	153,285,355	112,489	5
40P21	467,480	996,954	12,190	190,492,685	147,380	34
41I23	537,080	984,786	10,306	194,502,641	66,958	1
45G23	512,892	752,154	6,060	143,312,453	132,464	9
47M3	830,888	674,228	7,966	134,300,643	66,635	1
50G17	833,948	543,376	7,490	105,886,393	111,906	20
54E7	1,050,628	223,896	4,390	47,314,149	110,531	1
55E16	869,450	779,596	7,142	159,699,352	139,648	7
55P1	788,514	527,538	8,648	104,743,193	41,951	12
9D2	869,022	896,968	16,844	177,035,083	60,920	10

Tab. 9 – Paired-end Illumina reads statistics and the number of contig assembled for each BAC clones (continue).

CLONE	Raw reads	Quality trimmed reads	Contaminated reads	Good bases	Bases assembled	Number of sequences
60L21	1,000,668	739,326	13,264	142,425,774	89,356	8
62H9	972,036	1,014,210	12,722	198,017,590	74,327	3
66N22	728,026	363,308	5,062	73,621,175	66,159	21
6E20	961,844	944,628	14,046	178,164,612	94,358	3
6F3	948,942	1,380,344	13,186	258,567,072	88,586	7
70N14	1,022,926	1,261,722	12,840	227,385,529	78,521	15
71E15	898,534	465,620	7,426	95,828,159	91,777	3
71O15	785,770	944,564	19,124	182,520,945	55,117	3
72O22	723,638	1,706,688	8,936	340,174,775	135,213	33
73D20	684,380	297,636	10,786	59,800,941	56,426	7
73M21	771,036	1,086,046	5,914	218,984,156	91,154	24
78D22	824,054	1,696,836	39,768	322,466,548	90,949	28
5J1	223,900	223,276	221,593	147,620	0	0
7H1	355,784	354,683	261,623	23,780,274	87,510	7
7H5	365,624	364,477	289,507	19,215,742	110,408	4
26O17	280,236	279,527	200,464	20,339,621	95,747	1
30J17	352,050	350,890	285,375	16,460,688	155,480	7
31O22	535,228	534,016	394,555	36,480,991	99,095	14
36A21	239,506	238,888	193,325	11,612,973	100,696	28
40P23	331,122	330,023	253,083	19,575,888	59,908	4
41A1	244,810	243,916	241,807	130,221	0	0
45F7	162,816	162,160	129,401	8,165,766	84,371	16

Tab. 9 – Paired-end Illumina reads statistics and the number of contig assembled for each BAC clones (continue).

CLONE	Raw reads	Quality trimmed reads	Contaminated reads	Good bases	Bases assembled	Number of sequences
47K13	174,022	173,407	132,330	10,412,404	127,165	3
52J18	350,356	349,419	275,818	19,013,024	227,517	31
54P16	147,030	146,577	114,546	8,168,373	214,311	67
54L21	246,090	244,983	193,941	13,180,172	128,430	8
57A6	415,114	414,305	323,786	23,583,516	97,683	4
57A1	82,166	81,917	74,330	1,898,954	129,418	16
57E4	163,980	163,536	125,325	9,920,674	110,531	1
58N7	527,620	526,162	426,151	25,886,780	199,601	39
60M22	403,354	402,194	321,043	21,257,237	56,874	1
63O4	830,536	828,235	611,818	56,413,571	101,446	10
66A5	542,912	541,006	495,276	9,666,450	81,113	18

De novo assembly of the region carrying the PPV resistance in the LG1 of apricot

At the beginning, the information on the relative position of the molecular markers developed within Peach genome (http://www.rosaceae.org/species/prunus_persica/genome_v1.0) and the peach genome sequence assisted the ordering of the contig sequences of each apricot BAC clone. The aim was to determine the relative position and orientation of all BACs contigs and the reconstruction of the whole region of apricot genome for the resistant and susceptible haplotypes exploiting the collinearity of apricot sequence with the corresponding LG1 peach sequence from 6.2 to 8.3 Mbp.

The peach genome served initially as a good guideline, but lack of co-linearity between apricot and peach genome was found in several points of the region of interest and, in many cases, BAC clone sequences assembled with CLC turned out to be fragmented in several contigs because of the presence of several repeated sequences and this hindered solving the order and orientation of a number of sequences.

The order of BAC clone contigs was solved aligning the BAC clones against each other using Dot Plot, regardless of the peach genome. This similarity matrix can show the overlap between two sequences from the number and length of matching segments shown in the matrix.

Dot plots compare two sequences by organizing one sequence on the x-axis and another on the y-axis of a plot. Once the dots have been plotted, they will combine to form lines. The closeness of the sequences in similarity will determine how close the diagonal line is. This relationship is affected by certain sequence features such as frame shifts, direct and inverted repeats. Frame shifts include insertions, deletions, and mutations. The presence of one of these features, or the presence of multiple features, will cause for multiple lines to be plotted in different possible configurations, depending on the features present in the sequences.

In figure 13 is shown an example of the alignment work that was done between the BACs 6E20 and 47K13. Both of these BACs are made of three contig sequences.

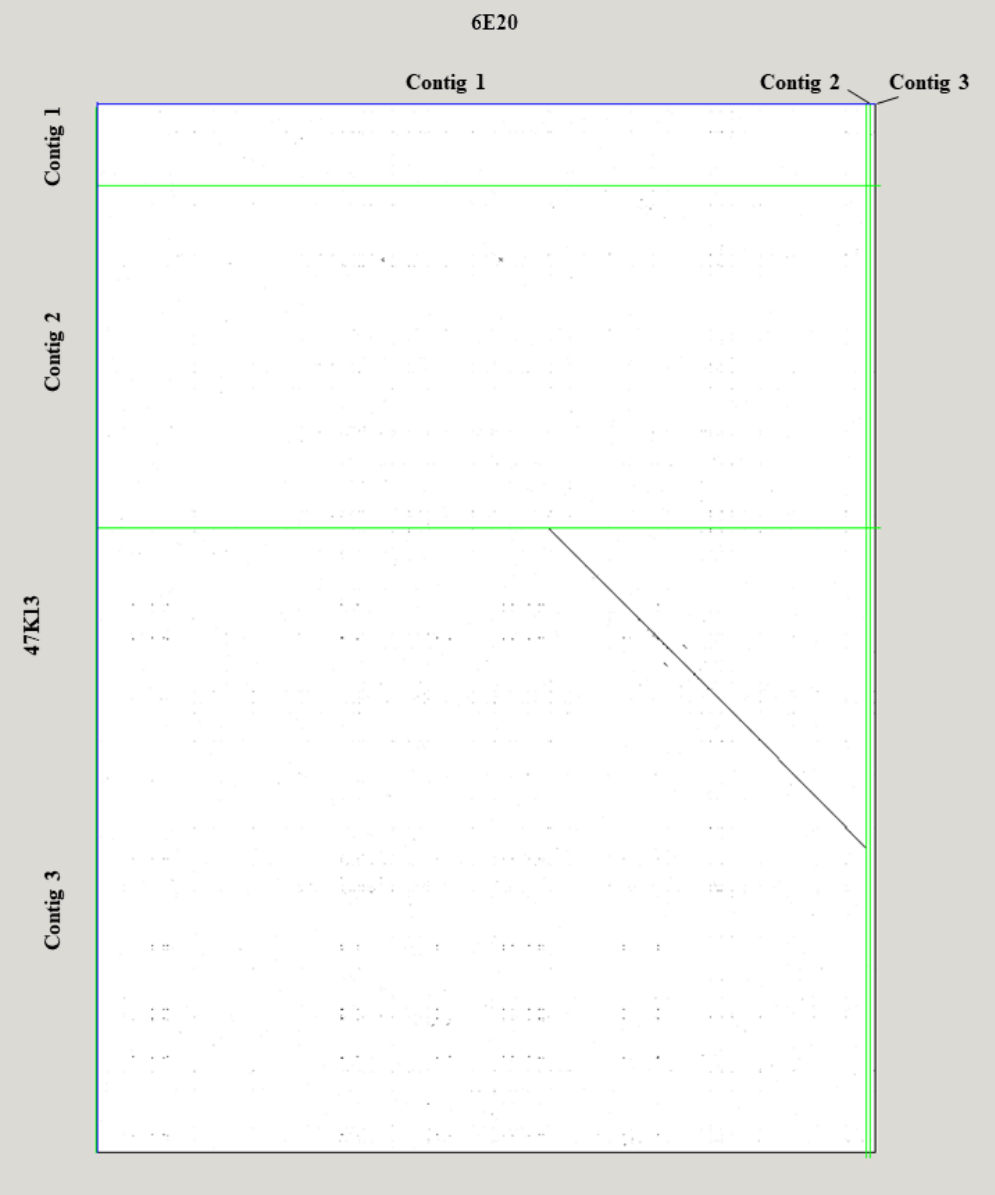


Fig. 13 – Example of dotplot alignment between two BAC clones. On x-axis the BAC clone 6E20, on y-axis the BAC clone 47K13. Green lines represent the start and end of the contig of each BACs. Diagonal lines represent the share sequence by the two BACs.

In this case, contigs 1 and 2 of 47K13 (y-axis) and 2 and 3 of 6E20 (x-axis) do not show regions with similarity. Instead, the final part of contig 1 of 6E20 is shared with the first part of contig 3 of 47K13, as shown by the diagonal line. From this plot it is possible to deduce and reconstruct the order of the sequences putting contigs 3 and 2 of 6E20 before

the first contig that shares the final part with the contig 3 of 47K13, and contigs 1 and 2 of 47K13 after this last one, as it's shown in the figure 14 below.

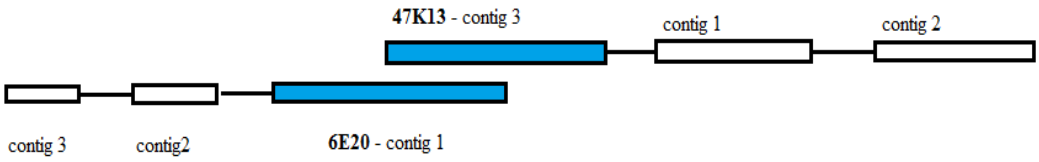


Fig. 14 – Graphical representation of the contigs order deduced from dotplot.

This work was done for all sequenced BACs, for both resistant and susceptible haplotypes. After the ordering of BAC sequences, overlapped sequences were assembled with iAssembler software. However, the assembly of the BAC supercontigs was hampered by the presence of several repeated region and in some cases supercontigs were still fragmented in several contigs of which order and orientation could not be solved.

To solve this problem and improve the assembly, all BACs were grown and DNA was extracted as described in material and methods, then BACs were pooled into two distinct pools, one containing the resistant BACs and the other the susceptible ones and mate-pair Hiseq Illumina reads from both pools were produced.

The mate-pair reads were aligned against the supercontigs and the correct order and orientation were verified by observing whether or not they aligned with the expected distance and orientation.

The alignment and assembly of the BAC clone sequences allowed to obtain the physical map of the region for the two haplotypes (resistant and susceptible). In particular, six supercontigs covering the susceptible region and five supercontigs for the resistant one were reconstructed.

The abundance of the repetitive fraction in this region hampered the assembly of the BAC supercontigs which remained fragmented in several contigs and despite the efforts and the alignment of mate-pair reads against the assembled supercontigs, the contigs order, in particular the smallest ones, has not been thoroughly solved. The number of contigs which

make up each supercontig sequences and the total length of these sequences are reported in table 10 for the resistant haplotype and in table 11 for the susceptible one. The gaps between contigs were replaced with 500 'N' characters.

Tab. 10 – Statistics for the supercontigs for the resistant haplotype.

SUPERCONTIG	Number of contig	Total length
52J18-54P16-57N7-58N7-36A21-40P21	30	256,767
31O22-60L21-66A5-50G17-6F3	25	227,866
66N22-55P1-45G23-59D2-16K16-70N14-40A13-35I18-72O22-28P4-7H1	69	692,568
41I23	1	66,958
37M10-47M3	7	151,245
ChrR - region	132	1,395,404

Tab. 11 – Statistics for the supercontigs for the susceptible haplotype

SUPERCONTIG	Number of contig	Total length
30J17	7	155,480
45F7	16	87,371
54E7	23	82,267
78D22	28	90,949
36E17-54L21-73M21-71E15-71O15-33B17-14B18-39E10-30M18-26O17-19F18-6E20-57A6-40P23-47K13	39	834,395
36C4-60O4-7H5-57A1-73D20-3L8-60M22-62H9-55E16	37	381,462
ChrS - region	150	1,631,924

Physical map of the resistant/susceptible region

The two figures below (fig 15, fig 16) represent the physical map of the region of interest between the molecular markers s1_6345556 and PGS1-24 in the LG1 of 'Lito' which is approx. 2 Mbs in length. These two physical maps (in green the resistant haplotype and blue the susceptible one) were obtained aligning BAC supercontigs against peach genome used as reference. They show the position of 'Lito' BAC sequences in the corresponding chromosome 1 of peach from 6,2 to 8,4 Mbp (coordinates are relative to Peach genome

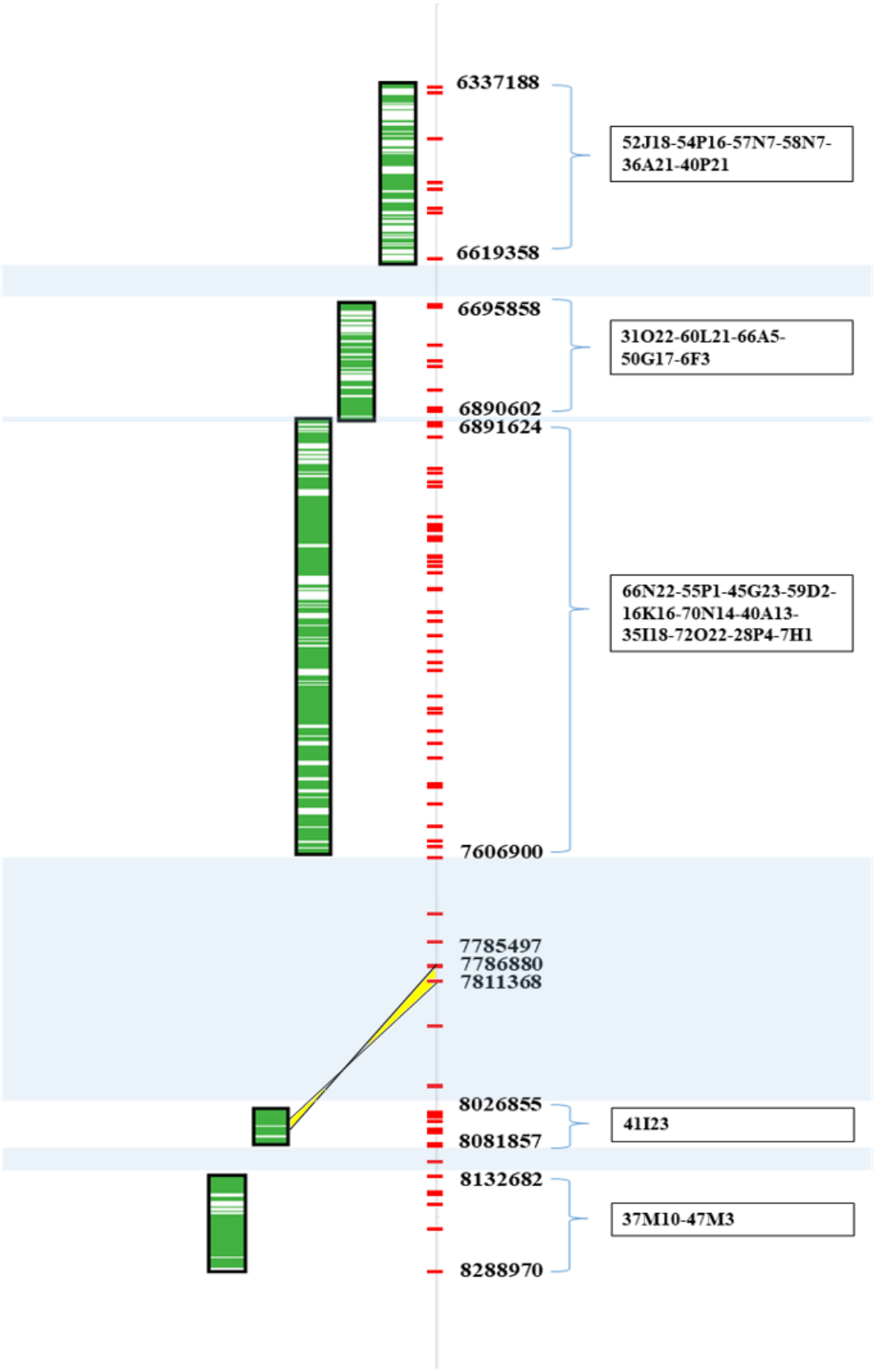
V.1). Markers from the genetic map of the region of interest are reported in red, according their physical distance in peach genome.

Peach genome acts as guideline and allows to understand how much of the sequence for the two haplotypes, the resistant and the susceptible one, has been reconstructed.

Approx. 1,6 Mbp of the susceptible and 1,4 Mbp of the resistant 'Lito' haplotypes were assembled but there are still gaps. The uncovered peach sequence are highlighted with light-blue background. In particular on the resistant region, three small gaps (6,695,858 – 6,619,358 bp, 6,891,624 – 6,890,602 bp, 8,132,682 – 8,081,857 bp) and a large gap from 7,6 Mbp to 8,0 Mbp still remain in the assembly. While, the susceptible region still have five gaps (6,497,763 – 6,507,974 bp, 6,594,101 – 6,606,102 bp, 6,675,774 – 6,727,267 bp, 6,773,413 – 6,854,605 bp, 7,749,801 – 7,939,359 bp).

The size of these gaps in the assembly can be only estimated thanks to the peach genome, but they cannot be safely determined because the apricot genome could be structurally different from the peach genome. Unfortunately, the molecular markers did not allow picking new BAC clones from the library to cover these gaps, and the design of new primers on the supercontigs ends was hampered by the presence of repeats in those regions.

This work permitted to highlight a possible inversion found in apricot chromosome with respect to the peach genome from 7,785,497 to 7,811,368 bp. This inversion is shown in yellow in both physical maps.


6337188
52J18-54P16-57N7-58N7-36A21-40P21
6619358
6695858
31O22-60L21-66A5-50G17-6F3
6890602
6891624
66N22-55P1-45G23-59D2-16K16-70N14-40A13-35I18-72O22-28P4-7H1
7606900
7785497
7786880
7811368
8026855
41I23
8081857
8132682
37M10-47M3
8288970

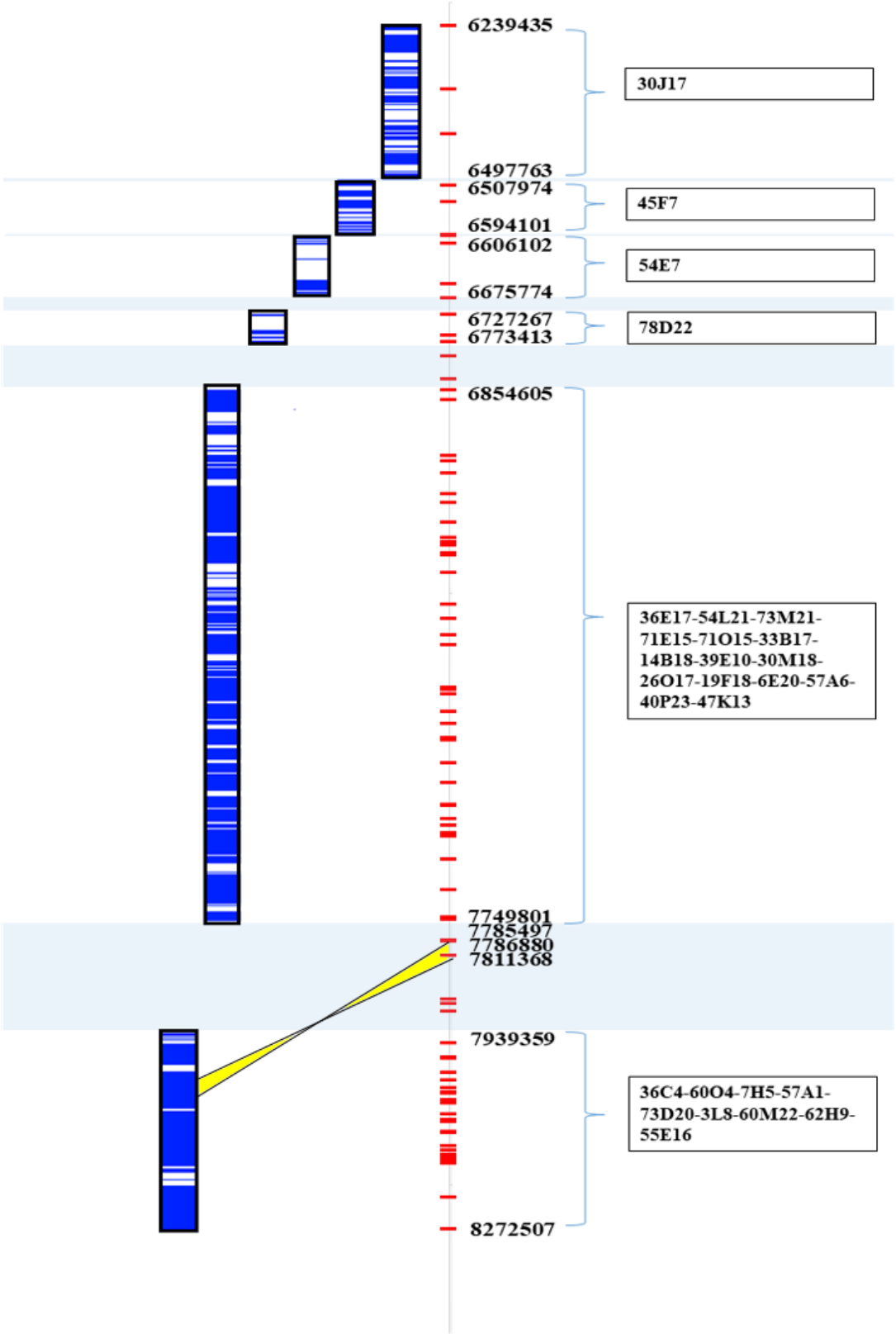

6239435
30J17
6497763
6507974
45F7
6594101
6606102
54E7
6675774
6727267
78D22
6773413
6854605
36E17-54L21-73M21-
71E15-71O15-33B17-
14B18-39E10-30M18-
26O17-19F18-6E20-57A6-
40P23-47K13
7749801
7785497
7786880
7811368
7939359
36C4-60O4-7H5-57A1-
73D20-3L8-60M22-62H9-
55E16
8272507

Fig. 15 – BAC supercontigs of the resistant 'Lito' haplotype aligned to the peach sequence V.1. from 6.2 to 8.3 Mbp. Markers of the 'Lito' genetic map between s1_6345556 and PGS1-24 are reported in red, according to their physical distance in peach genome. Supercontigs are represented by the green blocks on the left and their relative start/end positions on the peach genome are reported in the central part. Green color shows the shared regions between apricot and peach genome. The unreconstructed regions are highlighted in light-blue color. The possible inversion found in apricot is shown in yellow.

Fig. 16 – 'Lito' BAC supercontigs of the susceptible haplotype aligned to the peach sequence V.1 from 6.2 to 8.3 Mbp. Markers of the 'Lito' genetic map between s1_6345556 and PGS1-24 are reported in red, according to their physical distance in peach genome. Supercontigs are represented by the blue blocks on the left and their relative start/end positions on the peach genome are reported in the central part. Blue color shows the shared regions between apricot and peach genome. The unreconstructed regions are highlighted in light-blue color. The possible inversion found in apricot is shown in yellow.

DISCUSSION

PPV resistance in apricot is a quantitative trait (Soriano *et al.,* 2008; Lambert *et al*., 2007; Lalli *et al*., 2008; Dondini *et al*., 2011; Marandel *et al*., 2009°; Pilarova *et al*., 2010) particularly difficult to analyze because the phenotypic assignement to a dicotomic choice (resistant/susceptible) is not always easy and virologists prefer to use discrete classes (Babini and Fontana, 2012; Kegler *et al*., 1998; Faggioli and Barba, 1997; Karayannis, 2008).

PPV resistance phenotyping is a lengthy procedure, in which standardization is hindered by environmental factors and the physiological state of both plant and rootstock, that affect the manifestation of the trait. Beside environmental factors, translocation of the virus and development of the infection may be affected by minor as yet unknown factors (Soriano *et al.*, 2011). Differences between resistant cultivars in the restriction of virus movement upon inoculation are well documented (Ion-Nagy *et al*., 2006). This is the reason why, in spite of the large body of literature available, the genetic basis of Sharka resistance is still under debate.

Genetically, it is known from literature that the major determinant of resistance to Sharka is localized in the upper part of the linkage group 1 within the resistant apricot cultivars (Vera Ruiz *et al*., 2011; Marandel *et al.*, 2009a; Pilarova *et al*., 2010; Soriano *et al*., 2011). Since the function of the gene/s involved is unknown, the positional cloning has been expoited for the identification of resistance gene/s. In this clone-by-clone strategy, sequencing has been performed in libraries derived from individual genomic large-insert clones, selected in a minimum tile path according to physical and genetic map information. This approach benefited from work in small units, effectively reducing complexity and computational requirements.

The final goal of this thesis was the characterization of the resistant region to Sharka of the apricot cultivar 'Lito', isolating a chromosome region sufficiently restricted to be sequenced and assembled.

The approach started from a QTL mapped in the linkage group 1(LG1) of ‘Lito’, an apricot cultivar resistant to PPV.

However, the low map resolution did not supported the map-based cloning of the locus. For this reason, we enlarged the population of the controlled cross 'Lito' (resistant) X 'BO81604311' (susceptible) and saturated the region of the map with a further set of molecular markers isolated both from peach genome and the available scaffolds of 'Lito' sequenced at low coverage.

The screening of a 'Lito' BAC library with the markers of the map region allowed sorting out BAC clones suitable to produce a'Minimum Tiling Path' of the region, that is the minimum set of BAC clones that cover the region under study. In particular, 26 BAC clones covered the region of the resistance haplotype and 30 BAC clones covered the susceptible one around the QTL.

The saturation of the region of 'Lito' linkage group 1 with further molecular markers obtained by the BAC-end sequencing allowed to narrow down the region of the QTL to 1,7 cM between the markers GOL61 and pchcms4, which corresponds more or less to 2 Mbp of the genomic sequence of 'Lito'.

Through the process of *de novo* assembly, a genome is pieced together computationally, from overlapping randomly sequenced reads (Hunt, *et al*., 2014).

In our case, the *de novo* assembly process was performed in two steps. Firstly, each BAC clone was assembled individually from the Illumina reads; secondly, both the resistant and susceptible haplotypic region was reconstructed starting from contigs obtained in the first step. This multi-step process has required several cycles to reach a reliable assembly with the fewest conflicts.

Single BAC clones, assembled with CLC Genomics Workbench v3 using a de novo paired – end assembly algorithm, in many cases turned out to be fragmented in several contigs. This happened because the region under study was complex due to the presence of repeats.

The overlaps between reads were not long enough to distinguish between repeats present elsewhere. These repeat sequences were left unassembled and this solution split the contigs and made the analysis trickier. As shown in table n. 9, only seven BACs could be assembled in a single contig. The majority of BACs where fragmented in two to fifteen contigs and more than a few were extremely fragmented (20 - 67 contigs).

The peach genome taken as a reference was not completely suitable for the reconstruction of the ‘Lito’ region because of the relaxed sintheny, but it sped up the development of new molecular markers close to the QTL, enabling the use of the BAC library fo the production of a BAC minimum tiling path suitable for the positional cloning work.

The comparison of the position of anchor markers in several maps constructed with *Prunus* populations showed that the genomes of diploid (2*n*=16) species like almond, apricot, sweet cherry, peach, *Prunus cerasifera*, *P. davidiana* and *P. ferganinsis* revealed a rough co-linearity (Dirlewanger *et al.*, 2004a). Such collinearity ease the construction of framework maps and the saturation of chromosomal regions of interest virtually in any cross involving *Prunus* species by making use of markers of known position.

Despite the genetic collinearity, the variation at sequence level and the fact that the region under study was reach in repeats hampered the nucleotide sequence assembly of the region of interest in 'Lito'. In other words, peach genome served as a good guideline to anchor the BAC clones but it was not possible to use the peach genome to guide the assembly of 'Lito'. Indeed, in many cases the BAC contig alignment against peach genome provided multiple matches.

The order and orientation of all the BAC clones was not determined and it has been necessary the use a Dot Plot approach.

The manual assembly of BAC clones permitted to reconstruct 1,4 Mbp of ‘Lito’ sequence for the resistant haplotype and 1,6 Mbp for the susceptible one.

However, after the whole assembly process, supercontigs were still fragmented in contigs and some points of the sequence lacked of coverage. Yet, the highly presence of repeats within the sequences hampered the drawing of new unique primers on the supercontigs ends which could have permitted the identification of new BAC clones from the BAC library to close the gaps.

To solve these problems and to complete the sequence reconstruction of the region for the two haplotypes, we decided to modify the strategy of the program, by moving to the third-generation sequencing technology that has started to address some of the inherent limitations of sequencing and assembling complex regions in plant genomes.

CHAPTER 2 - PACBIO 'LITO' WHOLE GENOME SEQUENCING AND ASSEMBLY

INTRODUCTION

Within a research project aiming to map and clone candidate genes of resistance to Plum Pox Virus (PPV or Sharka), a QTL has been mapped in the linkage group 1 (LG1) of 'Lito', an apricot genotype resistant to Sharka. A fine genetic map has been then produced for the region of interest, that encompasses some 2 Mbp in the peach genome sequence taken as reference, by making use of the enlarged map population 'Lito' (resistant) x 'BO81604311' (susceptible) and new markers. Yet, a physical map of the region has been constructed with a mimimum tiling path of BAC clones selected from a 'Lito' BAC library. Sequencing those BAC clones with an Illumina NGS platform allowed the reconstruction of both resistant and susceptible haplotypes of 'Lito' with several gaps and fragmented supercontigs due to the presence of repetitive sequences, polymorphism, missing data and mistakes that limited the assembly. All these steps have been described in chapter 1.

In this chapter we discuss the assembly of the region of interest using either sequences from BAC clones and new long sequences produced with the PacBio technology.

Second-generation sequencing technologies are based on short reads and this makes them poorly suited for de novo assembly and annotation of complex regions.

The use of long reads is expected to address some of those shortcomings and to improve the overall quality of de novo assembly by ordering contigs, closing gaps, and improving scaffolding.

Single-molecule sequencing, developed by Pacific BioSciences (PacBio), offers the opportunity to overcome of these limitations producing reads with length from 10 kb up to some 30 kb (Rhoads and Fai Au, 2015).

The template, called a SMRTbell, is a closed, single-stranded circular DNA that is created by ligating hairpin adaptors to both end of a target double-stranded DNA molecule. When a sample is loaded to a chip called SMRT cell, a SMRTbell diffuses into a sequencing unit called a zero-mode waveguide (ZMW), which provides the smallest available

volume for light detection. In each ZMW, a single polymerase is immobilized at the bottom, which can bind to either hairpin adaptor of the SMRTbell and starts the replication. Four fluorescent-labeled nucleotides, which generate distinct emission spectrums, are added to the SMRT cell for the polymerase activity.

The replication processes are recorded by a “movie” of light pulses, and the pulses corresponding to each ZMW can be interpreted as a base of the growing sequence.

Because the SMRTbell forms a closed circle, after the polymerase replicates one strand of the target dsDNA, it can continue incorporating bases of the adapter and then the other strand. If the lifetime of the polymerase is long enough, both strands can be sequenced multiple times. In this scenario, the sequences can be split to multiple reads, called subreads, as the adaptor sequences are recognized and cutted.

MATERIALS and METHODS

Plant Material

The apricot cultivar 'Lito' was selected because it carries the resistance to Sharka and because the same cultivar was used to obtain the BAC library.
'Lito' is a Greek cultivar derived from the cross 'Stark Early Orange' (SEO, the donor of resistance) x the susceptible cultivar 'Early of Trynthos'.

PacBio whole genome sequencing of 'Lito'

PacBio 'Lito' whole genome sequencing has been commissioned to Amplicon Express. Some 20 g of plant young leaf tissue were sent on dry ice to Amplicon Express for HMW DNA extraction using a CTAB isolation method modified by R. Meilan (unpublished, rmeilan@purdue.edu), based on the original method Doyle & Doyle. Key features of Amplicon's NGS grade gDNA prep services were as follows: robust cell lysis (rough handling in strong detergents) followed by steps with gentle handling to prevent gDNA shearing. The gDNA was column purified with a traditional anion exchange resin and re-suspended in an Amplicon Express proprietary solution maximizing DNA quantity and quality.
A total of 5µg per sample were used as input into three libraries preparation. The SMRTbell libraries were constructed with SMRTbell™ Template Prep Kit 1.0 following the manufacturer's instructions (Pacific Biosciences). The small fragments lower than 20 kb of SMRTbell template were removed using Blue Pippin Size selection system for large – insert library. The constructed libraries were validated by Agilent 2100 Bioanalyzer. After a sequencing primer is annealed to the SMRTbell template, DNA polymerase is bound to the complex using DNA/Polymerase Binding kit P6. This polymerase – SMRTbell – adaptor complex is the loaded into SMRT cells.
The first two SMRTbell libraries were sequenced using three SMRT cells each other using C4 chemistry (DNA sequencing Reagent 4.0) and 2401-minute movies were captured for each SMRT cell using the PacBio RS II sequencing platform.

The last library was commissioned at a later stage and sequenced using the Sequel™ System, the newest Single Molecule, Real-Time sequencer. The Sequel System provides higher throughput, more scalability, a reduced footprint and lower sequencing project costs compared to the PacBio® RS II System, while maintaining the benefits of SMRT technology. The core of the Sequel System is the capacity of its redesigned SMRT Cells, which contain one million zero-mode waveguides (ZMWs) at launch, compared to 150,000 ZMWs in the PacBio RS II. Active individual polymerases are immobilized within the ZMWs, providing windows to observe and record DNA sequencing.

PacBio reads alignment to BAC supercontigs

Alignment of PacBio reads against the BAC assembly of the region for the two haplotypes (resistant and susceptible) and the peach genome was carried out with BLASR (https://github.com/PacificBiosciences/blasr). The presence of the peach genome in the reference helped the right alignment without forcing the alignment of the reads. This could prevent the alignment of the reads against wrong sites.
IGV3 (http://software.broadinstitute.org/software/igv/igv3.0) was used to visualize PacBio reads alignments against our reference sequences. IGV3 has two extra features compared to Tablet: "quick consensus mode" and "hide indels", to reveal biological variation in PacBio reads. The quick consensus mode shows mismatches only at positions where more than a specified fraction of reads disagrees with the reference. The "hide indels" feature suppresses the most common error in raw PacBio reads like random small indels.

'Lito' Whole genome de novo assembly

'Lito' whole genome *de novo* assembly was generated starting from PacBio reads using Canu software (Koren *et al.*, 2017).
Canu is a fork of the Celera Assembler designed for high-noise single-molecule sequencing such as the PacBio RSII. The Canu pipeline consists of three stages: correction, trimming and assembly, each of which can run independently or in series.

The correction stage selects the best overlaps to use for correction, estimates corrected read lengths, and generates corrected reads. The trimming stage identifies unsupported regions in the input and trims or splits reads to their longest supported range. The assembly stage makes a final pass to identify sequencing errors, constructs the best overlap graph and outputs contigs, an assembly graph and summary statistics.

A de novo assembly was generated setting correctedErrorRate=0.045 and corMax EvidenceErate=0.2.

CorrectedErrorRate is the maximum expected difference in the alignment of two corrected reads. For less than 30 X coverage, 0.075 value is recommended to adjust for inferior read correction. CorMaxEvidenceErate value limits read correction to only overlaps at or below this fraction error. A value of 0.2 is used for plants to speed up the assembly process. Statistics for the assembly were obtained using QUAST (http://quast.sourceforge.net/quast), a quality assessment tool for evaluating and comparing genome assembly.

Canu contigs extraction and assembly of the resistant and susceptible haplotypes

Canu contigs covering the region of interest were chosen from the *de novo* whole genome assembly of 'Lito' using the Basic Local Alignment Search Tool (Nucleotide-Nucleotide BLAST 2.2.27+) which finds regions of local similarity between sequences.

This program allowed comparing nucleotide sequences from BAC supercontigs assembly with the *de novo* contigs sequences obtained with Canu software. This work was done setting an e-value of 1e-200. This parameter gives a measure of the similarity of sequences: the lower is the e-value, the higher the congruity of the query sequence and the retrieved sequence.

Selected contig sequences of susceptible and resistant assembled supercontigs were aligned each other using dot-plot. This permitted to verify the concordance between the nucleotide sequences and understand which contig sequences belong to resistant or susceptible haplotype, to verify the contig order inside the supercontig sequences, to close gaps, and, where possible, to extent supercontig sequences close to the gaps and scaffolding a unique sequence.

The BACs Illumina reads were aligned against the final assembled sequences to verify the accuracy of the sequences, to correct possible errors and to obtain a better assembly. Alignment of Illumina reads was carried out with BWA. SNPs and small INDELs were called using default parameters of Unified Genotyper of GATK and manually checked and corrected.

RESULTS

PacBio Whole genome sequencing of 'Lito'

'Lito' genome sequencing with PacBio technology produced long sequences as expected although with a percentage of error rather high (10-15% error rate for a single read).
From six SMRTcell with libraries of 20 and 30 kb sequenced with PacBio RS II sequencing platform, after filtering, 312,605 and 286,895 reads were obtained respectively. Based on the estimated size of 240 Mb of the apricot genome, the theoretical genome coverage obtained with these two library was 18 X.
These data were enriched at a later time by commissioning five further SMRTcell with a library of 20 kb sequenced with Sequel System, which produced 2,412,177 reads. For this last library, the theoretical genome coverage was 80 X.
Sequencing performance for such a technology can be measured in read length and total throughput per experiment. Specific statistics for all tree libraries are reported in table 12. The figures 17, 18 and 19 show the read length distribution for the tree libraries.

Tab. 12 - Statistics of the tree libraries of 'Lito' sequenced with PacBio technology.

LIBRARY	TOTAL LENGTH	SUBREADS	AVERAGE	MAX	N50 SUBREAD LENGTH
FILTERED SUBREADS 30KB	2,027,049,302	341,983	5,927	67,025	13,006
FILTERED SUBREADS 20KB	3,183,302,676	347,671	9,156	50,985	14,884
FILTERED SUBREADS 30KB SEQUEL	20,657,074,788	2,142,177	9,643	91,274	15,559

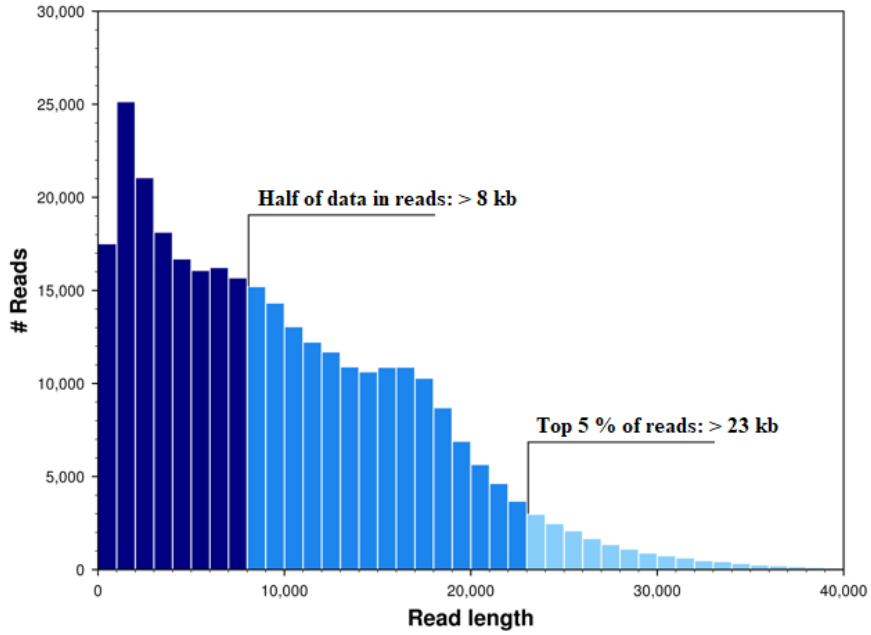


Fig. 17 – Read length distribution of 20 kb library sequenced with PacBio RS II sequencing platform.

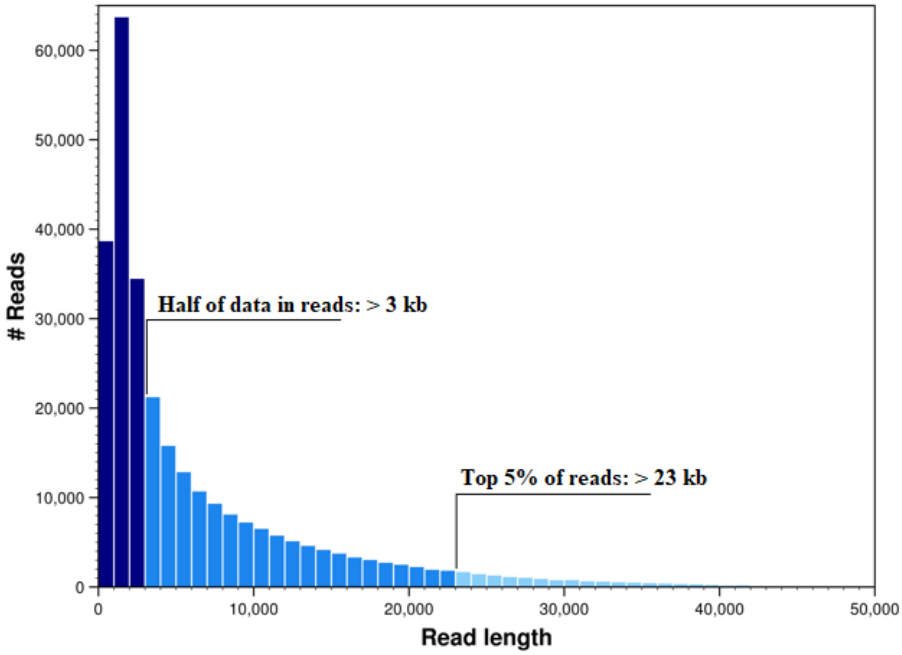


Fig. 18 - Read length distribution of 30 kb library sequenced with PacBio RS II sequencing platform.

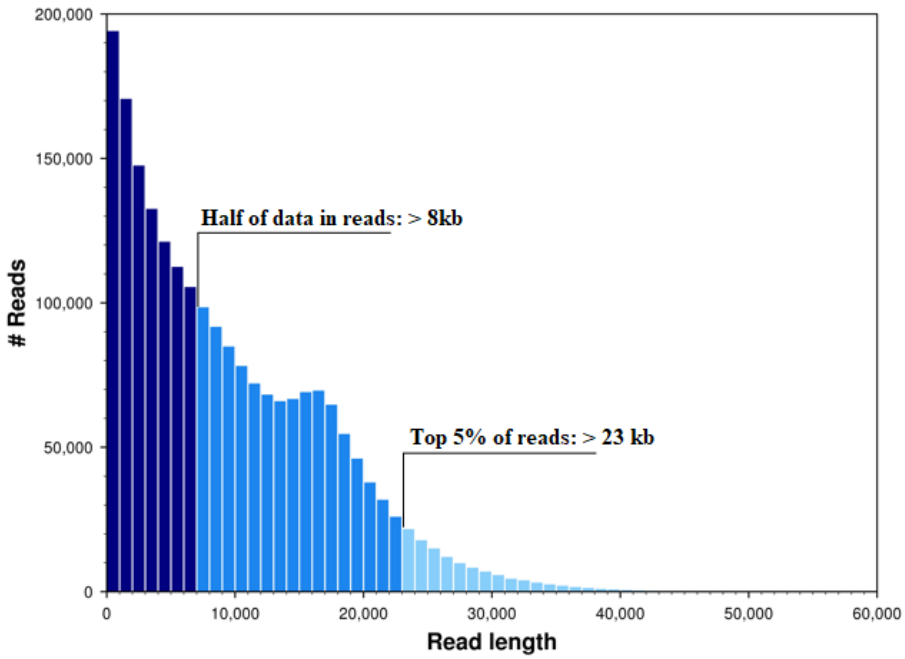


Fig. 19 - Read length distribution of 20 kb library sequenced with PacBio Sequel System sequencing platform.

'Lito' whole genome de novo assembly

Provided the quality of the statistics and the reads length distribution we decided to attempt a 'Lito' whole genome assembly using only the reads of the library sequenced trough Sequel System. Statistics for the assembly are reported in table 13.

A total of 2,142,177 PacBio reads were assembled in 3,762 contigs, with a N50 of 197,570 bp. Approx. 3,229 of these contigs (85%) had length greater than 25,000 bp and the largest contig was 2,358,793 bp in length. The assembled genome size was 360 Mb, about 50% larger than expected.

Canu splits haplotypes into separate contigs whenever the allelic divergence is greater than the post-correction overlap error rate. This threshold is typically 1,5% for recent PacBio data. This splitting is likely the cause of the assembly size larger than the expected haploid genome.

Tab. 13 – Statistics of 'Lito' whole genome assembly. The table shows: the number of assembled contigs and the length of the contigs in the assembly, the length of the largest and lower contig in the assembly, the average contig length, the total number of bases in the assembly and N50 and L50 statistics.

Apricot assembly statistics	
contigs (>= 0 bp)	3,762
contigs (>= 1000 bp)	3,762
contigs (>= 5000 bp)	3,743
contigs (>= 10000 bp)	3,723
contigs (>= 25000 bp)	3,229
contigs (>= 50000 bp)	1,686
Total length (>= 0 bp)	360,675,980
Total length (>= 1000 bp)	360,675,980
Total length (>= 5000 bp)	360,622,451
Total length (>= 10000 bp)	360,468,418
Total length (>= 25000 bp)	350,704,446
Total length (>= 50000 bp)	294,932,385
Largest contig	2,358,793
Lower contig	1,059
Average	95873,5
Total length (bp)	360,675,980
N50	359
L50	197,570

Canu contigs extraction and assembly of the resistant and susceptible haplotypes

Data from 20 and 30 kb libraries sequenced with PacBio RS II were the only ones available at first.

PacBio reads, being longer than Illumina reads, once aligned against the supercontigs of the two haplotypes, assembled using BACs, gave the possibility to understand the correct order of the contigs and closed the small gaps that remained in the supercontigs.

In particular, for each gap in the supercontigs, PacBio reads were extracted and aligned using dotplot to verify more precisely the presence of possible misassembles and bridge the gaps.

However, PacBio reads have a high error rate, and for that reason it was necessary to align the Illumina reads from the BACs and correct manually the sequence modified with PacBio reads each time (fig. 20).

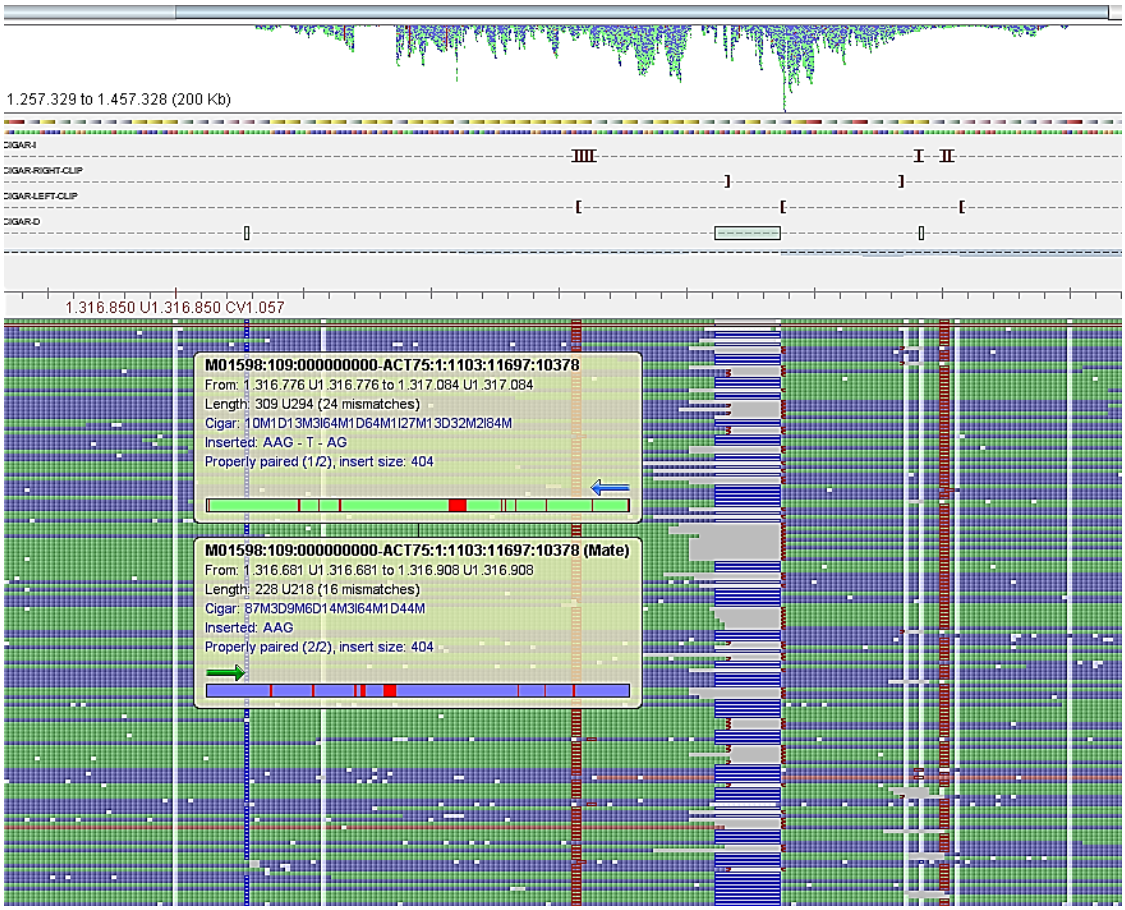


Fig. 20 – Example of alignment of Illumina paired-end reads of a BAC clone against the 'Lito' sequence post-modification with PacBio reads displayed using Tablet. SNPs are highlighted in white color, indels in blu, and insertions in red.

Despite the length of the reads, the amount of data from the first two libraries was too low to bridge all the gaps in the assembly of the region. In particular, the big gaps between the supercontigs. To solve this problem, the data were enriched with 20 further Gb of new PacBio sequences.

Using the new set of PacBio sequences, a whole genome assembly of 'Lito' was attempted. The whole idea was that, being the contigs longer than the individual PacBio reads and having fewer errors in the sequence, they could have been used as guideline to complete the reconstruction of the region (fig. 21).

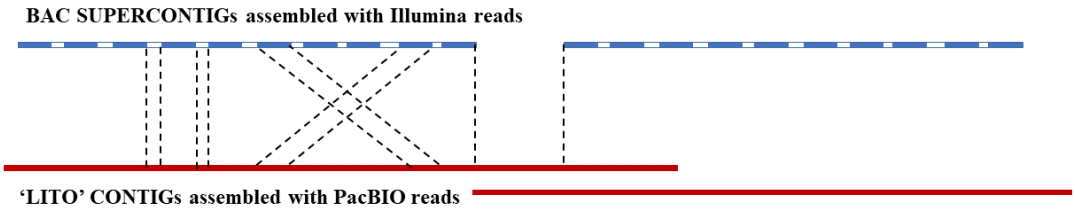


Fig. 21 – Representation of the strategy to complete the reconstruction of the region under study. Assembled PacBio contigs, being long, could be used to verify and understand the correct order of the contigs that made up the BAC supercontigs and closed the gaps inside and between supercontigs.

By aligning the contigs of the whole genome assembly of 'Lito' against the BAC supercontigs assembled using Illumina technology, it was possible to identify contigs covering the region under study.

In particular, 20 contigs were identified. One at a time, these sequences were aligned against the BAC supercontigs using dot plot. This work permitted to verify more efficiently the goodness of the assembly and to assign the contigs to the resistant/susceptible haplotype.

Overlapping between PacBio contigs and the BAC supercontigs allowed to bridge the gaps still present into and between these sequences.

The final contigs produced by Canu were almost perfect. Anyway, since the error rate of PacBio reads is high, few errors in the contigs remain despite the accurate correction performed by Canu. Therefore, at each modification of the assembled sequences, Illumina reads of BAC clones were aligned to verify the presence of small indels or SNPs.

In the regions where no BAC clones reads were available, few errors could be still present.

As mentioned above, Canu split haplotypes into separate contigs wherever the allelic divergence is greater than the post-correction overlap error rate. As a result redundant contigs covering the same region were obtained. The haplotype with more reads is often reconstructed in a large contig spanning the locus, while the haplotype with fewer reads is just the variant region. Less diverged regions are collapsed. In particular, Canu split the two haplotypes for almost the entire region. Only the final part of the region was reconstructed as a unique contig, due to the relatively low level of differences between the two haplotypes (fig.22).

The sequence of this main contig bridged the gap (fig.15 and 16 – chapter 1) from 7.6 Mbp to 8.0 Mbp (coordinates relative to peach genome V.1) in the resistant haplotypes and from 7.7 Mbp to 7.9 Mbp in the susceptible one. The gap, for both the haplotypes was closed using the sequence of the same contig. By aligning the PacBio reads against this region, it was possible to understand the phase of some small variants and to solve the two haplotypes.

Through this work, we are confident to have reconstructed with a substantial precision the resistant and susceptible region on the LG1 of 'Lito'. The assembly of both haplotypic regions consist in two continue supercontigs without gaps within the sequence.

The 132 contigs, which made up the four supercontigs of the resistance region assembled using BACs were assembled in two sequences of 264,556 and 1,419,143 bp in length, that totals overall 1,683,699 bp; while the 150 contigs that made up the five supercontigs of the susceptibility region were assembled in two sequences of 321,950 and 1,413,486 bp in length, totaling 1,735,436 bp.

A gap still remains in both haplotypes because of the high rate of repetitive sequences in that gap. Peach genome can not help to estimate precisely the size of the gap because the sequence appears highly different in the two species.

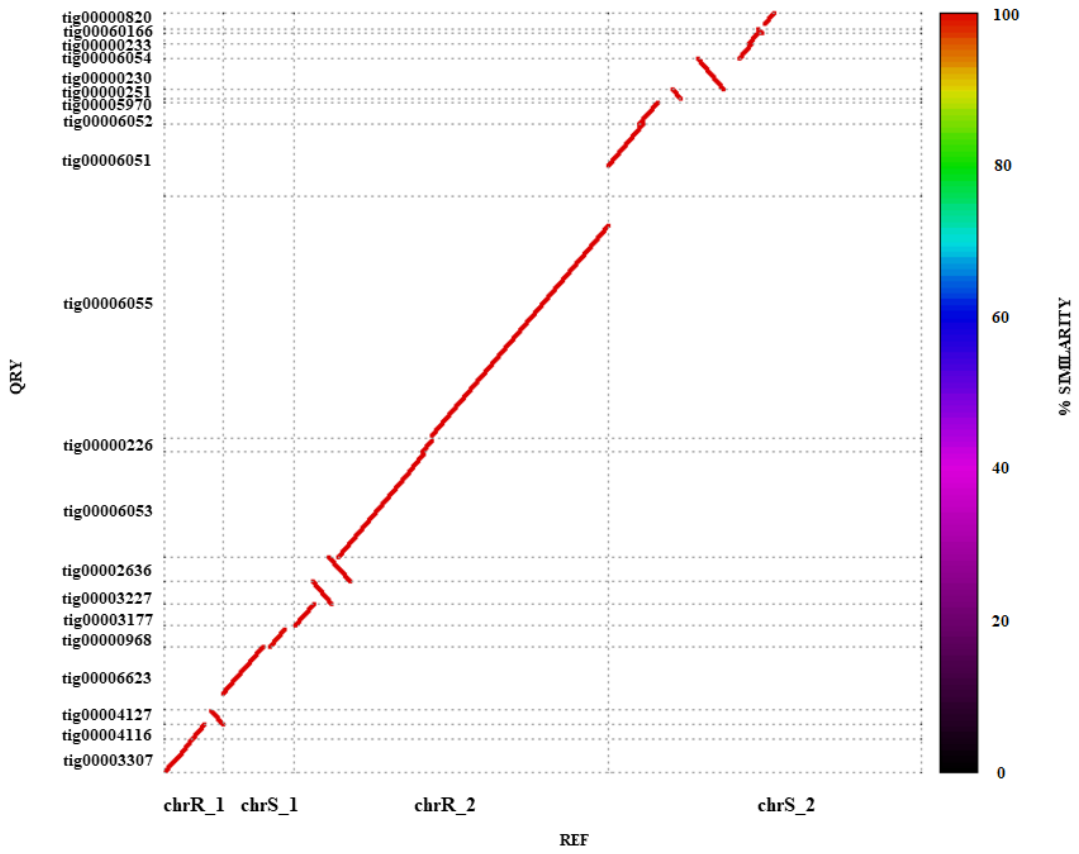


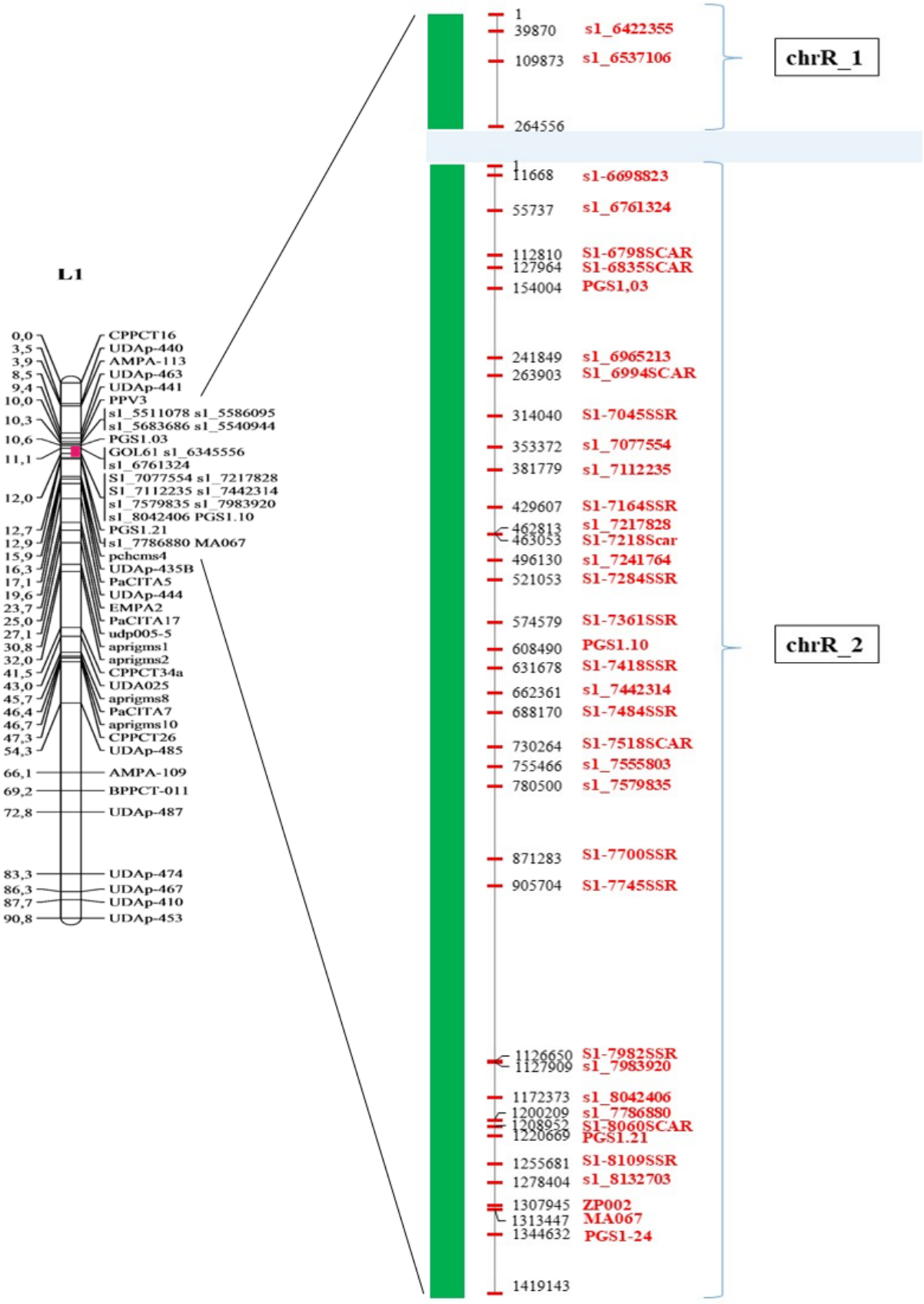
Fig. 22 – NUCMer plot (NUCleotide MUMer - http://mummer.sourceforge.net/) obtained aligning contigs (query sequences, QRY) from the Whole genome assembly of 'Lito' against the reference (REF) assembled sequences for the two haplotypes of the region (chrR_1, chrS_1, chrR_2, chrS_2). Contigs have similarity closest to 100% with the assembled sequences. Canu assembler was able to reconstruct different contigs covering the sequences chrR_1, chrS_1, and the first part of the sequences chrR_2 and chrS_2. The last part of those sequences is covered only by the contig tig00006055 due to the relatively low level of differences between the two haplotypes.

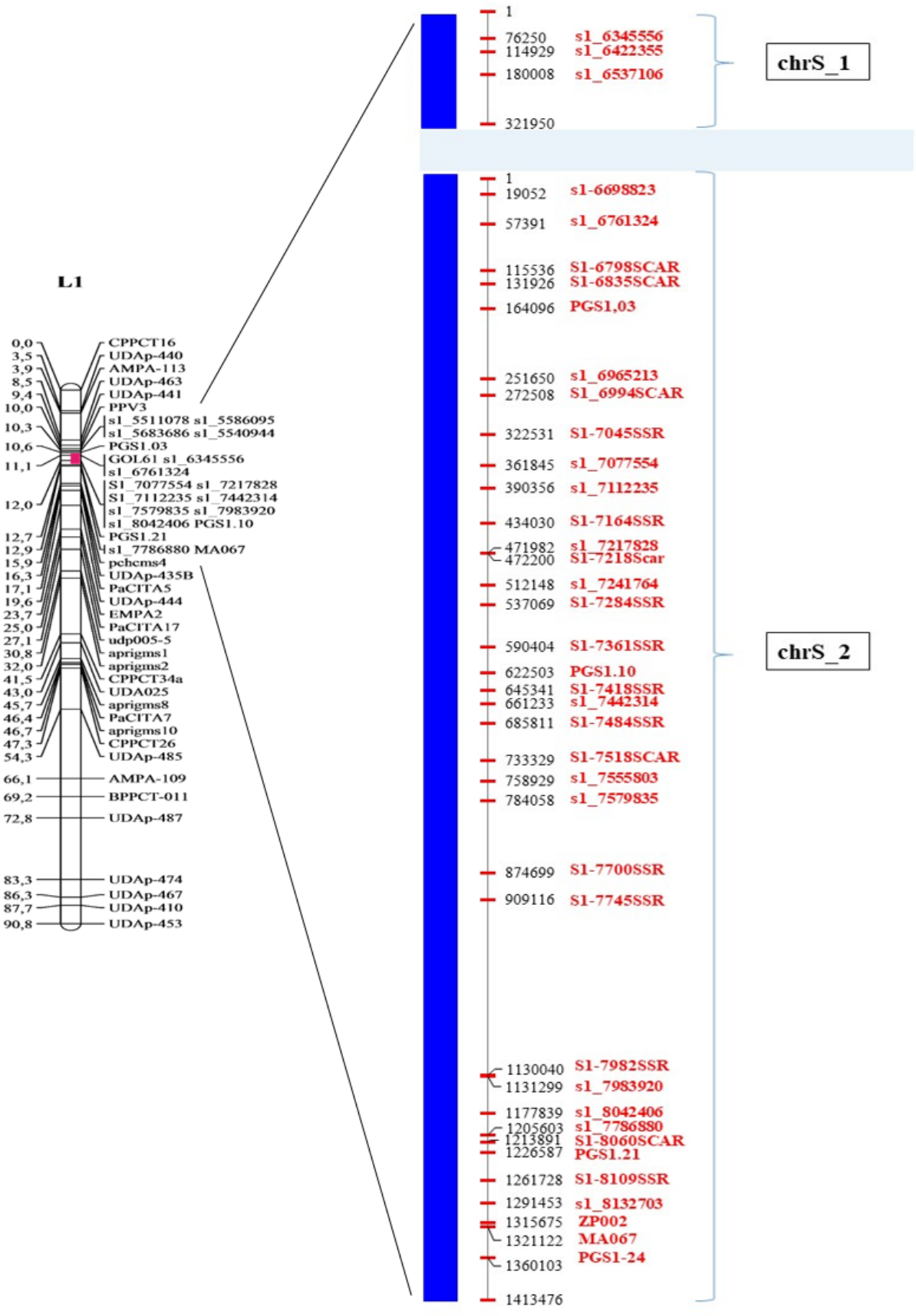
Physical map of resistance/susceptibility locus in 'Lito'

The figures 23 and 24 below are a graphical representation of the 'Lito' regions in which the assembled sequences for both resistant and susceptible haplotypes (in green the resistant haplotype and blue the susceptible one) were directly anchored to the apricot genetic map of the linkage group 1. These two physical maps were obtained by searching, within the ultimate assembled sequences, all molecular markers tightly linked to the resistance/susceptible locus that have been used in this work.
The information derived from this work will be important to better understand the complex recombination occurring in the region and to indentify the markers more closely linked with the putative candidate gene/s for Sharka resistance.

Fig. 23 – The assembled region of interest of the resistant haplotype in 'Lito' LG1. Molecular markers are located following their physical distance in the sequence. On the left, the genetic map of the LG1 of 'Lito' is reported and the locus under study is in red. Genetic distances in centimorgan (cM) are shown on the left of the map. On the right, physical distances of markers are reported within the two pieces of sequence (chrR_1, chrR_2) that make up the resistant haplotype. Molecular markers correspond to the relative position of the forward primer in the assembled sequence. Physical distances are measured in number of base pairs. A region lacking of coverage still remains in the final assembly. This region is highlighted in light-blue background color.

Fig. 24 – The assembled region of interest of the susceptible haplotype in LG1 'Lito'. Molecular markers are located following their physical distance in the sequence. On the left, the genetic map of the LG1 of 'Lito' is reported and the locus under study is in red. Genetic distances in centimorgan (cM) are shown on the left of the map. On the right, physical distances of markers are reported within the two pieces of sequence (chrS_1. chrS_2) that make up the susceptible haplotype. Molecular markers correspond to the relative position of the forward primer in the assembled sequences. Physical distances are measured in number of base pairs. A region lacking of coverage still remains in the final assembly. This region is highlighted in light-blue color.


L1
0,0 CPPCT16
3,5 UDAp-440
3,9 AMPA-113
8,5 UDAp-463
9,4 UDAp-441
10,0 PPV3
10,3 s1_5511078 s1_5586095
s1_5683686 s1_5540944
10,6 PGS1.03
11,1 GOL61 s1_6345556
s1_6761324
12,0 S1_7077554 s1_7217828
S1_7112235 s1_7442314
s1_7579835 s1_7983920
s1_8042406 PGS1.10
12,7 PGS1.21
12,9 s1_7786880 MA067
15,9 pchcms4
16,3 UDAp-435B
17,1 PaCITA5
19,6 UDAp-444
23,7 EMPA2
25,0 PaCITA17
27,1 udp005-5
30,8 aprigms1
32,0 aprigms2
41,5 CPPCT34a
43,0 UDA025
45,7 aprigms8
46,4 PaCITA7
46,7 aprigms10
47,3 CPPCT26
54,3 UDAp-485
66,1 AMPA-109
69,2 BPPCT-011
72,8 UDAp-487
83,3 UDAp-474
86,3 UDAp-467
87,7 UDAp-410
90,8 UDAp-453
1
39870 s1_6422355
109873 s1_6537106
264556
chrR_1
1
11668 s1-6698823
55737 s1_6761324
112810 S1-6798SCAR
127964 S1-6835SCAR
154004 PGS1,03
241849 s1_6965213
263903 S1_6994SCAR
314040 S1-7045SSR
353372 s1_7077554
381779 s1_7112235
429607 S1-7164SSR
462813 s1_7217828
463053 S1-7218Scar
496130 s1_7241764
521053 S1-7284SSR
574579 S1-7361SSR
608490 PGS1.10
631678 S1-7418SSR
662361 s1_7442314
688170 S1-7484SSR
730264 S1-7518SCAR
755466 s1_7555803
780500 s1_7579835
871283 S1-7700SSR
905704 S1-7745SSR
1126650 S1-7982SSR
1127909 s1_7983920
1172373 s1_8042406
1200209 s1_7786880
1208952 S1-8060SCAR
1220669 PGS1.21
1255681 S1-8109SSR
1278404 s1_8132703
1307945 ZP002
1313447 MA067
1344632 PGS1-24
1419143
chrR_2


L1
0,0 CPPCT16
3,5 UDAp-440
3,9 AMPA-113
8,5 UDAp-463
9,4 UDAp-441
10,0 PPV3
10,3 s1_5511078 s1_5586095
s1_5683686 s1_5540944
10,6 PGS1.03
11,1 GOL61 s1_6345556
s1_6761324
12,0 S1_7077554 s1_7217828
S1_7112235 s1_7442314
s1_7579835 s1_7983920
s1_8042406 PGS1.10
12,7 PGS1.21
12,9 s1_7786880 MA067
15,9 pchcms4
16,3 UDAp-435B
17,1 PaCITA5
19,6 UDAp-444
23,7 EMPA2
25,0 PaCITA17
27,1 udp005-5
30,8 aprigms1
32,0 aprigms2
41,5 CPPCT34a
43,0 UDA025
45,7 aprigms8
46,4 PaCITA7
46,7 aprigms10
47,3 CPPCT26
54,3 UDAp-485
66,1 AMPA-109
69,2 BPPCT-011
72,8 UDAp-487
83,3 UDAp-474
86,3 UDAp-467
87,7 UDAp-410
90,8 UDAp-453
1
76250 s1_6345556
114929 s1_6422355
180008 s1_6537106
321950
chrS_1
1
19052 s1-6698823
57391 s1_6761324
115536 S1-6798SCAR
131926 S1-6835SCAR
164096 PGS1,03
251650 s1_6965213
272508 S1_6994SCAR
322531 S1-7045SSR
361845 s1_7077554
390356 s1_7112235
434030 S1-7164SSR
471982 s1_7217828
472200 S1-7218Scar
512148 s1_7241764
537069 S1-7284SSR
590404 S1-7361SSR
622503 PGS1.10
645341 S1-7418SSR
661233 s1_7442314
685811 S1-7484SSR
733329 S1-7518SCAR
758929 s1_7555803
784058 s1_7579835
874699 S1-7700SSR
909116 S1-7745SSR
1130040 S1-7982SSR
1131299 s1_7983920
1177839 s1_8042406
1205603 s1_7786880
1213891 S1-8060SCAR
1226587 PGS1.21
1261728 S1-8109SSR
1291453 s1_8132703
1315675 ZP002
1321122 MA067
1360103 PGS1-24
1413476
chrS_2

Comparison between the resistant and susceptible haplotypes

Assembled sequences of the resistance haplotype are composed of two nucleotide sequences of 264,556 and 1,419,143 bp and called chrR_1 and chrR_2 respectively , while those of the susceptible haplotype are composed of two nucleotide sequences of 321,950 and 1,413,486 called chrS_1 and chrS_2 bp respectively. Both pair of sequences are interrupted by a gap of unknown length in apricot.

The assembled sequences of the resistance haplotype were aligned against those of the susceptible one to help visually comparing the two haplotypes.

Comparison shows that we were able to reconstruct 69,942 bp more at the beginning of chrS_1 with respect to chrR_1, while 12,469 more bases were assembled on the tail of chrR_1 compared with chrS_2

Comparison between chrR_2 and chrS_2 shows that chrS_2 sequence has been reconstructed 7,328 bp longer at the beginning with respect to the resistant haplotype, while 12,992 more bases were assembled on the tail of chrR_2 compared with chrS_2.

The NUCMer (NUCleotide MUMer - http://mummer.sourceforge.net/) tool was used to align the sequences of resistant and susceptible haplotypes and to identify differences and similarities in the shared regions. The percentage of similarity is shown in figs. 25 and 26 by the color gradient on the right of the plot. Overall, the two haplotypes are similar to each other. Indeed, shared similarity is almost close to 100%.

Anyway, there are some insertions within the sequences of the resistant haplotype that are missing in the susceptible one and vice versa.

These differences are shown more precisely by the graphs of figs. 27 and 28, obtained using Gevo comparative sequence alignment tool (genomevolution.org/coge/GEvo.pl). Gevo permitted also to highlight (fig. 29) more clearly that the piece of sequence from 310,401 to 321,959 bp of the chrS_1 is duplicated in the chrR_1 sequence (221,476 – 232,941, 241,881 – 252,255 bp) but with a percentage of identity of 95%. The NUCMer plot confirms this as the color of the shared regions in this point turns to the orange. Probably, in this point, the sequence for both the haplotypes is complexed by several repetitions and for the susceptible sequence we were able to reconstruct only one repetition.

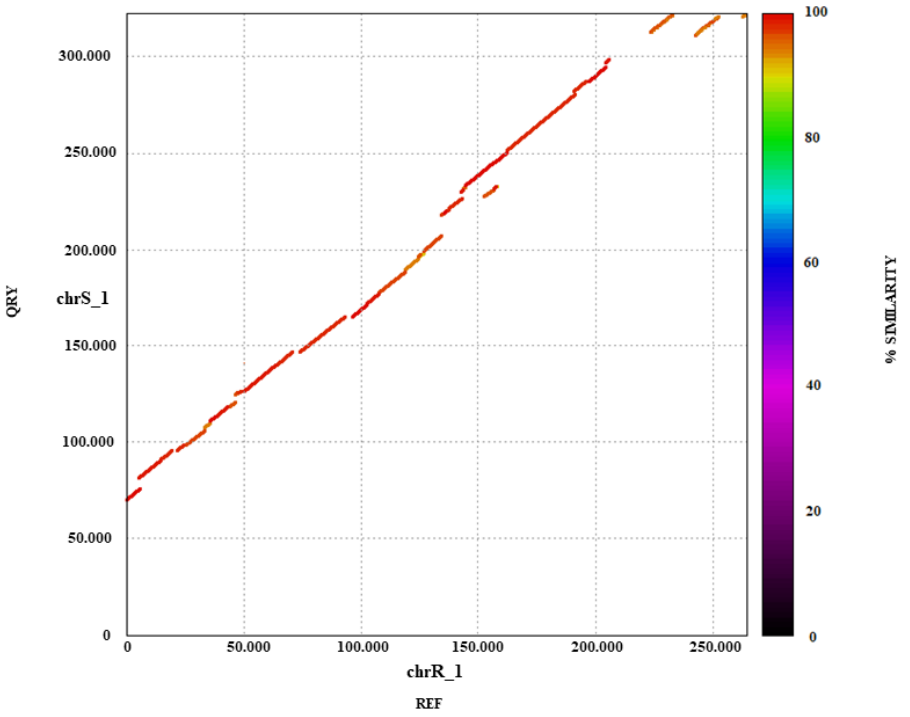


Fig. 25 – Alignment of chrR_1 against chrS_1. Plot was created using NUCMer.

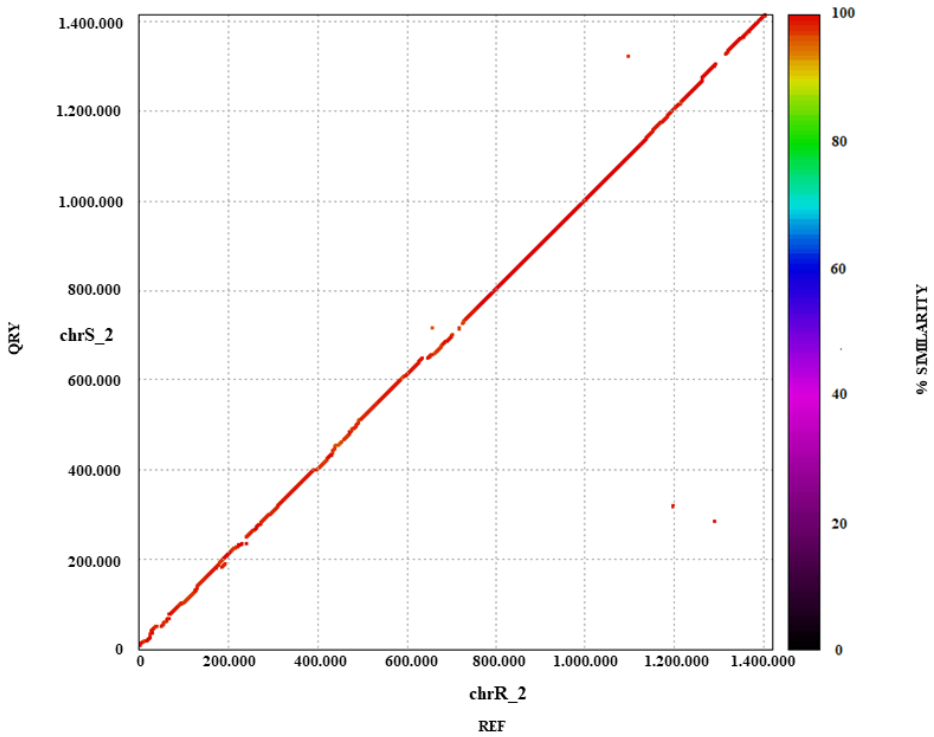


Fig. 26 - Alignment of chrR_2 against chrS_2. Plot was created using NUCMer.

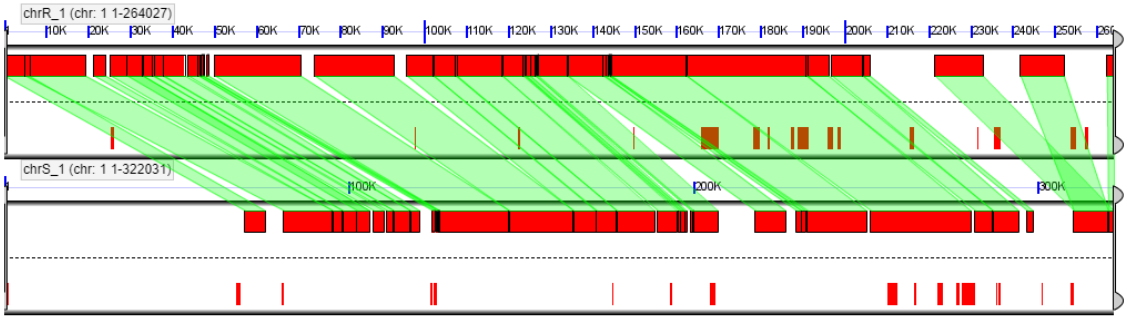


Fig. 27 – Alignment of chrR_1 against chrS_1. Graph was created using Gevo. Green connectors show the shared regions between the two haplotypes. White spaces highlight insertions within the sequences of the resistant haplotype compared to the susceptible ones and vice versa.

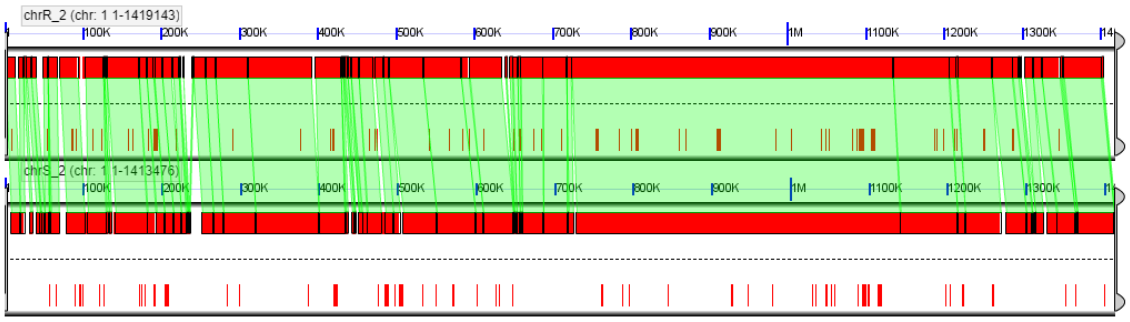


Fig. 28 - Alignment of chrR_2 against chrS_2. Graph was created using Gevo. Green connectors show the shared regions between the two haplotypes. White spaces highlight insertions within the sequences of the resistant haplotype compared to the susceptible ones and vice versa.

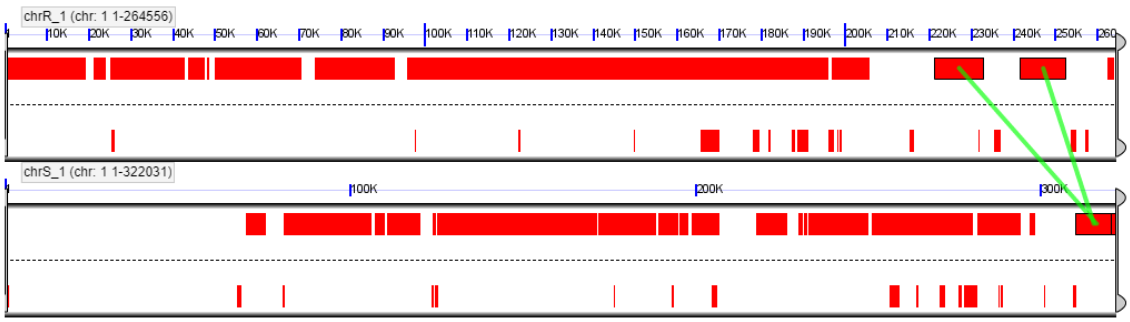


Fig. 29 - Alignment of chrR_1 against chrS_1. Graph was created using Gevo. Green connectors show the duplication in the chrR_1.

Comparison between resistance/susceptibility sequences of 'Lito' and peach genome

The assembled sequences for both R and S 'Lito' haplotypes were aligned onto the new release of peach genome (www.rosaceae.org/species/prunus_persica/genome_v2.0.a1). The reconstructed locus in 'Lito' spans in the peach chromosome 1 new release between 6,600,000 and 8,850,000 bp.

Comparison between resistance/susceptibility sequences of 'Lito' and Peach genomes, using NUCMer (fig. 30, 31), shows collinearity at the genome level. However, the similarity at sequence level, shown on the right of the plot, is between 80% and 95%. Moreover, the region is reach in repetitive sequences.

Gevo comparative sequence alignment tool permitted to highlight an inversion in apricot, in position 1,183,886 – 1,197,188 bp of the chrR_2 sequence (fig. 32) and 1,511,375 – 1,524,489 of the chrS_2 sequence (fig. 33), compared to the peach genome in the position 8,248,039 – 8,260,730 bp.

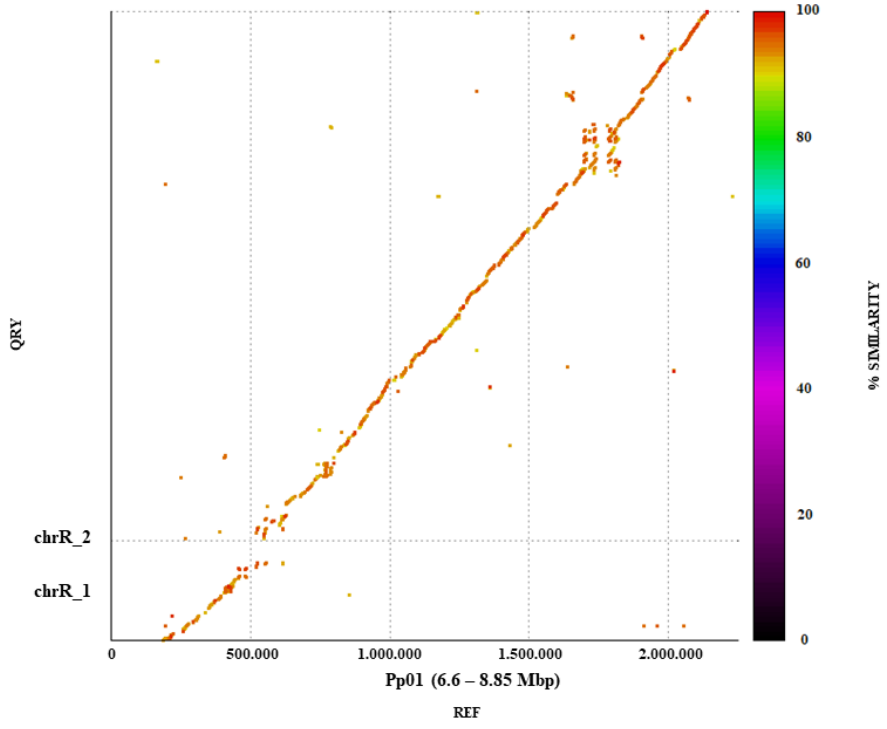


Fig. 30 - Alignment between the REF peach genome (6.6 – 8.85 Mbp) and 'Lito'(QRY) resistant haplotype. Plot was created using NUCMer. The percentage of similarity is shown by the color bar.

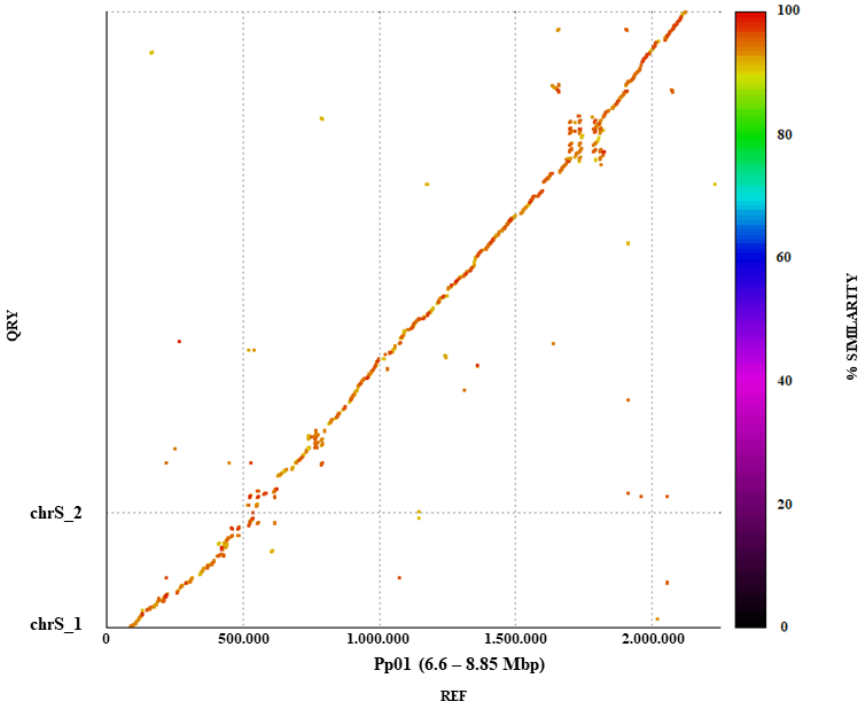


Fig. 31 - Alignment between the REF peach genome (6.6 – 8.85 Mbp) and 'Lito' (QRY) susceptible haplotype. Plot was created using NUCMer. The percentage of similarity is shown by the color bar.

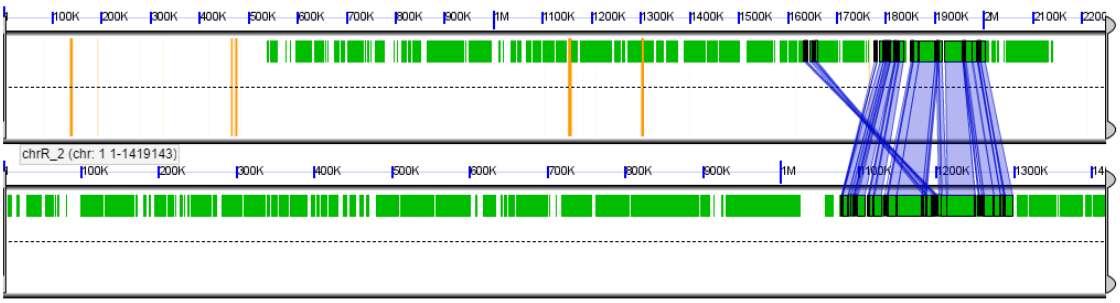


Fig. 32 - Alignment of chrR_2 against peach genome (6.6 – 8.85 Mbp). Graph was created using Gevo. Blue connectors show the shared regions between apricot and peach genome highlighting the inversion found in apricot. Orange bars indicate sequence gaps in the peach genome.

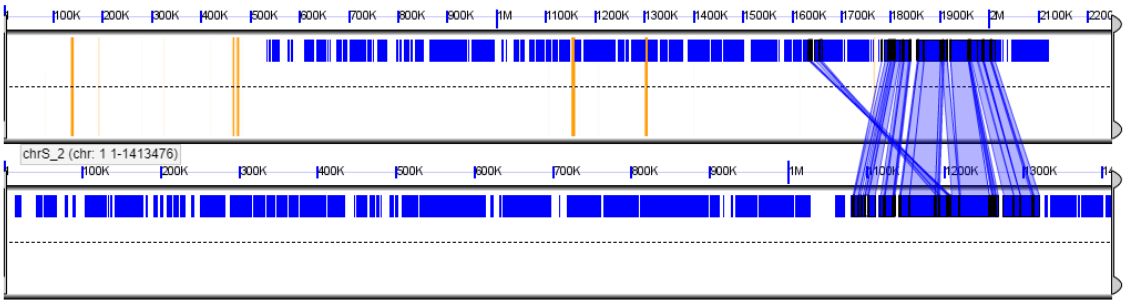


Fig. 33 - Alignment of chrS_2 against peach genome (6.6 – 8.85 Mbp). Graph was created using Gevo. Blue connectors show the shared regions between apricot and peach genome highlighting the inversion found in apricot Orange bars indicate sequence gaps in the peach genome.

DISCUSSION

The complexity of plant genomes remains a difficult challenge for *de novo* assembly for a variety of biological and computational reasons (Schatz *et al*. 2012).

One of the challenge in the assembly deals with the presence of large gene families and abundance of pseudogenes with nearly identical sequences derived from recent whole genome duplication events and transposon activity, which has been demonstrated in most plant assembled genomes (cit).

Another challenge depends on the fact that the plant genomes usually appear as gene islands among a background of high-copy repeats (usually >80%). The length of single-copy regions (always flanked by repeated sequences) varies widely among plant species (cit.).

Repeat sequences are difficult to assemble because reads with high identity levels could come from different regions of the genome. This generates gaps, ambiguities and collapses in alignment and assembly, which, in turn, can produce biases and errors when interpreting the results. Simply ignoring repeats is not an option, as this creates problems on its own and may mean that important biological phenomena are not taken into account (Claros et al., 2012).

These occurrences created several problems during the assembly of the resistance/susceptibility locus of 'Lito' using reads coming from BAC clones. These short reads produced by the Illumina NSG platform are commonly unable to span repetitive regions to include at least one unique flanking sequence. In these cases, the origin of a read cannot be precisely determined. The consequent multiple alignments and misalignments lead to problems in downstream analysis, including high fragmented assembly, identical tandem repeats collapse into fewer copies and the wrong estimation of the true copy number.

As a direct result, the primary assembly of the region under study was fragmented in many contigs, whose order could not be safely solved and coverage was not complete along the region tentatively reconstructed.

PacBio sequencing of 'Lito' offered the possibility to solve these problems.

PacBio sequencing provided long reads, that, in spite of the high error rate of the sequences, were merged into large contigs with high confidence due to the high genome coverage . Hence PacBio contigs gave the possibility to close the gaps within the BAC supercontigs assembled with Illumina reads and permitted to scaffold the supercontigs bridging the gaps between them.

The *de novo* assembly of the whole genome of 'Lito' is still largely fragmented, probably due to the high genetic heterozygosity shown by this genotype (Sánchez-Pérez et al., 2005).

Indeed, in cases of diploid sample, the boundaries between homozygous and heterozygous regions result in multiple assembly paths that are hard to resolve, leaving highly fragmented final assemblies (Pryszcz and Gabaldon, 2016).

Curating the whole genome assembly to obtain a high-quality draft genome of 'Lito' is still possible but it was out of our primary interest. Provided that the goal of this study was the fine mapping of the region of resistance/susceptibility to Sharka in apricot, we focused on the reconstruction of the sequence of the region under study. Therefore, only the contigs that covered that region were exctrated from the whole *de novo* assembly.

Limiting the region of interest, we were able to concentrate to overcome many of the obstacles faced by the assembly of Illumina reads.

For this scope, PacBio sequences have been of immense utility to improve the assembly of the resistance/susceptibility region and as a result, the final assembly contained only a single unresolved gap in both haplotypes.

The weakness of PacBio sequencing is the relatively high error rate of the long reads. Since sequencing errors are introduced randomly into the reads generated ad are thus largely non – context specific, they are likely to have minimal effect on the final assembled sequence if sufficient depth of coverage is adopted and error-correction is performed prior to assembly.

The sequences produced with the first 20 and 30 kb libraries, in this respect, were not enough to complete the reconstruction of the region under study. Indeed, the theoretical genome coverage from these two library (18 X) was too low to perform a *de novo* assembly.

Moreover, being ‘Lito’ heterozygous for the locus under study, the coverage of the two haplotypes was further halved. This hampered the discrimination of the real SNPs and indels/deletions from sequencing errors. Indeed, the sequences of BAC clones were fundamental to perform the sequence correction phase in points where PacBio reads have been used to bridge the gaps.

Several inlands of the region assembled with the BAC clone reads in both haplotypes lacked enough coverage and the chromosome-walking approach using PacBio reads should not have produced a whole sequence without introducing many errors in the sequence.

Increasing the depth of coverage with 20 Gb of new PacBio data and the de novo assembly of the genome was therefore the key point of success. PacBio and Illumina assemblies were mostly concordant. Moreover, the assembled contig lengths were higher than the single PacBio reads and this was crucial to bridge the big gaps between the supercontigs in the primary assembly. In any case, Illumina reads from the BAC clones were essential to verify the assessment of Canu assembly keeping out the probability of mis-assemblies between the two haplotypes.

The work resulted in the assembly of almost the entire region for both resistant and susceptible haplotypes. Both haplotypes consist of two continuous sequences, highly curated at nucleotide level, without the presence of ‘N’ filled in the sequences.

Fine mapping of the resistance/susceptibility locus to PPV in apricot was hampered for years by the limited efficiency of phenotypic assessment of the resistance/susceptibility (Llácer *et al*., 2007).

Using the multiple strategies described above let eventually to produce a detailed physical map of PPV resistance/susceptibility region of ‘Lito’.

The rough collinearity of apricot genome with the peach genome has been well-documented (Dirlewanger et al., 2004a; Vera Ruiz et al., 2011). This has been confirmed by the comparison between resistance/susceptibility ‘Lito’ assembled sequences and peach. Indeed, the alignment of apricot sequences to the region between 6.60 and 8.85 Mb of the Peach genome (v.2) showed synteny between the two species, except for the small inversion found in apricot with respect to peach genome.

The percentage of similarity at sequence level was between 80% and 95%. However, this appreciable level of similarity did not allow the use of peach sequence as a guide-reference for the apricot genome assembly at fine level, confirming the difficulties met in the first part of the work.

By the other hand, focusing onto the core objective of this work, aligning the sequence of resistant haplotype onto the susceptible one has permitted to understand that the two region have a high level of similarity.

The gene prediction and annotation of the region, that will be described in the next chapter, has taken great advantage from the goodness of the assembly of the resistant and susceptible haplotypes discussed in this chapter.

CHAPTER 3 - GENE PREDICTION AND GENOME ANNOTATION

INTRODUCTION

Genome annotation is the process of taking the raw DNA sequence produced by the genome-sequencing and adding the layer of analysis and interpretation necessary to extract its biological significance and place it into the context of our understanding of biological processes (Stein, 2001).

Generally, genome annotation of gene structure and function is divided into two distinct phases. In the first phase, the occurrence of genes and regulatory regions are predicted on the base of key features like the occurrence of ORFs (Open Reading Frames), stop codons etc. and validated by aligning expressed sequence tags (ESTs), proteins, coding DNA sequences (CDS), and RNA-seq data to the sequence. In the second phase, this information is synthesized into the gene annotation, that is the analysis of structural gene composition (introns, exons, alternative splicing etc.) and functional meaning of the coding sequence.

The flow-chart of gene prediction and annotation is provided more in details in the following paragraphs.

The process starts with the repeat identification and masking. This step is important because repeats left unmasked can seed millions of spurious BLAST alignments producing false evidence for gene annotations. The term 'masking' simply means transforming every nucleotide identified as a repeat to an 'N' or, in some cases, to a lower case a, t, g, or c ("soft-masked"), so that the repeat is no longer considered by the software.

After repeat masking, the following step involves the use of *ab initio* gene predictors or evidence-driven gene predictors. The first use mathematical models to identify genes and to determine their intron-exon structures. The greatest advantage of *ab initio* gene predictors for annotation is that they do not need any external evidence to identify a gene or to determine its intron-exon structure. However, this kind of tools has practical limitations, indeed most gene predictors find the single most likely coding sequence (CDS) and does not report untranslated regions (UTRs) or alternatively spliced transcripts. Moreover, *ab initio* gene predictors take organism-specific genomic traits information like codon frequencies and distributions of intron-exon lengths to determine

intron-exon structures. This could be an issue because unless the genome under study is very closely related to an organism for which precompiled parameter files are available, the gene predictor needs to be trained on the genome that is under study, as even closely related organisms can differ with respect to intron lengths, codon usage and GC content.
Instead, evidence-driven gene predictors (in contrast to *ab initio*) compare the sequence of interest to available reference annotations or external evidence in their prediction. The first step of these kinds of predictors involves the alignment of proteins, ESTs, CDS and RNA-seq data to the genome assembly. These sequences include previously identified transcripts and proteins from the organism whose genome has being annotated or other correlated organisms.
Evidence-driven gene prediction has great potential to improve the quality of the gene prediction in newly sequenced genomes compare to *ab-initio* predictions, but it can be difficult to use because it requires a lot of specialized software able to align the evidences against the genome under study, identify the splice sites, assemble and post-process the evidence before an outline of these data can be passed to the gene finder.
In order to identify the candidate gene/s related with resistance/susceptibility to Sharka disease in apricot, the MAKER pipeline (Yandell Lab, Institute of Human Genetics, University of Utah, U.S.A.) was used for gene annotation of the sequences of both resistant and susceptible haplotypes.
MAKER identifies repeats, aligns ESTs and proteins to a genome, produces *ab initio* gene predictions, and automatically synthesized these data into gene annotations having evidence-based quality indices.

MATERIALS and METHODS

Evidence Sources

Sequence evidences used for annotation by MAKER consisted of SwissProt protein data and EST, cDNA sequences, transcript assemblies and RNA-seq derived from publicly available data sets.

SwissProt data files containing protein sequences from *Prunus, Arabidopsis, Solanum,* and *Nicotiana* taxa were extracted from UniProt.

Files of EST sequences of *Prunus, Arabidopsis, Solanum* and *Nicotiana* were downloaded from the National Center for Biotechnology Information (NCBI) and a EST database that offers a collection of short single-read transcript sequences from GenBank respectively. CDS sequence files of *P. sibirica* and *P. mandshurica*, two species closely related to apricot, were downloaded from NCBI as well.

Files of repetitive elements, primary transcripts, predicted gene transcripts and predicted gene peptides of *P. persica* were obtained from the Peach Genome Browser of IGA.

Three RNA-Seq data sets from different *P. armeniaca* tissue, and two RNA-Seq data sets from different *P. mume*, the Japanese apricot, were extracted from the NCBI Short Read Archive.

The reads were first checked using the quality control tool for high throughput sequence data FastQC.

Then, the files were searched for possible contaminants with bbduk2 (https://github.com/BioInfoTools/BBMap/blob/master/sh/bbduk2.sh).

Reads were cleaned, masking possible residuals of adapter sequences using cutadapt https://cutadapt.readthedocs.io/en/stable/), trimmed by quality and filtered by possible contaminants using ERNE-FILTER (erne-soruceforge.net).

A reference of 'Lito' whole genome was created by joining all contigs from Canu 'Lito' whole genome assembly reads in a multifasta file. The contigs covering the region under study were removed from this file and replaced with the assembled sequences of the resistance/susceptibility regions of 'Lito'. This reference file was indexed using Bowtie2 (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml), which uses a data structure called FM index to store the reference genome sequence and allows it to be searched

rapidly. Bowtie doesn't allow alignments between a read and the genome that contains large gaps, consequently it cannot align reads that span introns. Hence, reads from each RNA-Seq data file were aligned against the reference file using TopHat (http://ccb.jhu.edu/software/tophat/index.shtml).

The next step was to assembly the individual transcripts from RNA-seq reads that had been aligned to the genome. This was done using Cufflinks (http://cole-trapnell-lab.github.io/cufflinks/), which also permits to align all assembled transcripts to the reference and to annotate them in a GFF3 file required by MAKER as input file.

MAKER pipeline

MAKER (http://www.yandell-lab.org/software/MAKER.html) is an annotation tool that allows to identify repeats, align ESTs, CDS, RNA-seq data and proteins to a genome, produces *ab initio* gene predictions, and automatically synthesize these data into gene annotations having evidence-based quality indices.

MAKER has a modular architecture (fig. 34) that allows breaking the annotation process into a series of five discrete activities: compute, filter/cluster, polish, synthetize, and annotate (Cantarel *et al*., 2008).

During the compute phase MAKER uses BLAST (Altschul *et al*. 1990; Korf *et al*. 2003) and RepeatMasker (http://www.repeatmasker.org/) to screen the genome for low-complexity repeats and soft-mask these regions. BLASTX is also used together with an internal library of transposon and viral encoding proteins to identify mobile elements.

This process is crucial in producing high-quality gene annotations. When not adequately masked, portions of transposable elements can be erroneously included in annotation of neighboring protein-coding genes. In the case of apricot, repetitive elements have not been yet analysed and classified. Therefore, repetitive elements library identified from peach genome, which is closely related to apricot genome, was provided to MAKER.

After repeat masking, BLAST was used to identify EST, mRNAs, and proteins with significant similarity to the input genomic sequence.

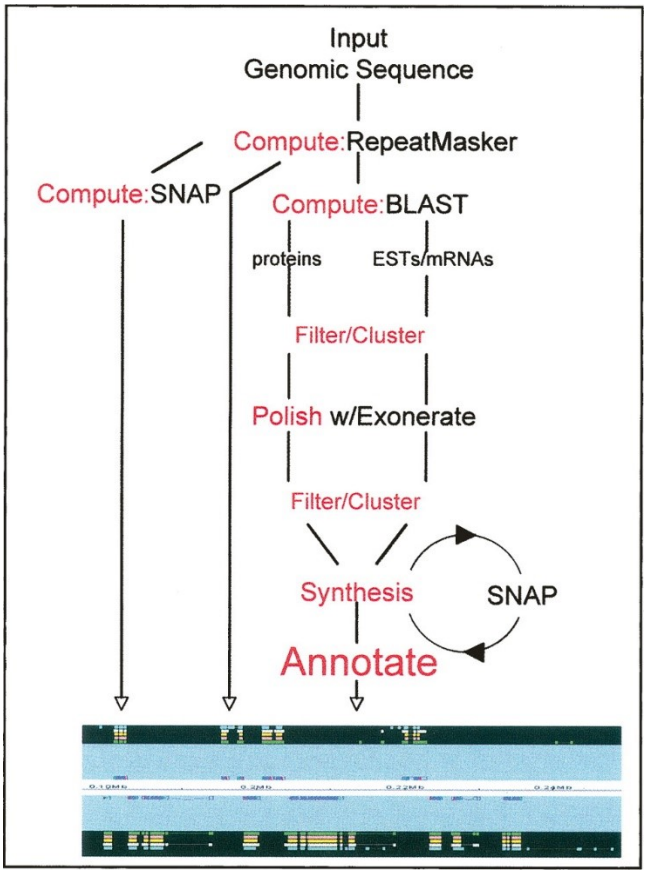


Fig. 34 – MAKER Overview. MAKER uses four external executables. RepeatMasker, BLAST, SNAP, and Exonerate. Actions corresponding to the five basic steps of automatic annotation are shown in red (Cantarel *et al*., 2008).

During filter/cluster phase, marginal predictions and sequence alignments on the bases of scores and percent identities are filtered and the remaining data are clustered against the genomic sequence under processing to identify overlapping alignments and predictions. In the clustering phase, different computational results are grouped into a single cluster of data, all of which support the same gene or transcript, and identify redundant evidence. Since BLAST doesn't take splice sites into account, MAKER exploits Exonerate (Slater and Birney 2005), a "splice-site aware alignment" algorithm to realign matching and highly similar proteins, ESTs, mRNAs to the genomic input sequence and polish the result. Exonerate, taking splice-sites into account, provides MAKER with information about splice donors and acceptors.

Of all forms of evidence, RNA-Seq data have the greatest potential to improve the accuracy of gene annotations, as these data provide copious evidence for better delimitation of exons, splice-sites and alternatively spliced exons.
Once a set of ESTs, transcripts and proteins alignment have been identified, positions on the genomic input sequence upstream and downstream of the alignments are labeled as possible intergenic regions. Those bases on the genomic input sequence that fall between exons are labeled as putative introns, and base overlapping the protein alignments are labeled as putative translated sequences.
For each of these nucleotides on the query sequence a score was calculated, based on the percentage of similarity of the alignment, type of alignment and a query nucleotide positions within the alignment.

Gene prediction

MAKER uses two types of evidences, intrinsic and extrinsic, to generate gene annotations. Extrinsic evidences have been described in the previous paragraph. Intrinsic evidences consist in start and stop codons and intron-exon boundaries predicted from gene predictors. The two gene predictors used by MAKER are SNAP and AUGUSTUS. Both are *ab initio* gene prediction program and come with pre-calculated parameter files that contain such information for different classic genomes, such as *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophyla melanogaster*, humans and mice. In our case, for both *ab initio* gene predictors, *Arabidopsis thaliana* was used as a model.
SNAP and AUGUSTUS offer the possibility to combine *ab initio* gene prediction with evidence-driven gene prediction. Moreover, SNAP can use external evidence to improve the accuracy of its prediction.
By default, a gene model must have half of its splice sites confirmed by an EST/mRNA-seq alignment, half of its exons must overlap an EST/mRNA-seq alignments.
In this way, each synthesis-generated SNAP prediction is checked against all ESTs and mRNAs, and 5’ and 3’ UTRs consistent with the prediction are identified based upon their coordinates relative to the predicted coding exons. The coordinates of the SNAP

prediction are then altered to include these regions. Finally, computed evidence supporting each exon is added, and alternatively spliced forms are documented.
Putative gene functions were added to the annotated genes using a protocol provided by MAKER. This protocol uses BLAST tool and the well-curated UniProt/Swiss-Prot set of proteins to assign putative functions to newly annotated genes.
Interproscan was used to identify functional domains within the annotated genes.

Assessing annotation quality

MAKER assigns to each annotation an Annotation Edit Distance (AED) score, which can be used to measure the congruency between an annotation and its supporting evidence.
AED is based on three measures of gene-finder performance, which are sensitivity, specificity and accuracy.
Sensitivity (SN) is the fraction of the reference feature that is predicted by the gene predictor and is calculated as SN = TP / (TP + FN), where TP are true positives and FN are false negatives.
Specificity (SP) is the fraction of the prediction overlapping the reference feature and is calculated as SP = TP / (TP + FP), where TP are true positives and FN are false negatives.
Sensitivity and specificity can be combined into a single measure called accuracy (AC) with this expression AC = (SN + SP) / 2.
These three measures can be combined to calculate AED (where AED = 1 – AC) which is used to compare two annotations to one another.
An AED of zero denotes perfect concordance with the available evidence and a value of one indicates a complete absence of support for the annotated gene model. In other words, the AED score provides a measure of the congruency of each annotated transcript, with its supporting evidence.

Analisys of the genes in the hot region of resistance

After the automatic gene prediction and annotation, I concentrate my efforts in the analysis of the predicted genes between and near the markers s1_7983920 and PGS1-24.

The analysis was focused in this region because recombinant with susceptible phenotypes (fig. 11 – chapter 1) should confine the hot region of resistance within those markers.
The analysis provided first the evaluation of the gene content and the breakdown of the genes by gene function. Then, each transcript of both the haplotypes were checked using ExPAsy translate tool (https://web.expasy.org/translate/) to verify the presence of start and stop codons within the predicted transcripts.
After this work, shared genes between resistant and susceptible haplotypes were compare using EMBOSS Needle (https://www.ebi.ac.uk/Tools/psa/emboss_needle/). This tool uses a Needleman-Wunsch alignment algorithm to find the optimum alignment (including gaps) of two sequences along their entire length at both nucleotide and protein level.

RESULTS

Reference of 'Lito' Whole genome

From the whole genome assembly of 'Lito', contigs spanned within the region under study (fig. 22 – chapter 2) were removed from the assembly and replace with the assembled sequences of the region of interest of both resistant and susceptible chromosomes, that is the sequences chrR_1, chrR_2, chrS_1 and chrS_2, being the region splitted in two sequences separated by a gap in both haplotypes. The final reference of 'Lito' whole genome consists of 3,746 contigs. The list of the removed contigs is reported in table 14.

Tab. 14 – List of contigs removed from the assembly. Contig length is reported in bp.

CONTIG NAME	Contig lenght
tig00003307	124,503
tig00000226	54,348
tig00000230	118,174
tig00000233	42,993
tig00000251	37,239
tig00000820	57,458
tig00000968	81,250
tig00002636	94,459
tig00003177	84,498
tig00003227	83,726
tig00004116	55,753
tig00004127	56,438
tig00005970	15,107
tig00006051	275,939
tig00006052	80,256
tig00006053	401,061
tig00006054	53,910
tig00006055	934,235
tig00006623	243,369
tig00060166	16,431

This reference was used to align each RNA-Seq data file. Since the RNA-Seq data don't come from 'Lito' but other cultivars, the presence of the entire 'Lito' genome as reference

helps the alignment of the reads without forcing the alignment only against the region under study.

Automatic annotation of resistance/susceptibility locus in 'Lito'

MAKER pipeline provided the annotation of 388 genes within the region of study.
In particular, 41 genes on the chrR_1 sequence, 33 genes on the chrS_1 sequence, 158 genes on the chrR_2 sequence and 156 genes on the chrS_2 sequence were annotated.
The graph in fig. 35 represents the cumulative distribution function curve of the annotated transcripts based on AED. This is a simple way to show the quality level of the annotated transcripts. The curve shows that approximately 80% of the annotations have AEDs less than 0.4. This means that 310 of the annotated gene are highly supported by the evidences used to perform the prediction.

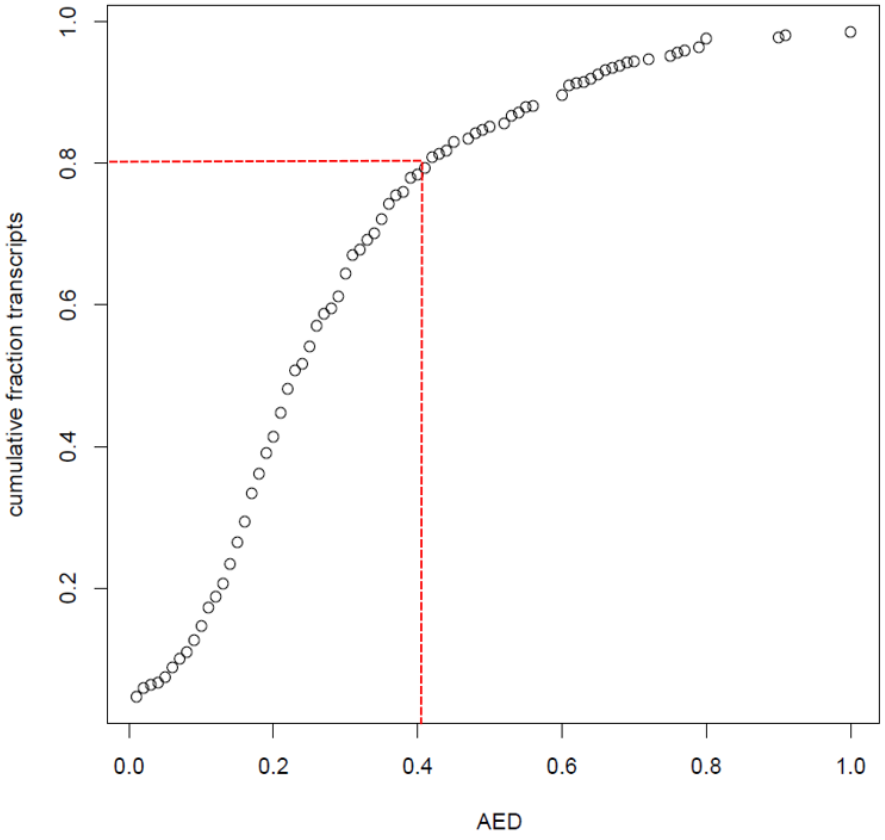


Fig. 35 – Distribution curve of the cumulative annotated transcripts based on AED.

The list of annotated genes for both haplotypes are reported in tables 4 and 5 (supplementary materials) and information includes their relative start/end positions. Genes shared by both resistant and susceptible haplotypes are reported in the same line to highlight the differences between them. In red are reported the molecular markers of the region.

Analisys of the genes in the hot region of resistance

The analysis of the genes annotate in the region between and near the markers s1_7983920 and PGS1-24 reveals that 41 genes and 38 genes are present in the resistant and susceptible haplotype respectively. The genes of the region are reported in table 15. Among the genes of the region, two genes are only present in the resistant haplotype (a protein of unknown fuction and a protein like SLX1 - Structure-specific endonuclease subunit SLX1) and one gene (protein of unknown function) is only present in the susceptible haplotype (fig. 36).

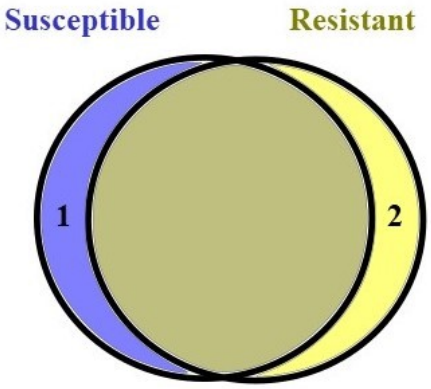


Fig. 36 – Venn diagram of the annotated genes among the hot region of Sharka resistance.

Whithin this restrict region, we found: genes that belong to S-adenosyl/methionine family protein, several genes with kinase activity, a LEA (Late embryogenesis abundant) protein, a gene involved in the constitution of the ribosome, some gene involved in biological process and DNA repair activity, several transcription factor, two PPR-like protein, two Pleiotropic drug resistance, an Ubiquitin-like superfamily protein, three proteins with unknown function and a cluster of MATH/TRAF-like family proteins (fig. 37).

Tab. 15 – Annotated genes in the hot region of resistance within the markers s1_7983920 and PGS1-24**.**

start	end		aa residues	chrS_2	chrR_2	start	end		aa residues
1135079	1136631	+	323	Similar to Late embryogenesis abundant protein D-29 (Gossypium hirsutum)	Similar to Late embryogenesis abundant protein D-29 (Gossypium hirsutum)	1131680	1133241	+	326
1146188	1150609	+	459	Similar to CIPK1: CBL-interacting serine/threonine-protein kinase 1 (Arabidopsis thaliana)	Similar to CIPK1: CBL-interacting serine/threonine-protein kinase 1 (Arabidopsis thaliana)	1141229	1145650	+	459
					Protein of unknown function	1145953	1146486	-	89
1149067	1154372	-	579	Similar to CRSH: Probable GTP diphosphokinase CRSH%2C chloroplastic (Arabidopsis thaliana)	Similar to CRSH: Probable GTP diphosphokinase CRSH%2C chloroplastic (Arabidopsis thaliana)	1146968	1149422	-	584
1154983	1159483	+	613	Similar to RPL3B: 50S ribosomal protein L3-2%2C chloroplastic (Arabidopsis thaliana)	Similar to RPL3B: 50S ribosomal protein L3-2%2C chloroplastic (Arabidopsis thaliana)	1150023	1154523	+	622
1161539	1163509	-	568	Similar to PME28: Putative pectinesterase/pectinesterase inhibitor 28 (Arabidopsis thaliana)	Similar to PME28: Putative pectinesterase/pectinesterase inhibitor 28 (Arabidopsis thaliana)	1156572	1158542	-	568
1164638	1174692	+	1074	Similar to PCMP-H61: Pentatricopeptide repeat-containing protein At5g66520 (Arabidopsis thaliana)	Similar to PCMP-H61: Pentatricopeptide repeat-containing protein At5g66520 (Arabidopsis thaliana)	1159682	1170025	+	1131
1175967	1178974	-	380	Similar to PDK: [Pyruvate dehydrogenase (acetyl-transferring)] kinase%2C mitochondrial (Arabidopsis thaliana)	Similar to PDK: [Pyruvate dehydrogenase (acetyl-transferring)] kinase%2C mitochondrial (Arabidopsis thaliana)	1170500	1173508	-	380
1184140	1188661	+	271	Similar to NPSN13: Novel plant SNARE 13 (Arabidopsis thaliana)	Similar to NPSN13: Novel plant SNARE 13 (Arabidopsis thaliana)	1178660	1183193	+	271

Tab. 15 – Annotated genes in the hot region of resistance within the markers s1_7983920 and PGS1-24 (continue)**.**

start	end		aa residues	chrS_2	chrR_2	start	end		aa residues
1189319	1191028	+	334	Similar to Slx1b: Structure-specific endonuclease subunit SLX1 (Mus musculus)	Similar to Slx1b: Structure-specific endonuclease subunit SLX1 (Mus musculus)	1183859	1185748	+	278
1191566	1192624	-	352	Similar to NAC071: NAC domain-containing protein 71 (Arabidopsis thaliana)	Similar to NAC071: NAC domain-containing protein 71 (Arabidopsis thaliana)	1186082	1187140	-	352
1194146	1207102	-	1173	Similar to ZRANB3: DNA annealing helicase and endonuclease ZRANB3 (Bos taurus)	Similar to smarcal1: SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily A-like protein 1 (Danio rerio)	1188993	1201785	-	1229
1208240	1209689	+	197	Protein of unknown function	Protein of unknown function	1202864	1204313	+	197
1210055	1211113	-	352	Protein of unknown function (Interproscan: NAC domain)	Protein of unknown function (Interproscan: NAC domain)	1204679	1205737	-	352
1212737	1216222	-	373	Similar to At3g17430: Probable sugar phosphate/phosphate translocator At3g17430 (Arabidopsis thaliana)	Similar to At3g17430: Probable sugar phosphate/phosphate translocator At3g17430 (Arabidopsis thaliana)	1207798	1211265	-	373
1226923	1229393	+	461	Similar to At5g18500: Probable receptor-like protein kinase At5g18500 (Arabidopsis thaliana)	Similar to At5g18500: Probable receptor-like protein kinase At5g18500 (Arabidopsis thaliana)	1220997	1223467	+	461
1230285	1234174	-	356	Similar to PTI1: Pto-interacting protein 1 (Solanum lycopersicum)	Similar to PTI1: Pto-interacting protein 1 (Solanum lycopersicum)	1226262	1227132	-	178
1240912	1254719	+	2098	Similar to SPL1: Squamosa promoter-binding-like protein 1 (Arabidopsis thaliana)	Similar to SPL1: Squamosa promoter-binding-like protein 1 (Arabidopsis thaliana)	1234864	1248671	+	2098

Tab. 15 – Annotated genes in the hot region of resistance within the markers s1_7983920 and PGS1-24 (continue)**.**

start	end		aa residues	chrS_2	chrR_2	start	end		aa residues
1191566	1192624	-	352	Similar to NAC071: NAC domain-containing protein 71 (Arabidopsis thaliana)	Similar to NAC071: NAC domain-containing protein 71 (Arabidopsis thaliana)	1186082	1187140	-	352
1194146	1207102	-	1173	Similar to ZRANB3: DNA annealing helicase and endonuclease ZRANB3 (Bos taurus)	Similar to smarcal1: SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily A-like protein 1 (Danio rerio)	1188993	1201785	-	1229
1208240	1209689	+	197	Protein of unknown function	Protein of unknown function	1202864	1204313	+	197
1210055	1211113	-	352	Protein of unknown function (Interproscan: NAC domain)	Protein of unknown function (Interproscan: NAC domain)	1204679	1205737	-	352
1212737	1216222	-	373	Similar to At3g17430: Probable sugar phosphate/phosphate translocator At3g17430 (Arabidopsis thaliana)	Similar to At3g17430: Probable sugar phosphate/phosphate translocator At3g17430 (Arabidopsis thaliana)	1207798	1211265	-	373
1226923	1229393	+	461	Similar to At5g18500: Probable receptor-like protein kinase At5g18500 (Arabidopsis thaliana)	Similar to At5g18500: Probable receptor-like protein kinase At5g18500 (Arabidopsis thaliana)	1220997	1223467	+	461
1230285	1234174	-	356	Similar to PTI1: Pto-interacting protein 1 (Solanum lycopersicum)	Similar to PTI1: Pto-interacting protein 1 (Solanum lycopersicum)	1226262	1227132	-	178
1240912	1254719	+	2098	Similar to SPL1: Squamosa promoter-binding-like protein 1 (Arabidopsis thaliana)	Similar to SPL1: Squamosa promoter-binding-like protein 1 (Arabidopsis thaliana)	1234864	1248671	+	2098
1261425	1263177	-	248	Protein of unknown function	Protein of unknown function	1255377	1257129	-	248

Tab. 15 – Annotated genes in the hot region of resistance within the markers s1_7983920 and PGS1-24 (continue)**.**

start	end		aa residues	chrS_2	chrR_2	start	end		aa residues
1274528	1276849	+	501	Similar to EMB2750: Pentatricopeptide repeat-containing protein At3g06430%2C chloroplastic (Arabidopsis thaliana)	Similar to EMB2750: Pentatricopeptide repeat-containing protein At3g06430%2C chloroplastic (Arabidopsis thaliana)	1261520	1263356	+	501
1278010	1285801	+	1453	Similar to PDR3: Pleiotropic drug resistance protein 3 (Nicotiana tabacum)	Similar to PDR3: Pleiotropic drug resistance protein 3 (Nicotiana tabacum)	1264938	1272745	+	1482
1287526	1294482	-	1444	Similar to PDR3: Pleiotropic drug resistance protein 3 (Nicotiana tabacum)	Similar to PDR3: Pleiotropic drug resistance protein 3 (Nicotiana tabacum)	1274477	1281431	-	1348
1297550	1299707	-	393	Similar to METK5: S-adenosylmethionine synthase 5 (Vitis vinifera)	Similar to METK5: S-adenosylmethionine synthase 5 (Vitis vinifera)	1284674	1285855	-	393
1304574	1306456	+	125	Similar to ATG8I: Autophagy-related protein 8i (Arabidopsis thaliana)	Similar to ATG8I: Autophagy-related protein 8i (Arabidopsis thaliana)	1293226	1295113	+	125
1306781	1318698	-	1004	Similar to At3g58210: MATH domain and coiled-coil domain-containing protein At3g58210 (Arabidopsis thaliana)	Similar to At3g58210: MATH domain and coiled-coil domain-containing protein At3g58210 (Arabidopsis thaliana)	1295626	1304726	-	518
					Protein of unknown function (Interproscan: TRAF-like protein)	1305114	1305750	-	85
					Similar to At3g58210: MATH domain and coiled-coil domain-containing protein At3g58210 (Arabidopsis thaliana)	1306469	1311005	-	565

Tab. 15 – Annotated genes in the hot region of resistance within the markers s1_7983920 and PGS1-24 (continue)**.**

start	end		aa residues	chrS_2	chrR_2	start	end		aa residues
1319766	1320352	+	124	Similar to At3g58210: MATH domain and coiled-coil domain-containing protein At3g58210 (Arabidopsis thaliana)	Protein of unknown function	1312087	1312665	+	96
1330290	1331326	+	244	Similar to UBP13: Ubiquitin carboxyl-terminal hydrolase 13 (Arabidopsis thaliana)	Similar to UBP13: Ubiquitin carboxyl-terminal hydrolase 13 (Arabidopsis thaliana)	1314280	1317229	+	270
1339639	1347374	-	729	Similar to UBP12: Ubiquitin carboxyl-terminal hydrolase 12 (Arabidopsis thaliana)	Similar to UBP12: Ubiquitin carboxyl-terminal hydrolase 12 (Arabidopsis thaliana)	1317863	1331925	-	605
1348497	1349955	-	161	Similar to PSMG4: Proteasome assembly chaperone 4 (Homo sapiens)	Similar to PSMG4: Proteasome assembly chaperone 4 (Homo sapiens)	1333386	1334449	-	161
1351151	1358216	-	1623	Similar to GFS12: Protein GFS12 (Arabidopsis thaliana)	Similar to GFS12: Protein GFS12 (Arabidopsis thaliana)	1335701	1342742	-	1667
1359522	1362252	-	404	Similar to ACLA-3: ATP-citrate synthase alpha chain protein 3 (Oryza sativa subsp. japonica)	Similar to ACLA-3: ATP-citrate synthase alpha chain protein 3 (Oryza sativa subsp. japonica)	1344049	1346782	-	404
1364384	1365469	+	361	Similar to NAC071: NAC domain-containing protein 71 (Arabidopsis thaliana)	Similar to NAC071: NAC domain-containing protein 71 (Arabidopsis thaliana)	1350553	1351611	+	352
1371430	1372823	-	122	Protein of unknown function					

Tab. 15 – Annotated genes in the hot region of resistance within the markers s1_7983920 and PGS1-24 (continue)**.**

start	end		aa residues	chrS_2	chrR_2	start	end		aa residues
					Similar to SLX1: Structure-specific endonuclease subunit SLX1 (Cryptococcus neoformans var. neoformans serotype D (strain JEC21 / ATCC MYA-565))	1352163	1353885	-	341
1374610	1376367	-	150	Similar to MED21: Mediator of RNA polymerase II transcription subunit 21 (Arabidopsis thaliana)	Similar to MED21: Mediator of RNA polymerase II transcription subunit 21 (Arabidopsis thaliana)	1360497	1364062	-	301
1375823	1377529	+	258	Similar to METTL13: Methyltransferase-like protein 13 (Bos taurus)	Similar to METTL13: Methyltransferase-like protein 13 (Bos taurus)	1363518	1365231	+	258
1379702	1387111	+	1012	Similar to CMTA1: Calmodulin-binding transcription activator 1 (Arabidopsis thaliana)	Similar to CMTA1: Calmodulin-binding transcription activator 1 (Arabidopsis thaliana)	1367497	1374906	+	1012

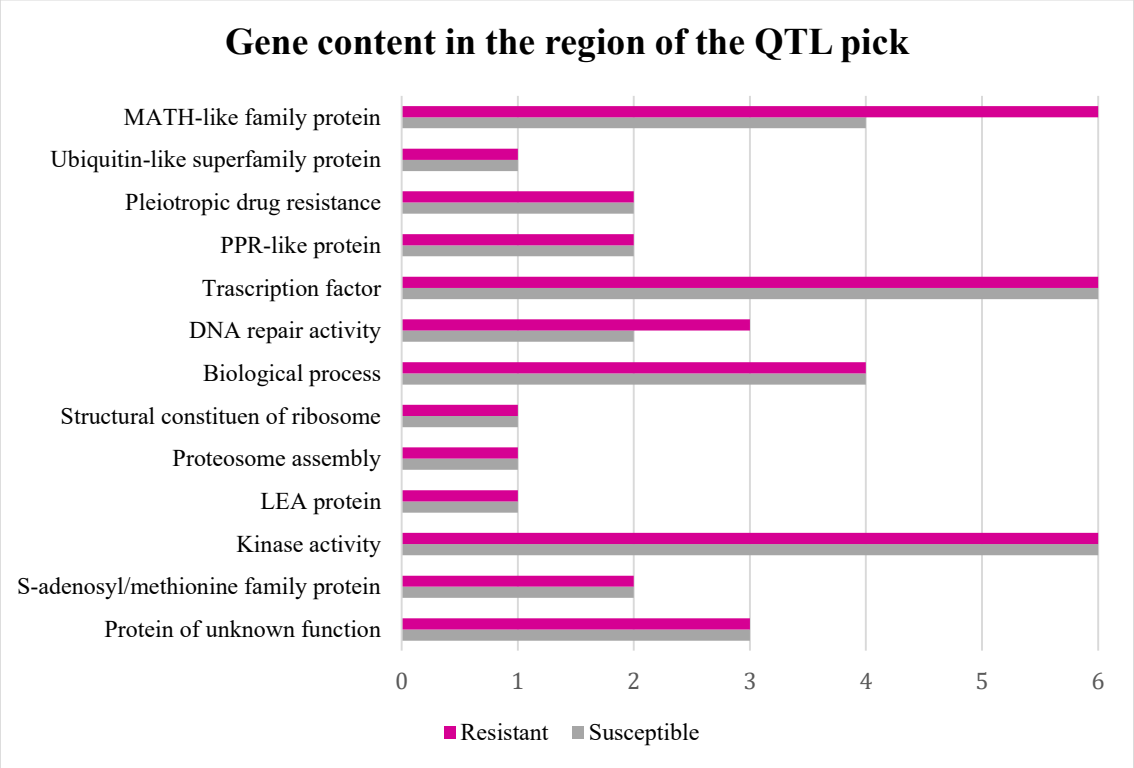


Fig. 37 – Annotated genes of the hot region for Sharka resistance divided by function.

After this work, share genes between resistant and susceptible haplotypes were compared. Comparison allowed the identification of the predicted proteins with 100% of sequence identity and those showing differences.

The analysis was focused first on the genes with higher differences between resistant and susceptible haplotype.These genes are reported in table 16.

The above alignments underlines the differences between the proteins in the susceptible and resistant haplotype.

Table 16 – Genes with higher differences in terms of protein sequence between resistant and susceptible haplotype. Genes are ordered based on their start/end position. In the table, for each genes aa residues and strand (+/-) of prediction are reported. In addition, A.thaliana Gene ID for each predicted gene is specified.

start	end		A.thaliana Gene ID	aa residues	Susceptible haplotype	Resistant haplotype	aa residues		start	end
1189319	1191028	+	NA	334	Similar to Slx1b: Structure-specific endonuclease subunit SLX1 (Mus musculus)	Similar to Slx1b: Structure-specific endonuclease subunit SLX1 (Mus musculus)	278	+	1183859	1185748
1230285	1234174	-	AT2G47060	356	Similar to PTI1: Pto-interacting protein 1 (Solanum lycopersicum)	Similar to PTI1: Pto-interacting protein 1 (Solanum lycopersicum)	178	-	1226262	1227132
1306781	1318698	-	AT3G58210	1004	Similar to At3g58210: MATH domain and coiled-coil domain-containing protein At3g58210 (Arabidopsis thaliana)	Similar to At3g58210: MATH domain and coiled-coil domain-containing protein At3g58210 (Arabidopsis thaliana)	518	-	1295626	1304726
						Protein of unknown function (Interproscan: TRAF-like protein)	85	-	1305114	1305750
						Similar to At3g58210: MATH domain and coiled-coil domain-containing protein At3g58210 (Arabidopsis thaliana)	565	-	1306469	1311005
1319766	1320352	+	AT3G58210	124	Similar to At3g58210: MATH domain and coiled-coil domain-containing protein At3g58210 (Arabidopsis thaliana)	Protein of unknown function	96	+	1312087	1312665
1330290	1331326	+	AT3G11910	244	Similar to UBP13: Ubiquitin carboxyl-terminal hydrolase 13 (Arabidopsis thaliana)	Similar to UBP13: Ubiquitin carboxyl-terminal hydrolase 13 (Arabidopsis thaliana)	270	+	1314280	1317229
1339639	1347374	-	AT5G06600	729	Similar to UBP12: Ubiquitin carboxyl-terminal hydrolase 12 (Arabidopsis thaliana)	Similar to UBP12: Ubiquitin carboxyl-terminal hydrolase 12 (Arabidopsis thaliana)	605	-	1317863	1331925
1364384	1365469	+	AT4G17980	361	Similar to NAC071: NAC domain-containing protein 71 (Arabidopsis thaliana)	Similar to NAC071: NAC domain-containing protein 71 (Arabidopsis thaliana)	352	+	1350553	1351611
1374610	1376367	-	AT4G04780	150	Similar to MED21: Mediator of RNA polymerase II transcription subunit 21 (Arabidopsis thaliana)	Similar to MED21: Mediator of RNA polymerase II transcription subunit 21 (Arabidopsis thaliana)	301	-	1360497	1364062

Protein sequence similar to S1x1b annotated in resistant and susceptible haplotype. Differences between the two proteins are highlighted in yellow.

Line 1: chrR_2 - Similar to Slx1b: Structure-specific endonuclease subunit SLX1

Line 2: chrS_2 - Similar to Slx1b: Structure-specific endonuclease subunit SLX1

```
chrR_2      1 MGQRRKIGSEIPETLIEEEEETEEGRFFACYLLTSRSPRYKGHTYIGAWG     50
              |||||||||||||||.||||||||||||||||||||||||||||...
chrS_2      1 MGQRRKIGSEIPETLIAEEEETEEGRFFACYLLTSRSPRYKGHTYIGFTV     50

chrR_2     51 NKRAHLNSVPLFVLIWVILIQIHSEPTHGEIAQGAWRTKRKRPWEMVLCI    100
              |.|..                |:.|    :||||||||||||||||||||
chrS_2     51 NPRRR---------------IRQH----NGEIAQGAWRTKRKRPWEMVLCI     82

chrR_2    101 YGFPTNVSALQFEWAWQHPTVSKAVRQAAASFKSLRGLVSKIKLAYTMLT    150
              ||||||||||||||||||||||||||||||||||||||||||||||||||||
chrS_2     83 YGFPTNVSALQFEWAWQHPTVSKAVRQAAASFKSLRGLVSKIKLAYTMLT    132

chrR_2    151 LPPWQSLNITVKFFSTQYTKHSAGCPRLPEQMKVKVCSMDELPSCTKLSD    200
              ||||||||||||||||||||||||||||||||:|||||||||||||||||
chrS_2    133 LPPWQSLNITVKFFSTQYTKHSAGCPRLPEQMKVEVCSMDELPSCTKLSD    182

chrR_2    201 DLLENKDDWCHERECDEDMNSSTLPEETLLDFRTHNSADDQQSDSGIRMN    250
              |||||||:||||||||||||||||||||||||||||||||||||||||||
chrS_2    183 DLLENKDEWCHERECDEDMNSSTLPEETLLDFRTHNSADDQQSDSGIRMN    232

chrR_2    251 EEYGCSKEVGKDEWYNGKECDEAMKDGT----------------------    278
              |||||||||||||||||||||||||..
chrS_2    233 EEYGCSKEVGKDEWYNGKECDEAMKDADDQQDDTGKIINETYGCSEVVGE    282

chrR_2    279 --------------------------------------------------    278

chrS_2    283 DCTEQMALPHLTQKPAREQSTAIVADNDQSPSRSYLRPCGAEVIDLTTPA    332

chrR_2    279 --    278

chrS_2    333 PL    334
```

Slxb1 is a catalytic subunit of the SLX1-SLX4 structure-specific endonuclease that resolves DNA secondary structures generated during DNA repair and recombination processes (http://www.uniprot.org/uniprot/Q8BX32).

It has endonuclease activity towards branched DNA substrates, introducing single-strand cuts in duplex DNA close to junctions with ss-DNA. In addiction, Slxb1 prefers 5'-flap structures, and promotes symmetrical cleavage of static and migrating Holliday junctions (HJs). It also resolves HJs by generating two pairs of ligatable, nicked duplex products.

In the resistant haplotype, the predicted protein has 278 aa, in the susceptible one it has 334 aa. The protein contains a catalytic domain GIT_YIG_SLX1 underlined in red in the alignment. This domain in the susceptible haplotype includes the endonuclease YhbQ, that is missed in the resistant one.

Despite the differences observed between the two haplotypes, the protein being involved in DNA repair processes makes its hypotetical involvement in Sharka resistance unlikely.

Protein sequence similar to PTI1 annotated in resistant and susceptible haplotype. Differences between the two proteins are highlighted in yellow.

Line 1: chrR_2 - Similar to PTI1: Pto-interacting protein 1

Line 2: chrS_2 - Similar to PTI1: Pto-interacting protein 1

```
chrR_2      1 MSCFSCCVQDDIRKASDNGPFVANNSAGSSGGYYHRETAPKDTQTVNILP     50
              ||||||||||||||||||||||||||||||||||||||||||||||||||
chrS_2      1 MSCFSCCVQDDIRKASDNGPFVANNSAGSSGGYYHRETAPKDTQTVNILP     50

chrR_2     51 IAVPAIPVDELKDLTDNFGTKSLIGEGSYGRVYHGVLKSGPAAAIKKLDS    100
              ||||||||||||||||||||||||||||||||||||||||||||||||||
chrS_2     51 IAVPAIPVDELKDLTDNFGTKSLIGEGSYGRVYHGVLKSGPAAAIKKLDS    100

chrR_2    101 SKQPDQEFLSQVSMVSRLKHENVVELVGYCIDGPLRLLAYEYAPNGSLMI    150
              ||||||||||||||||||||||||||||||||||||||||||||||||
chrS_2    101 SKQPDQEFLSQVSMVSRLKHENVVELVGYCIDGPLRLLAYEYAPNGSL--    148

chrR_2    151 FSIKTFLLLIKLLISFTTVAYHLIINPP----------------------    178
                                      |.|::..
chrS_2    149 ----------------------HDILHGQKGVKGAQPGPVLSWVQRVKIAV    177

chrR_2    179 --------------------------------------------------    178

chrS_2    178 GAARGLEYLHEKAQPHIIHRDIKSCNILLFDDDVAKIADFDLSNQAPDMA    227

chrR_2    179 --------------------------------------------------    178

chrS_2    228 ARLHSTRVLGTFGYHAPEYAMTGQLSSKSDVYSFGVVLLELLTGRKPVDH    277

chrR_2    179 --------------------------------------------------    178

chrS_2    278 TLPRGQQSLVTWATPKLSEDKVKQCVDARLNGEYPSKAVAKLAAVAALCV    327

chrR_2    179 ----------------------------    178

chrS_2    328 QYEADFRPNMSIVVKALQPLLNARSGPHH    356
```

PTI1 (Pto – interacting protein 1) is a member of the PTI1-like serine/threonine protein kinases that share strong sequence identity to the tomato PTI1 kinase (http://www.uniprot.org/uniprot/Q41328). This protein is involved in cell surface receptor signaling pathway, protein phosphorylation and response to oxidative stress. In tomato, it is involved in the hypersensitive response (HR)-mediated signaling cascade.
In the resistant haplotype, the predicted protein has 178 aa, while in the susceptible one it has 356 aa. The protein contains a serine/threonine kinase catalytic domain underlined in red in the alignment. This domain in the protein predicted in the susceptible haplotype is predicted from 68 to 348 aa, while it seems to be truncate in the resistant one.

Protein sequence similar to NAC071 annotated in resistant and susceptible haplotype. Differences between the two proteins are highlighted in yellow.

Line 1: chrR_2 - Similar to NAC071: NAC domain-containing protein 71
Line 2: chrS_2 - Similar to NAC071: NAC domain-containing protein 71

```
chrR_2      1 MEESLVPFGFRFRPSDEEIVGSFLYPFLVESKPFMSLYNNFFHACNLFGN     50
              |||||||||||||||||||||||||||||.||||||||||||||||||||
chrS_2      1 MEESLVPFGFRFRPSDEEIVGSFLYPFLVECKPFMSLYNNFFHACNLFGN     50

chrR_2     51 NTEPSEIWKKYGGPQLVDTDLYFISKLKKLTPKRMDRRIGNGGTWSETES    100
              ||||||||||||||||||||||||||||||||||||||||||||||||||||
chrS_2     51 NTEPSEIWKKYGGPQLVDTDLYFISKLKKLTPKRMDRRIGNGGTWSETES    100

chrR_2    101 SKLVHEKVSGNPNPNPIGRKRKFRYENKGSEDHTGWLLDEYSLFDGPKND    150
              |||||||.||||||||||||||||||||||||||||||||||||||||||||
chrS_2    101 SKLVHEKASGNPNPNPIGRKRKFRYENKGSEDHTGWLLDEYSLFDGPKND    150

chrR_2    151 YNQRSYDFDFVICRMRKNDRVGIKATNLKRGSQDKEEKNMTTNKKMKKDD    200
              |||||||||||||||||||||||||||||||||||||.||||||||||||
chrS_2    151 YNQRSYDFDFVICRMRKNDRVGIKATNLKRGSQDKEEKKMTTNKKMKKDD    200

chrR_2    201 QMGSTESSSQQGCSSSPIGGDLVGFDQIDLTIFEENTMADMEQLLGEAWS    250
              ||.||||||||||||||||||||||||||||||||||||||||||||||||||
chrS_2    201 QMRSTESSSQQGCSSSPIGGDLVGFDQIDLTIFEENTMADMEQLLGEAWS    250

chrR_2    251 PSNFEDAVSYDVDPIGETQINFENEENTMADMEQLLGEDWSPSNFENEEN    300
              ||||||||||||||||||||||||||||||||||||||||||||||||||||
chrS_2    251 PSNFEDAVSYDVDPIGETQINFENEENTMADMEQLLGEDWSPSNFENEEN    300

chrR_2    301 TTANMEQLLGEAWSPSNFEN-------VVSHDVDPIGETQSSQLSNW---    340
              ||||||||||||||||||.:       :....:.|:..:.....|.:
chrS_2    301 TTANMEQLLGEAWSPSNFLSCTIFHSFITFFSIIPLIFSYFFTASIYFIH    350

chrR_2    341 SQAILDQLLVGV    352
              |.|.: .||:..
chrS_2    351 SNATI-TLLISC    361
```

NAC domain-containing protein 71 is a transcription factor involved in cell proliferation in incised inflorescence stems. Is is also involved in cellular response to auxin stimulus, multicellular organisms development. NAC acronym is derived from three genes that were initially discovered to contain a particular domain: NAM /for no apical meristem, ATAF1-2 ad CUC2 for cup-shaped cotyledon (Souer *et al.,* 1996; Aida *et al.*, 1997).

NAC proteins commonly possess a conserved NAC domain at the N-terminus. In contrast, the C-terminal regions of NAC proteins are highly divergent and are responsible for the observed regulatory differences between transcriptional activation acrivity of NAC proteins (Nuruzzaman *et al*., 2013).

In the resistant haplotype, the predicted protein has 352 aa, in the susceptible one it has 361 aa. The protein contains a NAM domain, underlined in red in the alignment, which is identical between the two proteins. On the contrary, differences were observed in the C-terminal. Whether these difference are linked to the resistance it is hard to guess.

Protein sequence similar to MED21 annotated in resistant and susceptible haplotype. Differences between the two proteins are highlighted in yellow.

Line 1: chrR_2 - MED21: Mediator of RNA polymerase II transcription subunit 21

Line 2: chrS_2 - MED21: Mediator of RNA polymerase II transcription subunit 21

```
chrR_2      1 MDAISQLQEKVNTIATIAFTTIGTLQRDAPPVRISPNYPESGSGPTPAPA     50
              |||||||||||||||||||||||||||||||||||||||||||||||||||
chrS_2      1 MDAISQLQEKVNTIATIAFTTIGTLQRDAPPVRISPNYPESGSGPTPAPA     50

chrR_2     51 PNPNPNPTPTPAADSDADFAKQPKLMSAELVKAAKQFDALVAALPLSEGG    100
              ||||    ||||||||||||||||||||||||||||||||||||||||||||
chrS_2     51 PNPN----PTPAADSDADFAKQPKLMSAELVKAAKQFDALVAALPLSEGG     96

chrR_2    101 EEAQLKRIAQLEAENDAVGQQLEKQLEAAELELWVYSLGLVNRFGFRERI    150
              ||||||||||||||||||||||||||||||.||
chrS_2     97 EEAQLKRIAQLEAENDAVGQQLEKQLEAAEREL-----------------    129

chrR_2    151 ARGQRVVWTSSRSLFELEETRMKSLVQAAVIDVYVNGSRAESFSIFLLFC    200
                 |.|      |.||.           |||   |..:|..:.|
chrS_2    130 ---QEV-----RELFG----------QAA--DHCLNLKKPF---------    150

chrR_2    201 LSNHHIRSEKGPTQMTTMEDDATRLVIQLEDNLGELRNGIPLIGVLMADK    250

chrS_2    151 --------------------------------------------------    150

chrR_2    251 LPNRGAVEGILRKAWEPFGEVKISVVKDNLFAITVESEDMAGRILERGPW    300

chrS_2    151 --------------------------------------------------    150
```

```
chrR_2    301 AVMGYAFSTHPWEEGMAICHNFFFKKLKYSRGLFKRIAARLYCLLKKFK    349
chrS_2    151 --------------------------------------------------    150
```

MED21 is a mediator of RNA polymerase II transcription subunit 21. It is a component of the Mediator complex, a coactivator involved in transcriptional regulation.
Mediator is recruited to promoters by direct interactions with regulatory proteins and serves as a scaffold for the assembly of a functional preinitiation complex with RNA polymerase II and the general transcription factors.
In addiction MED21 can interact physically with the E3 ligase HUB1 and this interaction may be important in mediation defence response, expecially against fungal pathogens (http://www.uniprot.org/uniprot/C0LU16).
In the resistant haplotype, the predicted protein has 301 aa, in the susceptible one it has 150 aa. The protein contains a MED21 domain, underlined in red in the alignment.
This domain is predicted in the susceptible haplotype from 2 to 141 aa, while it seems shorted in the resistant one (from 2 to 133 aa). In addiction, the susceptible haplotype has a deletion of 4 aa, within the domain, in position 55 compared to the resistant one.
In the case of resistant haplotype, the predicted protein is longer than the susceptible one and contains another domain (DUF4283) underlined in green in the alignment. The function of this domain is unknown. DUF domain family is found in plants and is approximately 100 aa in length. It is possible that this domain is a binding/guiding region.

MATH genes

Within the genes with large differences between resistant and susceptible haplotype, a cluster of MATH-like family proteins is worthy of mention.
MATH – like genes are concentrated in the region between 1,295,626 and 1,331,925 bp (coordinates in the sequence chrR_2) of the resistant haplotype and in the region between 1,306,781 and 1,347,374 bp (coordinates in the sequence chrS_2) in the susceptible haplotype. The landscape of this region is complex as shown in fig. 38.
Comparison between the two sequences shows that resistant and susceptible haplotypes have differences in terms of sequence similarity and display structural variants as well

(fig 38). Indeed, Nucmer plot shows that for the first part of the region the similarity at sequence level is near 90% while in the final part is near 100%.

Moreover, in the susceptible region with respect to the resistant one there is: an insertion of 2,233 bp in position 1,309,560 with a repetitive sequence of 310 bp at the start and end, an insertion of 483 bp in position 1,313,042, another insertion of 6,187 bp in position 1,321,223, an insertion of 245 bp in position 1,328,219 and an insertion of 1,386 bp in position 1,339,179. In the resistant region, there is an insertion of 6,423 bp in position 1,297,982 not present in the susceptible haplotype.

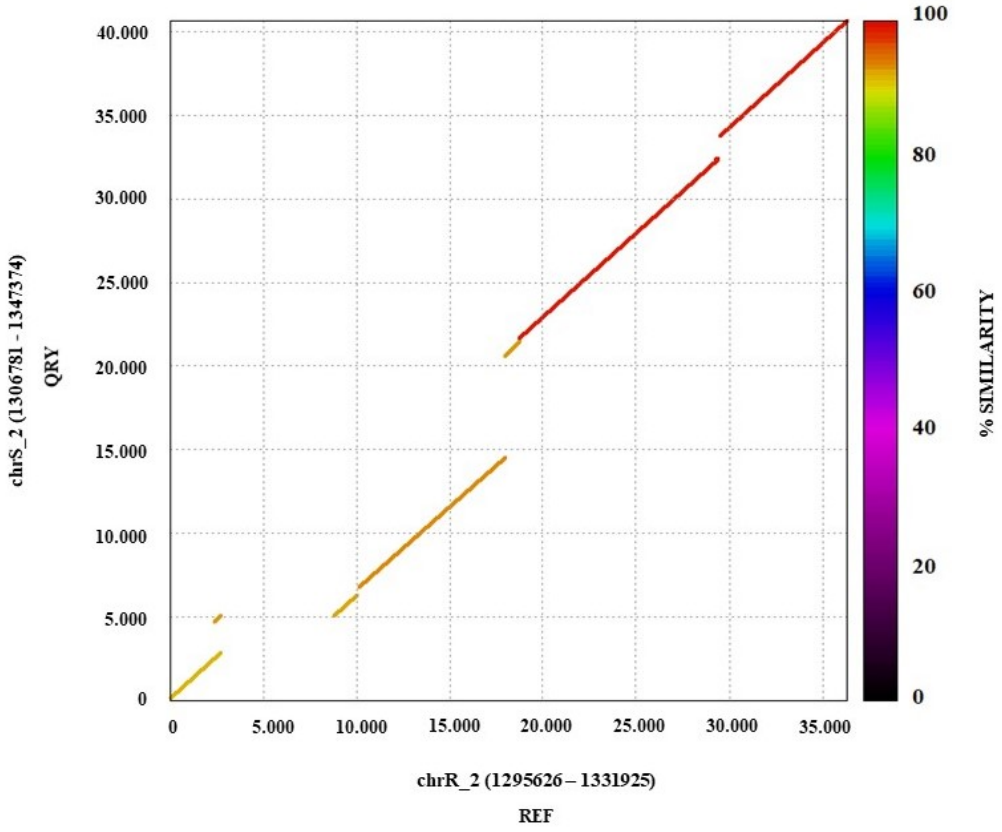


Fig. 38 - Alignment between the REF sequence chrR_2 (1,295,626 – 1,331,925 bp) and QRY sequence chrS_2 (1,306,781 – 1,347,374). Plot was created using NucMer. The percentage of similarity is shown by the color bar.

This complex landscape is reflected in the annotation of the MATH – like proteins. We found 4 MATH-like proteins in the susceptible haplotype and 6 MATH-like proteins in the resistant haplotype (tab 17). A comparison between MATH – like proteins of resistant and susceptible haplotype is reported in supplementary material. A graphical representation of the structure of these genes is reported in fig. 39.

Tab. 17 – MATH – like proteins found in the region under study for the reistant and susceptible haplotypes. The table shows the aa residues and the number of MATH domain of each MATH – like protein.

aa residues	domain	Susceptible haplotype	Resistant haplotype	domain	aa residues
1004	6	Similar to At3g58210: MATH domain and coiled-coil domain-containing protein At3g58210 (Arabidopsis thaliana)	Similar to At3g58210: MATH domain and coiled-coil domain-containing protein At3g58210 (Arabidopsis thaliana)	2	518
			Protein of unknown function (Interproscan: TRAF-like protein)	0	85
			Similar to At3g58210: MATH domain and coiled-coil domain-containing protein At3g58210 (Arabidopsis thaliana)	4	565
124	1	Similar to At3g58210: MATH domain and coiled-coil domain-containing protein At3g58210 (Arabidopsis thaliana)	Protein of unknown function	0	96
244	2	Similar to UBP13: Ubiquitin carboxyl-terminal hydrolase 13 (Arabidopsis thaliana)	Similar to UBP13: Ubiquitin carboxyl-terminal hydrolase 13 (Arabidopsis thaliana)	2	270
729	5	Similar to UBP12: Ubiquitin carboxyl-terminal hydrolase 12 (Arabidopsis thaliana)	Similar to UBP12: Ubiquitin carboxyl-terminal hydrolase 12 (Arabidopsis thaliana)	4	605

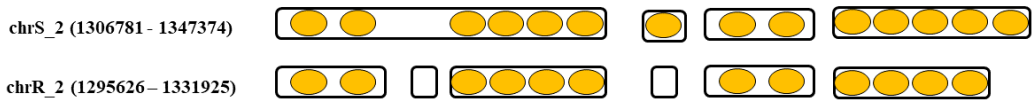


Fig. 39 – Graphical representation of the MATH – like genes organization found in apricot between 1,295,626 and 1,331,925 in the resistant haplotype and between 1,306,781 and 1,347,374 in the susceptible one. Yellow figures represent the MATH domain found within the proteins.

The comparison of the MATH proteins from both the resistant and susceptible haplotype, allowed underlining several differences in the MATH domains both in the number and the sequence of the amino acids.

MATH-like proteins found in this region show similarity with At3g58210 MATH domain, coiled – coil domain containing protein At3g58210 and Ubiquitin carboxyl-terminal hydrolase proteins found in *Arabidopsis thaliana*.

The alternative name of At3g58210 protein is RTM3-like protein (http://www.uniprot.org/uniprot/Q9M2J5). The RTM resistance genes are atypical R genes which restrict the long-distance movement of several potyviruses in *Arabidopsis thaliana*. In this process, viral replication and cell-to-cell movement in inoculated leaves appear unaffected, hypersensitive response (HR) and systemic acquired resistance are not triggered and salicylic acid in not involved (Cosson *et al*., 2012) as observed in case of Sharka infection in apricot.

Ubiquitin carboxyl-terminal hydrolase proteins are protease involved in the jasmonic acid mediated signalin pathway, protein de-ubiquitination and in ubiquitin-dependent protein catabolic processes. The Ubiquitin proteasome system (UPS), in the context of virus – plant interactions, is targeted by many viruses to maintain suitable levels of viral proteins and to achieve a successful infection. However, the UPS also acts as a host defence mechanism to eliminate viral components. (Calil and Fontes, 2017).

Provided that we have no evidence of which MATH domains are expressed in 'Lito' plants challenged with the virus, we are not able to say anything more about the involvement of this gene family in the restriction of the PPV movement in apricot.

DISCUSSION

The accurate identification of the role of genetic determinants in the apricot defense mechanism against Sharka is one of the most intriguing aspects of this work.

The accurate reconstruction of both resistant and susceptible haplotypes of the region that control the resistance to PPV in apricot is the new information compared with the information available in the literature. Consequently, this permitted to annotate separatedly both haplotypic regions and to find out the major differences in their annotated genes.

We annotated the entire assembled region, but we focused on the analysis of the genes between and near the markers s1_7983920 and PGS1-24 of the sequence. This because, according to the phenotyped recombinants (data shown in chapter 1), in this region at least one of the major determinants of Sharka resistance is documented.

Our attention went first on the major differences between the two haplotypes found in the annotated genes. In particular, we focused on a *Structure-specific endonuclease subunit SLX1 like-protein*, a *Pto-interacting protein 1 like-protein*, a *NAC domain-containing like-protein*, a *Mediator of RNA polymerase II transcription subunit 21 like-protein* and a cluster of *MATH-like proteins*.

In the case of the *SLX1 like-protein* and *NAC domain-containing like-protein*, despite the differences observed between the two haplotypes, the proteins, being involved in general DNA repair processes and in cell proliferation mechanisms in incised inflorescence stems, are unlikely involved in the resistance to Sharka.

On the other hand, differences in the other proteins, like, for instance, the PTI-like gene and the MATH – like genes, seem more interesting.

The first gene, the PTI-like gene, encodes for a protein kinase. Over the past years, protein kinases have been studied for their role in the induction of defense responses: they participate in the direct perception of elicitors and avirulence (Avr) products, they mediate signaling required for the induction of defense mechanisms, including the activation of transcription factors, systemic responses, and the function as negative regulator. Yet, they are involved in desensitization of defense responses (Romeis T., 2001). *Pto-interacting*

protein 1 like-protein, in tomato, encodes a serine/threonine kinase that is phosphorylated by *Pto protein* ad is involved in the hypersensitive response.

Plants have developed different mechanisms to resist viruses. Passive resistance generally results from lack of or incompatible interactions between plant and viral factors, causing a block in one of the viral cycle steps, while active resistance is generally triggered by the recognition of the viruses in plants and can be controlled by at least two types of mechanisms. One well-known mechanism is associated with the hypersensitive response or extreme resistance at initial infection sites and is controlled by resistance genes through a gene-for-gene model. The second mechanism concerns the general antiviral defense system of RNA interference, which recognizes and targets the viral nucleic acids.

From a physiological point of view, we remind that resistance of apricot to Sharka is not based on a hypersensitive response triggered by recognition genes. However, due to the multiple roles of the protein kinases, this gene deserves consideration, if not for the large differences observed between the resistant and susvceptible haplotypes.

MATH – like genes in the most recent literature on potyvirus resistance and Sharka in particular, have been reported to play a possible key role (Zuriaga *et al*., 2013; Manuel Rubio *et al*., 2015; Mariette *et al*., 2016).

In our particular case, we found 4 MATH-like genes in the susceptible haplotype and 6 MATH-like genes in the resistant one. The comparison between proteins encoded by these genes revealed several differences not only in the number of the domains but also in the amino acid composition of these proteins.

The MATH genes found in apricot are similar to MATH domain and coiled-coil domain-containing protein At3g58210 and Ubiquitin carboxyl-terminal hydrolase proteins. The alternative name of At3g58210 protein is RTM3-like protein. The RTM genes have already been described as an atypical class of disease resistance genes (Martin *et al*., 2003; Cosson *et al.*, 2010).

RTM genes in *Arabidopsis* does not correspond to any of the known resistance mechanisms described previously. They seem involved in the restriction of long-distance movement of potyviruses.

However, Cosson *et al.* (2012) reported that, as the classical R proteins, many RTM protein domains are involved in protein-protein interaction and some of them are known

to be involved in plant defense, or chaperone activity. In addition, the cluster organization of RTM3 and the RTM3-like genes in the *Arabidopsis* genome, showing signatures of gene duplication and deletion events, presents similarity to the cluster organization of the more typical R genes.

Furthermore, Cosson *et al.* (2012), suggested that mutations in RTM non-functional proteins disrupt interactions necessary for the functionality of these proteins. Another suggestion would be that these mutations alter the stability of these proteins either by destabilizing their structure or by increasing their degradation. In this regard, two of the annotated MATH-like genes are similar to Ubiquitin carboxyl-terminal hydrolase proteins.

Ubiquitin carboxyl-terminal hydrolase proteins are protease involved in the jasmonic acid mediated signaling pathway, protein de-ubiquitination and in ubiquitin-dependent protein catabolic processes.

The ubiquitin proteasome system (UPS) is a complex machinery that plays a central role in a wide range of fundamental plant processes, including degradation and functional modification of cellular proteins, and signaling in response to abiotic and biotic *stimuli.* Different studies report that many positive strand RNA plant viruses interact with UPS to regulate the infection in a manner that promotes replication and movement, but also modulates the levels of RNA accumulation to ensure successful biotrophic interactions. Concomitantly, plants use this pathway as another layer of resistance, mainly targeting viral proteins for degradation through the ubiquitin pathway.

These observations are important to speculate about a role of the MATH-like genes in a UPS complex machinery that could be involved in Sharka resistance.

The restriction of virus movement and the ubiquitination machinery could be correlated. Reichel and Beachy (2000) investigated the role of ubiquitination of movement proteins in Tobacco mosaic tobamovirus (TMV). They suggested that polyubiquitination of movement proteins (MP), which are proteins produced by the plant viruses to facilitate cell-cell-transport, and subsequent degradation by the 26S proteasome may play a substantial role in regulation of virus cell-to-cell spread.

In addition in the region under study, other genes similar to *Autophagy-related protein 8i* (ATG8I) and genes similar to *Proteasome assembly chaperone 4* (PSMG4) were

detected. Both these genes are involved in the proteasome system. Predicted proteins for both these genes are identical in the two haplotypes. However, we do not know anything about their expression level during PPV infection.

Among the genes showing major differences between the two haplotypes there is also another interesting gene, similar to *Mediator of RNA polymerase II transcription subunit 21 like-protein* (MED21). The protein encoded by this gene is a subunit of an evolutionarily conserved multisubunit Mediator complex, regulating the function of RNA polymerase II.

Dhawan *et al.* (2009) investigated about the role of MED21 suggesting that, in addition to the role of this protein during the embryo development, it may be activated by microbial infection and other factors involved in stress signaling. They observed also that HUB1 (histone monoubiquitination1) interacts with MED21 and this interaction may mediate the defense response, especially against fungal pathogens. A dual role of a protein connecting defense and embryo development has been documented also in *Drosophila* (Lemaitre *et al*., 1996), where the mediator subunits acts also as transcriptional factors (Kim *et al*., 2004).

In conclusion, we identified a complex landscape in the region of study, with presence/absence of genes in either haplotype, differences in the number of domains of a gene family, like the MATH-like genes, and large differences also in the gene sequences. In this primary analysis, we focused only on the major differences found within the predicted genes in the resistant and susceptible haplotypes, but the state of our knowledge does not permit to exclude neither the genes with minor differences nor those with 100% of identity between the two haplotypes in this region (41 for the resistant haplotype and 38 for the susceptible one) because we have not evidence of gene expressed in 'Lito' when challenged with the pathogen and the work is going to be completed only with the analysis of gene transcripts (RNAseq).

Certainly the analysis of gene expression would shed light to the differences between resistant and susceptible haplotypes in term of intron/exon composition of genes and alternative splicing, and the rate of transcription efficiency linked to the transcription factors not investigated and therefore not considered up to now as potential players in the mechanism of resistance to Sharka.

This work is being accomplished and will allow to better understand the different landscape of resistance and susceptible haplotypic region of ‘Lito’ and to move to the next natural step, that is the test of each candidate gene through transformation of susceptible genotypes.

LITERATURE

Aida M, Ishida T, Fukaki H, Fujisawa H, Tasaka M, 1997. Genes involved in organ separation in *Arabidopsis*: an analysis of the cup-shaped cotyledon mutant. *Plant Cell* 9, pp. 841-857.

Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ, 1990. Basic local alignment search tool, *Journal of Molecular Biology*, 215, pp. 403-410.

Aranzana MJ, Pineda A, Cosson P, Dirlewanger E, Ascasibar J, Cipriani G, Ryder CD, Testolin R, Abbott A, King GJ, Iezzoni AF and Arùs P, 2003. A set of simple-sequence repeat (SSR) markers covering the Prunus genome, *Theoretical and Applied Genetics*, 106, pp. 819-825.

Atanasoff D, 1932. Plum pox. A new virus disease, *Yearbook university of Sofia*, vol. 11, University of Sofia pp. 49-69.

Audergon JM, 1995. Variety and breeding, *Acta Horticolturae*, 384, pp. 35-45.

Babini AR and Fontana F, 2012. Aggiornamenti sulla tolleranza varietale di cultivar e selezioni avanzate di drupacee nei confronti della Sharka. Convegno "Sharka aggiornamenti tecnici e normativi" Cesena, 23 marzo 2012.

Badenes MI, Cuenca J, Romero C, Martinez J et al., 2002. Description of peach cultivars from Spain: identification of closely related clones by SSR markers, *Acta Horticolturae*, 592, pp. 211-216.

Barba M, Hadidi A, Candresse T and Cambra M, 2011. Plum Pox virus, *Virus and Virus-like disease of Pome and Stone Fruits,* Hadidi A, Barba M, Candresse T andhJelkmann W eds, pp- 185-197, St.Paul, Minnesota, APS Press.

Bassi D, Bellini D, Guerriero R, Monastra Fand Pennone F, 1995. Apricot breeding in Italy, *Acta Horticolturae*, 384, pp. 47-54.

Calil IP and Fontes EPB, 2017. Plant immunity against viruses: antiviral immune receptors in focus, *Annals of Botany* 119, pp. 711 – 723.

Cambra M, Capote N, Myrta A, Llàcer G, 2006. Plum pox virus and the estimate costs associated with sharka disease. *Bulletin OEPP/EPPO* Bulletin 36, pp. 202-204.

Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, Holt C, Alvarado AS and Yandell M, 2008. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes, *Genome Research*, 18, pp. 188-196.

Castelló M, Carrasco J and Vera P, 2010. DNA-binding protein phosphatase AtDBP1 mediates susceptibility to two potyviruses in Arabidopsis, *Plant Physiology*, 153, pp. 1521.1525.

Cervera MT, Riechmann JL, Martin MT and García JA, 1993. 3′-Terminal sequence of the plum pox virus PS and o6 isolates: Evidence for RNA recombination within the potyvirus group, *Journal of General Virology*, 74, pp. 329-334.

Chung BYW, **Miller WA**, **Atkins JF and Firth AE 2008.** An overlapping essential gene in the Potyviridae, *Proceedings of the National Academy of Sciences USA*, 105, pp. 5897-5902.

Claros MG, Bautista R, Guerrero-Fernández D, Benzerki H, Seoane P and Fernández-Pozo N, 2012. Why assembling plant genome sequences is so challenging, *Biology*, 1, pp. 439-459.

Cosson P, Sofer L, Le QH, Leger V, Schurdi-Levraud V, Whitham SA, Yamamoto ML, Gopalan S, Le Gall O, Candresse T *et al*., 2010. RTM3, which controls long-distance movement of potyviruses, is a member of a new plant gene family encoding a meprin and TRAF homology domain-containing Protein, *Plant Physiology*, 154, pp. 222-232.

Cosson P, Schurdi-Levraud V, Le QH, Sicard O, Caballero M, Roux F, Le Gall O, Candresse T, Revers F, 2012. The RTM Resistance to Potyviruses in *Arabidopsis thaliana*: Natural Variation of the RTM genes and evidence for the implication of additional genes, *Plos one*, vol. 7.

Damsteegt VD, Scorza R, Stone AL, Schneider WL, Webb K, Demuth M and Gildow FE, 2007. *Prunus* host range of *Plum pox virus* (PPV) in the United States by aphid and graft inoculation, *Plant Disease Journal*, 91, pp. 18-23.

Dangl JL and Jones JD, 2001. Plant pathogens and integrated defense responses to infection, *Nature*, 411, pp. 826-833

Decroocq V, Foulongne M, Lambert P, Le Gall O, Mantin C, Pascal T, Schurdi-Levraud V, Kervella J, 2005. Analogues of virus resistance genes map to QTLs for

resistance to Sharka disease in Prunusdavidiana, *Molecular Genetics and Genomics*, 272, pp. 680-689.

Dicenta F, Martinez-Gomez P, Burgos L and Egea J, 2000. Inheritance of resistance to Plum pox potyvirus (PPV) in apricot, (*Prunus armeniaca*), *Plant Breeding*, 119, pp. 161-164.

Dirlewanger E, Graziano E, Joobeur T, Garriga-Caldere F, Cosson P, Howad W and Arus P, 2004. Comparative mapping and marker-assisted selection in Rosaceae fruit crops, *Epub*, 26 pp. 9891- 9896.

Dondini L, Lain O, Geuna F, Banfi R, Gaiotti F, Tartarini S, Bassi D and Testolin R, 2007. Development of a new SSRbased linkage map in apricot and analysis of synteny with existing Prunus maps, *Tree Genetics and Genomes*, 3, pp. 239-249.

Dondini L, Lain O, Vendramin V, Rizzo M, Vivoli D, Adami M, Guidarelli M, Gaiotti F, Palmisano F, Bazzoni A, Boscia D, Geuna F, Tartarini S, Negri P, Castellano M, Savino V, Bassi D and Testolin R, 2011. Identification of QTL for resistance to plum pox virus strains M and D in Lito and Harcot apricot cultivars, *Molecular Breeding*, 27, pp. 289-299.

Duprat A, Caranta C, Revers F, Menand B, Browning KS and Robaglia C, 2002. The Arabidopsis eukaryotic initiation factor (iso)4E is dispensable for plant growth but required for susceptibility to potyviruses, *The Plant Journal*, 32, pp. 927-934.

Egea J, Burgos L, Martínez-Gómez P and Dicenta F, 1999. Apricot breeding for Sharka resistance at the CEBAS-CSIC, Murcia (Spain), *Acta Horticolturae*, 488, pp. 153-157.

Faggioli F, Barba M 1997. Valutazione del germoplasma di albicocco per la resistenza alla "Vaiolatura delle drupacee" ("Sharka"), *Riv. Frutticolt*, 7, 8, pp. 73-75.

Fanigliulo A, **Comes S**, **Maiss E**, **Piazzolla P and Crescenzi A, 2003.** The complete nucleotide sequence of Plum pox virus isolates from sweet (PPV-SwC) andsour (PPV-SoC) cherry and their taxonomic relationships within the species, *Archives of Virology,* 148, pp. 2137-2153.

García JA, Glasa M, Cambra M, Candresse T, 2014. Plum Pox Virus and Sharka: A Model Potyvirus and a Major Disease. Available from:

https://www.researchgate.net/publication/257532079_Plum_Pox_Virus_and_Sharka_A_Model_Potyvirus_and_a_Major_Disease [accessed Oct 29 2017].

Garçia JA, Riechmann J L, Lain S, Martin MT, Guo H, Simon L, Fernandez A, Dominguez E, Cervera MT, 1994. Molecular characterization of plum pox potyvirus, *Bulletin OEPP/EPPO* Bulletin 24, pp. 543-553.

Giunchedi L, 2003. Malattie da virus, viroidi e fitoplasmi degli alberi da frutto, *Edagricole*, p. 338.

Glasa M, Candresse T and The SharCo Consortium, 2012. A large scale study of Plum pox virus genetic diversity and of its geographical distribution, *22nd International Conference on Virus and Other Graft Transmissible Diseases of Fruit Crops: Book of Abstracts*, June 3-8, Rome, p. 38.

Glasa M, Candresse T, 2005. Partial sequence analysis of an atypical Turkish isolate provides further information on the evolutionary history of Plum pox virus (PPV). *Virus Research*, 108, pp. 199-206.

Glasa M, Malinowski T, Predajna L, Pupola N, Dekena D, Michalczuk L and Candresse T, 2011. Sequence variability, recombination analysis and specific detection of the W strain of *Plum pox virus*, *Phytopathology*, 101, pp. 980-985.

Glasa M, Palkovics L, Kominek P, Labonne G, Pittnerova S, Kudela O, Candresse T and Šubr Z, 2004. Geographically and temporally distant natural recombinant isolates od Plum pox virus (PPV) are genetically very similar and form a unique PPV subgroup, *Journal of General Virology*, 85, pp. 2671- 2681.

Glasa M, Šubr ZW, 2005. The complete nucleotide sequence of a natural recombinant *Plum pox virus* (PPV) isolate, *Phytopatology ,* 36, pp. 41-46.

Guillet-Bellanguer I, Audergon JM, 2001. Inheritance of the Stark Early Orange apricot cultivar resistance to Plum pox virus, *Acta Horticolturae*, 550, pp. 111-115.

Gupta R, Huang Y, Kieber J and Luan S, 1998. Identification of a dual-specificity protein phosphatase that inactivates a MAP kinase from Arabidopsis, *The Plant Journal*, 16, pp. 581-589.

Howad W, Yamamoto T, Dirlewanger E, Testolin R, Cosson P, Cipriani G, Monforte AJ, Georgi L, Abbott AG and Arùs P, 2005. Mapping with a few plants:

using selective mapping for microsatellite saturation of the Prunus reference map, *Genetics*, 171, pp. 1305-1309.

Hunt M, Newbold C, Berriman M and Otto T, 2014. A comprehensive evaluation of assembly scaffolding tools, *Genome Biology*, 15, (3), p. 42.

Hurtado MA, Romero C, Vilanova S, Abbott AG, Llacer G and Badenes ML, 2002. Genetic linkage map of two apricot cultivars (*Prunus armeniaca* L.) and mapping of PPV (Sharka) resistance, *Theoretical and Applied Genetics*, 105, pp. 182-191.

Ion-Nagy L, Lansac M, Eyquard JP, Salvador B, Garcia JA, Le Gall O, Hernould M, Schurdi-Levraud V and Decroocq V, 2006. PPV long-distance movement is occasionally permitted in resistant apricot hosts, *Virus Research*, 120, pp. 70-78.

James D and Thompson D, 2006. Hosts and symptoms of *Plum pox virus*: ornamental and wild *Prunus* species, *Bulletin OEPP/EPPO,* Bullettin 36, pp. 222-224.

James D and Varga A, 2005. Nucleotide sequence analysis of Plum pox virus isolate W3174: evidence of a new strain, *Virus Research,* 110, pp. 143-150.

Jones JDG and Dangl JL, 2006. The plant immune system. Nature, 444, pp. 323-329.

Joobeur T, Viruel MA, de Vicente MC, Jauregui B, Ballester J, Dettori MT, Verde I, Truco MJ, Messeguer R, Battle I, Quarta R, Dirlewanger E and Arùs P, 1998. Construction of a saturated linkage map for Prunus using an almond 9 peach F2 progeny, *Theoretical and Applied Genetics*, 97 pp. 1034-1041.

Karayiannis I, Thomidis T, Tsaftaris A, 2008. Inheritance of resistance to Plum pox virus in apricot (*Prunus armeniaca* L.), *Tree Genetics and Genomes*, 4, pp. 143-148.

Karayiannis I, Mainou A and Tsaftaris A, 1999. Apricot breeding in Greece for fruit quality and resistance to plum pox virus, *Acta Horticolturae*, 488, pp. 111-117.

Karayiannis I, Thomidis T and Tsaftaris A, 2008. Inheritance of resistance to *Plum pox virus* in apricot (*Prunus armeniaca* L.), *Tree Genetics and Genomics*, 4, pp. 143-148.

Kegler H, Fuchs E, Gru¨ntzig M and Schwarz S, 1998. Some results of 50 years of research on the resistance to Plum Pox Virus, *Acta Virologica*, 42, pp. 200-215.

Kim TW, Kwon YJ, Kim JM, Song YH, Kim SN, 2004. MED16 and MED23 of Mediator are coactivators of lipopolysaccharide- and heat-shock-induced transcriptional activators, *Proc. Natl. Acad. Sci. USA*, 101, pp. 12153-12158.

Koren S, Walenz BP, Berlin K, Miller JR and Phillippy AM, 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation, *Genome Research*, 27, pp. 722-736.

Korf I, Yandell M and Bedell M, 2003. BLAST: An essential guide to the basic local alignment search tool, O'Reilly&Associates Inc, Sebastopol, CA.

Laín S, Riechmann JL &García JA, 1989a. The complete nucleotide sequence of plum pox potyvirus RNA, *Virus Research*, 13 pp. 157-172.

Laín S, **Riechmann JL and García JA, 1989.** The complete nucleotide sequence of plum pox potyvirus RNA, *Virus Research,* 13, pp. 157-172.

Laín S, Riechmann JL, Méndez E &García JA, 1988. Nucleotide sequence of the 3′ terminal region of plum pox potyvirus RNA, *Virus Research*, 10, pp. 325-342.

Lalli DA, Abbott AG, Zhebentyayeva TN, Badenes ML, Damsteegt V, Pola´k J, Krsˇka B and Salava J, 2008. A genetic linkage map for an apricot (*Prunus armeniaca* L.) BC1 population mapping Plum pox virus resistance, *Tree Genetics and Genomes*, 4, pp. 481-493.

Lambert P, Dicenta F, Rubio M and Audergon JM, 2007. QTL analysis of resistance to Sharka disease in the apricot (*Prunus armeniaca* L.) 'Polonais' 9 'Stark Early Orange' F1 progeny, *Tree Genetics and Genomes*, 3, pp. 299-309.

Lambert P, Hagen LS, Arus P and Audergon JM, 2004. Genetic linkage maps of two apricot cultivars (*Prunus armeniaca* L.) compared with the almond Texas 9 peach Early gold reference map for Prunus, *Theoretical and Applied Genetics*, 108, pp. 1120-1130.

Levy L, Damsteegt V, Scorza R, Kolber M, 2000. Plum Pox Potyvirus Disease of Stone Fruits. *APSnetFeatures*, Online.

Llácer G, Cambra M and Lavina A, 1985. Detection of plum pox virus in Spain, *Bulletin OEPP/EPPO*, Bulletin 15, pp. 325-329

Lommel SA, Mccain AH and Morris TJ, 1982. Evaluation of indirect enzyme-linked immunosorbent assay for the detection of plant viruses, *Phytopathology*, 72, pp. 1018-1022.

Maejima K, **Himeno M**, **Komatsu K**, **Takinami Y**, **Hashimoto M**, **Takahashi S**, **Yamaji Y**, **Oshima K and Namba S, 2011.** Molecular epidemiology of *Plum pox virus* in Japan, *Phytopathology*, 101, pp. 567.574.

Maiss E, **Timpe U, Brisske A, Jelkmann W, Casper R, Himmler G, Mattanovich D and Katinger HWD, 1989.** The complete nucleotide sequence of plum pox virus RNA, *Journal of General Virology,* 70, pp. 513-524.

Manoussopoulos IN, Maiss E, Tsagris M 2000. Native electrophoresis and Western blot analysis (NEWeB): a method for characterization of different forms of potyvirus particles and similar nucleoprotein complexes in extracts of infected plant tissues, *Journal of General Virology*, 81, pp. 2295-2293.

Marandel G, Pascal T, Candresse T and Decroocq V 2009a. Quantitative resistance to Plum pox virus in Prunus davidiana P1908 linked to components of the eukaryotic translation initiation complex, *Plant Pathology*, 58, pp. 425-435.

Marandel G, Salava J, Abbott A, Candresse T and Decroocq V, 2009b. Quantitative trait loci meta-analysis of Plum pox virus resistance in apricot (*Prunus armeniaca* L.): new insights on the organization and the identification of genomic resistance factors, *Molecular Plant Pathology*, 10, pp. 347-360.

Mariette S, Wong Jun Tai F, Roch G, Barre A and Decroocq V, 2015. Genome-wide association links candidate genes to resistance to *Plum pox Virus* in apricot (*Prunus armeniaca*), *New Phytologist*, 209, pp. 773-784.

Martin GB, Bogdanove AJ, Sessa G, 2003. Understanding the function of plant disease resistance proteins, *Annual Rev Plant Biol* 54:23-61.

Martinez-Gomez P, Dicenta F and Audergon JM, 2000. Behaviour of apricot (*Prunus armeniaca* L.) cultivars in the presence of Sharka (plum pox potyvirus): a review. *Agronomie*, 20, pp. 407-422.

Martínez-Gómez P, Rubio M, Dicenta F and Gradziel TM, 2004. Utilization of almond as source of plum pox virus resistance in peach breeding, *Acta Horticolturae*, 657, pp. 289-293.

Mavrodieva V, James D, Williams K, Negi S, Varga A, Mock R and Levy L, 2013. Molecular analysis of a Plum pox virus W isolate in plum germplasm hand carried into the USA from the Ukraine shows a close relationship to a Latvian isolate. *Plant Disease Journal*, 97, pp. 44-52

Mohammed N, Akhter MS, Shoshi K, 2013. Roles of NAC transcription factors in the regulation of biotic and abiotic stress responses in plants. *Frontiers in Microbiology,* volume 4, article 248.

Moreno A, **Fereres A and Cambra M, 2009.** Quantitative estimation of plum pox virus targets acquired and transmitted by a single *Myzuspersicae*, *Archives of Virology*, 154, pp. 1391.1399.

Moustafa TA, Badenes ML, Martı´nez-Calvo J and Lla´cer G, 2001. Determination of resistance to Sharka (plum pox) virus in apricot, *Horticoltural Science*, 91, pp. 59-70.

Myrta A, Varga A and James D, 2006. The complete genome sequence of an El Amar isolate of plum pox virus (PPV) and its phylogenetic relationship to other PPV strains, *Archives of Virology,* 151, pp. 1189-1198.

Palkovics L, **Burgyán J and Balázs E, 1993.** Comparative sequence analysis of four complete primary structures of plum pox virus strains, *Virus Genes*, 7, pp. 339-347.

Palmisano F, Boscia D, Minafra A, Myrta A and Candresse T, 2012. An atypical Albanian isolate of Plum pox virus could be the progenitor of the Marcus strain, *22nd International Conference on Virus and Other Graft Transmissible Diseases of Fruit Crops*: *Book of Abstracts*, June 3-8, Rome, p. 33.

Pasquini G and Barba M, 2006. The question of seed transmissibility of *Plum pox virus*, *Bulletin OEPP/EPPO*, Bulletin 36, pp. 287-292.

Pilarova P, Marandel G, Decroocq V, Salava J, Krka B and Abbott AG, 2010. Quantitative trait analysis of resistance to plum pox virus in the apricot F1 progeny 'Harlayne' x 'Vestar', *Tree Genetics and Genomes*, 6, pp. 467-475.

Pryszcz LP and Gabaldon T, 2016. Redundans: an assembly pipeline for highly heterozygous genomes, *Nucleic Acids Research*, 44 (12).

Public Health Security and Bioterrorism Act of 2002.

Rhoads A and Kin Fai Au, 2015. PacBio Sequencing and its Applications, *Genomics Proteomics and Bioinformatics*, 13, pp. 278-289.

Reichel C, Beachy RN, 2000. Degradation of Tobacco mosaic virus movement protein by the 26s proteasome, *Journal of Virology*, 74, pp. 3330-3337.

Riechmann JL, Laín S &García JA, 1989. The genome-linked protein and 5′ end RNA sequence of plum pox potyvirus, *Journal of General* Virology, 70, pp. 2785.2789.

Romeis T, 2001. Protein kinases in the plant defence response, *Current Opinion in Plant Biology,* vol 4, issue 5, pp. 407-414.

Roudet-Tavert G, German-Retana S, Delaunay T, Delécolle B, Candresse T and Le Gall O, 2002. Interaction between potyvirus helper component-proteinase and capsid protein in infected plants, *Journal of General Virology*, 83, pp. 1765-1770.

Roy AS, Smith IM, 1994. Plum pox situation in Europe. *Bulletin OEPP/EPPO* Bulletin 24, pp. 515- 523.

Rubio M, Audergon JM, Martinez-Gomez P and Dicenta F, 2007. Testing genetic control hypotheses for Plum pox virus (Sharka) resistance in apricot, *Scientia Horticolturae*, 112, pp. 361-365.

Rubio M, Rodríguez-Moreno L, Ballester AR, Moura MC, Bonghi C, Candresse T et al., 2015. Analysis of gene expression changes in peach leaves in response to Plum pox virus infection using RNA-Seq. *Molecular Plant Pathology*, 16, pp. 164-176.

Sáenz P, Cervera MT, Dallot S, Quiot L, Quiot JB, Riechmann JL and García JA, 2000. Identification of a pathogenicity determinant of Plum pox virus in the sequence encoding the C-terminal region of protein P3+6K1, *Journal of General Virology*, 81 pp. 557-566.

Sánchez-Pérez R, Ruiz D, Dicenta F, Egea J and Martínez-Gómez P, 2005. Application of simple sequence repeat (SSR) markers in apricot breeding: molecular characterization, protection, and genetic relationships, *Scientia Horticulturae*,103, pp. 305-315

Scholthof KBG, Adkins S, Czosnek H *et al.,* 2011. Top 10 plant viruses in molecular plant pathology. *Molecular Plant Pathology* 12, pp. 938-954.

SharCo database,http://w3.pierroton.inra.fr:8060/

Sheveleva A, Ivanov P, Prihodko Y, James D and Chirkov S, 2012. Occurrence and genetic diversity of Winona-like Plum pox virus isolates in Russia, *Plant Disease Journal*, 96, pp. 1135-1142.

Sicard O, Marandel G, Soriano JM, Lalli DA, Lambert P, Salava J, Badenes ML, Abbott A and Decroocq V, 2008. Flanking the major Plum pox virus resistance locus in apricot with co-dominant markers (SSRs) derived from candidate resistance genes, *Tree Genetics and Genomes*, 2, Vol. 4, p. 359.

Simon L, Saenz P, Garçia JA, 1997. Plum pox virus, *Recent Research Development in Plant Pathology-Filamentous Viruses of Woody Plants*, Veseley,D. and Monette. P. Eds., Research Signpost, Trivandrum, India Chapter 7, pp. 75-86.

Slater GS and Birney E, 2005. Automated generation of heuristics for biological sequence comparison, *BMC Bioinformatics*, 6, p. 31.

Soriano JM, Vera-Ruiz EM, Vilanova S, Martı´nez-Calvo J, Lla´cer G, Badenes ML and Romero C, 2008. Identification and mapping of a locus conferring Plum pox virus resistance in two apricot-improved linkage maps, *Tree Genetics and Genomes*, 4, pp. 391-402.

Souer E, van Houwelingen A, Kloos D, Mol J, Koes R, 1996. The no apical meristem gene of Petunia is required for pattern formation in embryos and flowers and is expressed at meristem and primordia boundaries. *Cell* 85, pp. 159-170.

Stein L, 2001. Genome annotation: From sequence to biology, *Nature Reviews Genetics*, 2(7), pp. 493-503.

Teycheney PY, **Tavert G**, **Delbos R**, **Ravelonandro M and Dunez J, 1989.** The complete nucleotide sequence of plum pox virus RNA (strain D), *Nucleic Acids Research,* **17**, pp. 10115-10116.

Ulubas˘Serçe C, **Candresse T**, **Svanella-Dumas L**, **Krizbai L**, **Gazel M and Çag˘layan K, 2009.** Further characterization of a new recombinant group of *Plum pox virus* isolates, PPV-T, found in orchards in the Ankara province of Turkey, *Virus Research,* 142, pp. 121-126.

Urcuqui-Inchima S, Haenni A, Bernardi F, 2000. Potyvirus proteins: a wealth of functions. *Virus Research*, 74, pp. 157-175.

Vera-Ruiz EM, Soriano JM, Romero C, Zhebentyayeva T, Teril J, Zuriaga E et al., 2011. Narrowing down the apricot Plum pox virus resistance with the peach genome syntenic region, *Molecular Plant Pathology*, 12, pp. 535-547.

Verde I, Abbott AG, Scalabrin S, Jung S, Shu S, Marroni F, Zhebentyayeva T, Dettori MT, Grimwood J, Cattonaro F, Zuccolo A, Rossini L, Jenkins J, Vendramin E, Meisel LA, Decroocq V, Sosinski B, Prochnik P, Mitros T, Policriti A, Cipriani G, Dondini L, Ficklin S, MGoodstein D, Xuan P, Del Fabbro C, Aramini V, Copetti D, Gonzalez S, SHorner D, Falchi R, Lucas S, Mica E, Maldonado J, Lazzari B,

Bielenberg D, Pirona R, Miculan M, Barakat A, Testolin R, Stella A, Tartarini S, Tonutti P, Arús P, Orellana A, Wells C, Main D, Vizzotto G, Silva H, Salamini F, Schmutz J, Morgante M, SRokhsar D, (2013). The high-quality draft genome of peach (Prunus persica) identifies unique patterns of genetic diversity domestication and genome evolution, *Nature Genetics*, pp. 487-494.

Vilanova S, Romero C, Abbott AG, Llacer G and Badenes ML, 2003. An apricot (*Prunus armeniaca* L.) F2 progeny linkage map based on SSR and AFLP markers mapping Plum pox virus resistance and self-incompatibility traits, *Theoretical and Applied Genetics*, 107, pp. 239-247.

Vogel JM, Rafalski A, Powell W, Morgante M, Andre C, Hanafey M and Tingey SV, 1996. Application of genetic diagnostics to plant genome analysis and plant breeding, *Horticoltural Science* 31, pp. 1107-1118

Wetzel T, Candresse T, Ravelonandro M and Dunez J, 1991. A polymerase chain reaction assay adapted to plum pox potyvirus detection, *Journal of Virological Methods*, 33, pp. 355-365.

Zhang HB, Martin GB, Tanksley SD and Wing RA, 1994. Map-based cloning in crop plants: tomato as a model system II. Isolation and characterization of a set of overlapping yeast artificial chromosomes encompassing the jointless locus, *Molecular Genetics and Genomics*, 244, pp. 613-621.

SUPPLEMENTARY MATERIALS

Tab. 1 – List of primers used for the multiplexing genotyping using a primer single base extension chemistry. A tale represented by lowercase characters were added in a few primers to balance the oligonucleotide mass.

SNP	Primer single base extention
s1_5511078	ccCAGTAAGCTTTGGTCACTAAGTT
s1_5540944	ggttAGAGGTTTTTGTCTAAGAGTT
s1_5586095	cAATTTATCAATTTCTATTATTGTCTTG
s1_5616242	AACTATACAAACAGGTTTAAGCATTC
s1_5683686	aaaaTCAAAGTTCAGTTACTGCT
s1_5801422	TGTGTTTCTTACGAAAACATAAT
s1_5820311	TTGATTTATTCTGAACAACTTC
s1_5864975	CTTTCCCCTTGTTTCTTT
s1_5878914	TCAAAGTTCAGTTACTGCT
s1_6127634	TGCAACTGTCCCATCTTTTATCAA
s1_6180800	gtgGTTAGGAGCTCTCCATT
s1_6223532	ggggTAAATCTCAGCCTAATCAGAAAG
s1_6280616	gGAGCCGGAAGCTGCT
s1_6345556	GTGGATTTCATTGACAATTCAT
s1_6422355	ccACACAATGACCCACTAT
s1_6537106	CCCTTCTCGGGAATTTA
s1_6698823	ccCAGGTATGCTCTGTACAC
s1_6761324	aAGAGTTTTTTATTAGACAACACAAG
s1_6828246	GAAAGAGGCCATTTTTTTAG
s1_6965213	TTTATATAGTTTGTTTCCTCCA
s1_7077554	GAATGTAGTGCTGAAAAATTATG
s1_7112235	tGGTCATGACATTTCAATTTTTTT
s1_7217828	ctccGGCTGTATGACCTTCTT

SNP	Primer single base extention
s1_7241764	GAGTTTCGTAACATTGGG
s1_7267206	GGATGAGGAATCATAGATAATAATTTAT
s1_7316247	TAATTGTTCTTTTCACCTTTGAAAT
s1_7442314	AAAACAGTTTCTCGTTCAT
s1_7473604	TCCTTAATCCCAATACAGATA
s1_7505322	gTAGGGGACTTTCTAAGGATG
s1_7526901	TAATGCTAAATGACGCAATAA
s1_7555803	TGTCCAAATGCCAATAATTTATTTTTT
s1_7579835	TATCCTTGGGCCACA
s1_7786880	gacGACAATGGTGTTGTTGAACAT
s1_7805039	tAGCATCAGCTATGCCT
s1_7960013	aTTTTACATCTTTAAATATTGAGGAG
s1_7983920	ctATTTTTATGCAAAACACATGTAG
s1_8042406	TGGAATTCGCATGCT
s1_8105025	CCATTTGTTTCTGTTGCC
s1_8132703	GTTCGACATATGACAAATGG

Tab. 2 – Molecular markers and primer sequences per BAC library screening.

MARKER	TYPE	FORWARD PRIMER	REVERSE PRIMER
s1_6345556	SNP	TGTGCAGTCAAAATTGGTGG	ACAAGATGCCGACATCACTG
s1_6422355	SNP	AAATTCAACTACTGCGGCCT	TTTGAGAGAATTTGGGGAAG
s1_6537106	SNP	CACTTTCGTAGAAGCTCTCC	CATCAAAAAGCCCTAGAACG
Gol061	SSR	TGGCTCAACCACAAAGTGAC	GGAGCTAGTCTTCTGTCCAAGG
s1_6698823	SNP	AAGCCACCCTCATGATGTTG	GGTTAGATAATTTCTTGCGG
s1_6761324	SNP	GTTACCAGCTAACATTCAGAG	TCTACGTACTTTACATAGC
S1-6798SCAR	SCAR	TCTAGATTAGCTTCCCAACtTTCCT	GAAATTTGGAAAACCACGCAACA
S1-6835SCAR	SCAR	TCTCTTATACAAAACAAGTGAAAGCA	TGAAGAAGCTGAGCCTCACT
PGS1.03	SSR	GCTCTCTCCCTGCCATTTTT	CCATCCTCCACTTCTCAACC
s1_6965213	SNP	GTGATTTCGATTTCTTCATGC	GGATGGATGGAATTCAAAGC
S1-6994SCAR	SCAR	CACTACGTGTTCAACCTCCA	AGCATtCAAGGAGCAAGAGT
S1-7045SSR	SSR	gGAGAACACACGCATACATGAT	GATGCTTCACGTCTACTCCAAA
s1_7077554	SNP	AACACGAGAATGTAGTGCTG	CAACGTGGTTTCGGCAAATG
s1_7112235	SNP	CCTTGTGTTAGGTCATGACA	TGCAAGTTGTAAAATGTCG
S1-7164SSR	SSR	TGCGCCATATTATCTTGCCT	GGTGCGAATTCCACACATCA
s1_7217828	SNP	GTGGTACATTACATTCAGGC	TGGTGAAGTCGATAATGATG
S1-7218Scar	SCAR	CAGGTGTGGGTGGGATCTTA	TTTGTCCACCCAACCCCAAT
s1_7241764	SNP	TCACCTGATGATATGACGAG	GTGTCCATCAAATTCTTCCC
S1-7284SSR	SSR	AGACACGCTTTTCTTGCAGG	CCTTCTGGGTTTTAAAGGAAGCT
S1-7361SSR	SSR	ACtACCATGGCTTGACTaGT	TTTCTGGGCTAGGCTGGTTT

MARKER	TYPE	FORWARD PRIMER	REVERSE PRIMER
PGS1.10	SSR	GCCCTTTAATCCCAAGGAAG	GCAGGGCTTGCTCTATTCAC
S1-7418SSR	SSR	GCTCGTGAAACCATGTGAAC	ACGTAGAAGGCcGTCGGTAC
s1_7442314	SNP	GTACTCACTCATATCTTGG	GTAGCAGCAACCATGGATCT
S1-7484SSR	SSR	AATCCCTCTCTTCCCATGGC	GGTCGTCCTCTGCAACAAAT
S1-7518SCAR	SCAR	TCCCTCCTTTGGCTGCAAAA	ATTGTCAGGTGCGAACCCAT
s1_7555803	SNP	GTGCTACTATATCTCATGTCC	TAATTCATCCCCTGGTTTGC
s1_7579835	SNP	GCTGTAATTACCAACTTTTCC	GTGGGTAGAAATGGATAGAG
S1-7700SSR	SSR	AtCATCCCCaTCCACAATTG	TCTACCTAGACGCACAGCCT
S1-7745SSR	SSR	CGAACCtAAACCAGGCTTGT	GTTCTGGCAGATCCCTCAGA
S1-7982SSR	SSR	AAGTCATtACACGGTTCGCT	GGaACCCTAGATTGCATGGA
S1_7983920	SNP	TCCCGATGAGATTTTTATGC	TAGTACTGTGCTTGTGTGCC
s1_8042406	SNP	GGATCTCGATGAGATTCACC	CACAGTAAGAAGGCAATCTG
s1_7786880	SNP	GCCACTCAATTTCAGACAATG	GCACTTATGGCAAACATAAC
S1-8060SCAR	SCAR	GCAGTtGCCTAAATTGCAAT	CACTGACAAACCAAGGTGCG
PGS1.21	SSR	CCCTGGTGTTCTGCTCTCTC	CATCCACAAATGGGAAGCAT
S1-8109SSR	SSR	CCTCCCTTCCTCCATCGTTG	CAGGCTGCAACTCAAATCCC
s1_8132703	SNP	GGGCGCATTGTTTTATACAC	GTAACAGTCAAGGACAGCTC
ZP002	SSLP	AACATTTTCTGATTCAATGCCA	TGTATCCTCCAGCTTCAAAGTC
MA067	SSR	AAGGAGAGGAAAGAAAGGGAGA	CACTTCACCCTCACTTCACTCA
PGS1-24	SSR	GTAAATGAGTGCCTGCGTGT	TGCGAGAGTTGTGATTGATG

*Gol61, PGS markers, MA067 and ZP002 were obtained from literature (Soriano *et al*., 2012; Zuriaga *et al*., 2013; Decrooq *et al*., 2014).

Tab. 3 – Primer sequences developed on the BAC ends.

BAC clone	FORWARD PRIMER	REVERSE PRIMER
11P3_SR	AGTAGACGTGAAGCATCCACT	TCATCAATAAACTCAGCACAGC
16K16_SR	GTCTTTGACGTGCTTCACCATA	TTTTCGTAAAGCCAATACAAACA
28P4_SL	AATGGGGTTTGGGAATGTTTGA	CCCGTGGAAGGTTGCAATTG
36C4_SR	CCAATAGCTGTCATCGCGAG	AGACACATCTTGATTTTCCGGT
36E17_SR	GGTTTTCTGCCGACCCAATT	CCCCTAAACTGTACTTAAGCTGC
37M10_SR	GCGCTTGTCAAGGGTGCTTGTTCATTCACAGC	GACGTTTTGTGATAATAAGCCCT
40I18_SL	TTGCTCAACTCCTTGGCCTA	GATCTATGGTGTGGGCATGC
40P21_SL	CAACACGCGGGCTTCTAAG	CTTAATACTCATGCGGTTGGGC
41I23_SL	ACAAACTGGCATAACAATTTCCA	GGAGTGGCATTAGATGGGGA
41I23_SR	TAGAACTGGCGATGCGTTTG	TGAGAAGGATGGATTGGCTGT
45G23_SR	TGTAGACGGCGAAGGGTTTT	CAAAAGCCCATCGGACAGAC
47M3_SR	GCGCTTGTCAAATTCACAGC	TTCAAAGCTTGCATGGCCAG
54E7_SL	GCTTAAGACTGGGCGCATTT	GGCTCAAGTGGGATTAGGGT
54E7_SR	TTTTGTGTGGGTGCTGCTG	CATACACTGCAAGGCGAGTC
55E16_SR	TCAACTTCGCCTTCCTCCTT	CTTTCGCAGTTTCACCAGCT
55P1_SL	CACCACCATCACCCACATTG	TGATGGTTTGTGGAGGAGGT
55P1_SR	GGACAATCCCCTGCTTAACC	TGCCCTTTGGTTTGTGCAAT
59D2_SR	TAAGCCAATCTCCAGGACGG	CCTCTCCTCCAGAACTACCG
63O4_SL	TGGGAAGATGAGATGAGCTTGT	GAGGGAGCAAAGTACACGGA
66N22_SR	TCCAGGACACAGAAACCACT	AGGCAATCGTAGCTGTCCTT
6E20_SR	CGGGACTTATGGTATGTTCACA	ATAGTCTAAGGGTGGCCACG
6F3_SL	GATGACCATTAATTTACCCCGTAA	AGCCACCTCAACAAGACTGA
6F3_SR	CCACAGCATTTCCACAGCAA	CAGGTCAAGCAGAGGAAAGC
70N14_SR	GCGTGAGGTAAGGCTAGTGT	TGCCATGAAACTTGAAGCACA
74P8_SR	TTGCCGATTCCAACAAGTCC	GGAAGCATGCATCAACAGCT

Tab. 4 - List of annotated genes for chrR_1 and chrS_1 with their start/end positions. Genes shared by both resistant and susceptible haplotypes are reported in the same line. In red are reported the molecular markers that saturate the region.

start	end	**ChrS_1**	**ChrR_1**	start	end
2041	7251	Similar to fggy: FGGY carbohydrate kinase domain-containing protein (Xenopus laevis)			
11055	11489	Protein of unknown function			
11715	17612	Similar to DDB_G0272254: Probable serine/threonine-protein kinase DDB_G0272254 (Dictyostelium discoideum)			
28644	30139	Similar to MYB36: Transcription factor MYB36 (Arabidopsis thaliana)			
32792	33177	Protein of unknown function			
36464	36985	Protein of unknown function			
46810	47426	Protein of unknown function			
53455	61308	Similar to DTX16: Protein DETOXIFICATION 16 (Arabidopsis thaliana)			
69517	69801	Protein of unknown function			
72663	82024	Similar to DTX16: Protein DETOXIFICATION 16 (Arabidopsis thaliana)	Similar to DTX16: Protein DETOXIFICATION 16 (Arabidopsis thaliana)	2732	5708
76250	76344	**S1_6345556**		????	????
84577	104604	Similar to UGT78G1: Flavonoid 3-O-glucosyltransferase (Medicago truncatula)	Similar to UGT78G1: Flavonoid 3-O-glucosyltransferase (Medicago truncatula)	8247	14661
112214	112855	Similar to UGT78D2: UDP-glycosyltransferase 78D2 (Arabidopsis thaliana)	Similar to FGT: Anthocyanidin 3-O-glucosyltransferase 2 (Fragaria ananassa)	28263	37381
114929	115092	**S1_6422355**		39870	40033
115053	115562	Similar to PR-1: Pathogenesis-related protein PR-1 (Medicago truncatula)	Similar to PR-1: Pathogenesis-related protein PR-1 (Medicago truncatula)	39994	40503
117024	117605	Similar to PR-1: Pathogenesis-related protein PR-1 (Medicago truncatula)	Similar to PR-1: Pathogenesis-related protein PR-1 (Medicago truncatula)	41961	42542

start	end	ChrS_1	ChrR_1	start	end
			Protein of unknown function	44878	45292
127027	132668	Similar to ROG1: Putative lipase ROG1 (Saccharomyces cerevisiae (strain ATCC 204508 / S288c))	Similar to ROG1: Putative lipase ROG1 (Saccharomyces cerevisiae (strain ATCC 204508 / S288c))	50902	56501
134298	137400	Similar to CIPK21: CBL-interacting serine/threonine-protein kinase 21 (Arabidopsis thaliana)	Similar to CIPK21: CBL-interacting serine/threonine-protein kinase 21 (Arabidopsis thaliana)	58126	61430
158072	164446	Similar to XYLA: Xylose isomerase (Arabidopsis thaliana)	Similar to XYLA: Xylose isomerase (Arabidopsis thaliana)	85342	91739
166307	166588	Similar to CYP71B20: Cytochrome P450 71B20 (Arabidopsis thaliana)	Similar to CYP71B20: Cytochrome P450 71B20 (Arabidopsis thaliana)	96913	97194
168365	174385	Similar to ARF1: ADP-ribosylation factor 1 (Brassica rapa subsp. pekinensis)	Similar to ARF1: ADP-ribosylation factor 1 (Brassica rapa subsp. pekinensis)	98972	104591
177938	187277	Similar to MDL3: (R)-mandelonitrile lyase 3 (Prunus serotina)	Similar to MDL3: (R)-mandelonitrile lyase 3 (Prunus serotina)	107561	117192
180008	180181	**S1_6537106**		109873	110046
200994	203052	Similar to MDL2: (R)-mandelonitrile lyase 2 (Prunus serotina)	Similar to MDL2: (R)-mandelonitrile lyase 2 (Prunus serotina)	127992	147377
221280	223402	Similar to MDL2: (R)-mandelonitrile lyase 2 (Prunus serotina)			
233906	236018	Similar to MDL2: (R)-mandelonitrile lyase 2 (Prunus serotina)			
246578	248744	Similar to MDL4: (R)-mandelonitrile lyase 4 (Prunus serotina)	Similar to MDL4: (R)-mandelonitrile lyase 4 (Prunus serotina)	158582	160749
256836	259003	Similar to MDL1: (R)-mandelonitrile lyase 1 (Prunus dulcis)	Similar to MDL1: (R)-mandelonitrile lyase 1 (Prunus dulcis)	167628	169793
278041	307037	Similar to MDL1: (R)-mandelonitrile lyase 1 (Prunus dulcis)	Similar to MDL1: (R)-mandelonitrile lyase 1 (Prunus dulcis)	187250	192692
311518	312748	Protein of unknown function	Similar to MDL2: (R)-mandelonitrile lyase 2 (Prunus dulcis)	202591	205212
314392	315858	Similar to SLSG: S-locus-specific glycoprotein S6 (Brassica oleracea)	Protein of unknown function	206204	207062
315955	316302	Protein of unknown function	Similar to TPR1: Topless-related protein 1 (Arabidopsis thaliana)	207441	207786
316402	316890	Protein of unknown function	Protein of unknown function	208563	209842
317547	318725	Similar to TPR1: Topless-related protein 1 (Arabidopsis thaliana)	Similar to TPR1: Topless-related protein 1 (Arabidopsis thaliana)	210298	210628

start	end	**ChrS_1**	**ChrR_1**	start	end
318791	319883	Protein of unknown function	Similar to TPR1: Topless-related protein 1 (Arabidopsis thaliana)	211883	212308
320331	320670	Similar to TPR1: Topless-related protein 1 (Arabidopsis thaliana)	Protein of unknown function	212616	212910
			Protein of unknown function	213272	214571
			Protein of unknown function	218875	219818
			Similar to At4g27290: G-type lectin S-receptor-like serine/threonine-protein kinase At4g27290 (Arabidopsis thaliana)	223980	226644
			Protein of unknown function	226886	227200
			Protein of unknown function	227303	228115
			Similar to TPR1: Topless-related protein 1 (Arabidopsis thaliana)	228490	228842
			Similar to TPR1: Topless-related protein 1 (Arabidopsis thaliana)	231308	231630
			Similar to TPR1: Topless-related protein 1 (Arabidopsis thaliana)	232642	233067
			Protein of unknown function	233375	233625
			Protein of unknown function	234032	235331
			Protein of unknown function	239280	240223
			Protein of unknown function	243093	244161
			Similar to At4g27290: G-type lectin S-receptor-like serine/threonine-protein kinase At4g27290 (Arabidopsis thaliana)	244668	247904
			Protein of unknown function	248007	248257
			Similar to TPR1: Topless-related protein 1 (Arabidopsis thaliana)	249199	249551
			Protein of unknown function	251344	251562
			Protein of unknown function	258497	259050
			Protein of unknown function	264234	264515

Tab. 5 – List of annotated genes for chrR_2 and chrS_1 with their start/end positions. Genes shared by both resistant and susceptible haplotypes are reported in the same line. In red are reported the molecular markers that saturate the region.

start	end	ChrS_2	ChrR_2	start	end
57	1240	Similar to MDL3: (R)-mandelonitrile lyase 3 (Prunus serotina)			
18994	21129	Similar to MDL2: (R)-mandelonitrile lyase 2 (Prunus dulcis)	Similar to MDL2: (R)-mandelonitrile lyase 2 (Prunus dulcis)	11610	19904
19052	19203	S1_6698823		11668	11819
33472	34578	Similar to PME63: Putative pectinesterase 63 (Arabidopsis thaliana)	Similar to PPME1: Pectinesterase PPME1 (Arabidopsis thaliana)	25158	26231
43228	62942	Similar to MDL1: (R)-mandelonitrile lyase 1 (Prunus serotina)	Similar to MDL2: (R)-mandelonitrile lyase 2 (Prunus dulcis)	31150	33700
			Similar to MDL2: (R)-mandelonitrile lyase 2 (Prunus dulcis)	51609	60154
57391	57570	**S1_6761324**		55737	55916
68552	69541	Similar to At3g03770: Probable inactive leucine-rich repeat receptor-like protein kinase At3g03770 (Arabidopsis thaliana)			
73217	75441	Similar to AGL16: Agamous-like MADS-box protein AGL16 (Arabidopsis thaliana)	Protein of unknown function	68018	76794
97338	98627	Similar to MDL2: (R)-mandelonitrile lyase 2 (Prunus serotina)	Similar to MDL2: (R)-mandelonitrile lyase 2 (Prunus serotina)	88347	89572
107373	107788	Protein of unknown function	Protein of unknown function	104648	104872
108449	109683	Similar to COL5: Zinc finger protein CONSTANS-LIKE 5 (Arabidopsis thaliana)	Similar to COL5: Zinc finger protein CONSTANS-LIKE 5 (Arabidopsis thaliana)	105771	107197
113239	113475	Similar to At2g29880: Uncharacterized protein At2g29880 (Arabidopsis thaliana)	Similar to At2g29880: Uncharacterized protein At2g29880 (Arabidopsis thaliana)	110523	110783
115536	115774	S1-6798SCAR		112810	113048
124227	131245	Protein of unknown function	Protein of unknown function	121556	137180
131926	132030	**S1-6835SCAR**		127964	128127

start	end	ChrS_2	ChrR_2	start	end
147271	147732	Protein of unknown function			
148152	155575	Similar to At5g57670: Probable receptor-like serine/threonine-protein kinase At5g57670 (Arabidopsis thaliana)	Similar to At5g57670: Probable receptor-like serine/threonine-protein kinase At5g57670 (Arabidopsis thaliana)	137600	145024
163569	164111	Similar to GDU3: Protein GLUTAMINE DUMPER 3 (Arabidopsis thaliana)	Similar to GDU3: Protein GLUTAMINE DUMPER 3 (Arabidopsis thaliana)	153477	154019
164096	164241	**PGS1,03**		154004	154169
175313	177679	Similar to SNRPE: Small nuclear ribonucleoprotein E (Sus scrofa)	Similar to SNRPE: Small nuclear ribonucleoprotein E (Sus scrofa)	165257	167627
180574	180979	Protein of unknown function	Protein of unknown function	170395	170800
183485	221020	Similar to GLR2.7: Glutamate receptor 2.7 (Arabidopsis thaliana)	Similar to GLR2.8: Glutamate receptor 2.8 (Arabidopsis thaliana)	172550	208617
231561	233486	Similar to GLR2.9: Glutamate receptor 2.9 (Arabidopsis thaliana)	Similar to GLR2.7: Glutamate receptor 2.7 (Arabidopsis thaliana)	239799	242198
233504	235335	Similar to GLR2.1: Glutamate receptor 2.1 (Arabidopsis thaliana)			
244605	247245	Protein of unknown function			
247306	248082	Protein of unknown function			
251650	251792	**s1_6965213**		241849	241991
253141	261465	Similar to DGK4: Diacylglycerol kinase 4 (Arabidopsis thaliana)	Similar to DGK7: Diacylglycerol kinase 7 (Arabidopsis thaliana)	243161	251956
269980	277052	Similar to Trmo: tRNA (adenine(37)-N6)-methyltransferase (Mus musculus)	Similar to Trmo: tRNA (adenine(37)-N6)-methyltransferase (Mus musculus)	261718	268489
272508	272687	**S1_6994SCAR**		263903	264082
290837	299132	Similar to SCY1: Preprotein translocase subunit SCY1%2C chloroplastic (Arabidopsis thaliana)	Similar to SCY1: Preprotein translocase subunit SCY1%2C chloroplastic (Arabidopsis thaliana)	282724	291017
322531	322593	**S1-7045SSR**		314040	314102
323386	328135	Similar to SMAX1: Protein SUPPRESSOR OF MAX2 1 (Arabidopsis thaliana)	Similar to SMAX1: Protein SUPPRESSOR OF MAX2 1 (Arabidopsis thaliana)	314897	319817
358574	362459	Protein of unknown function	Protein of unknown function	350101	353959
361845	361993	**S1_7077554**		353372	353520

start	end	ChrS_2	ChrR_2	start	end
366288	366476	Protein of unknown function	Protein of unknown function	357788	357976
366565	373926	Similar to CNGC17: Cyclic nucleotide-gated ion channel 17 (Arabidopsis thaliana)	Similar to CNGC17: Cyclic nucleotide-gated ion channel 17 (Arabidopsis thaliana)	358065	364551
379956	382940	Similar to TPS9: Probable alpha%2Calpha-trehalose-phosphate synthase [UDP-forming] 9 (Arabidopsis thaliana)	Similar to TPS9: Probable alpha%2Calpha-trehalose-phosphate synthase [UDP-forming] 9 (Arabidopsis thaliana)	371468	374452
390356	390542	**S1_7112235**		381779	381965
393009	396201	Protein of unknown function	Protein of unknown function	384436	387612
397210	398073	Protein of unknown function	Protein of unknown function	388623	389345
399511	402388	Protein of unknown function	Protein of unknown function	390933	391872
408832	409770	Protein of unknown function	Protein of unknown function	398409	399365
			Protein of unknown function	405707	406645
417248	424870	Similar to ATL56: RING-H2 finger protein ATL56 (Arabidopsis thaliana)	Similar to ATL56: RING-H2 finger protein ATL56 (Arabidopsis thaliana)	414024	414710
427265	427758	Protein of unknown function	Protein of unknown function	414922	424490
450414	451715	Protein of unknown function	Protein of unknown function	424849	426655
452037	453185	Protein of unknown function	Protein of unknown function	426682	427471
434030	434136	**S1-7164SSR**		429607	429713
			Protein of unknown function	439975	440720
			Protein of unknown function	443395	444325
			Protein of unknown function	444373	445519
468164	479322	Similar to MDL3: (R)-mandelonitrile lyase 3 (Prunus serotina)	Similar to MDL3: (R)-mandelonitrile lyase 3 (Prunus serotina)	458980	470572
471982	472138	**s1_7217828**		462813	462969
472200	472307	**S1-7218Scar**		463053	463160
493166	493927	Similar to Uncharacterized protein RJ39 (Fragment) (Fragaria ananassa)	Similar to Uncharacterized protein RJ39 (Fragment) (Fragaria ananassa)	481828	482589
497909	498236	Protein of unknown function			

start	end	ChrS_2	ChrR_2	start	end
499742	501569	Similar to RPL30: 60S ribosomal protein L30 (Lupinus luteus)	Similar to RPL30: 60S ribosomal protein L30 (Lupinus luteus)	488547	490374
512148	512333	**s1_7241764**		496130	496315
513198	517220	Similar to CjBAp12: EG45-like domain containing protein (Citrus jambhiri)	Similar to CjBAp12: EG45-like domain containing protein (Citrus jambhiri)	497102	501160
521623	523107	Similar to ATL13: RING-H2 finger protein ATL13 (Arabidopsis thaliana)	Similar to ATL13: RING-H2 finger protein ATL13 (Arabidopsis thaliana)	505588	507072
537069	537327	**S1-7284SSR**		521053	521306
536976	542966	Similar to XBOS32: Probable E3 ubiquitin-protein ligase XBOS32 (Oryza sativa subsp. japonica)	Similar to XBOS32: Probable E3 ubiquitin-protein ligase XBOS32 (Oryza sativa subsp. japonica)	522014	526961
551336	553888	Protein of unknown function	Protein of unknown function	535660	538212
559829	563439	Similar to PIGB: GPI mannosyltransferase 3 (Bos taurus)	Similar to Pigb: GPI mannosyltransferase 3 (Mus musculus)	544159	547769
565200	567849	Protein of unknown function	Protein of unknown function	549527	552180
577690	578235	Similar to IBL1: Transcription factor IBH1-like 1 (Arabidopsis thaliana)	Similar to IBL1: Transcription factor IBH1-like 1 (Arabidopsis thaliana)	561995	562540
588046	588297	Protein of unknown function	Protein of unknown function	572242	572493
590404	590527	**S1-7361SSR**		574579	574698
592532	599069	Protein of unknown function	Protein of unknown function	576898	582935
602164	607035	Similar to CER3: Protein ECERIFERUM 3 (Arabidopsis thaliana)	Similar to CER3: Protein ECERIFERUM 3 (Arabidopsis thaliana)	585570	589889
			Protein of unknown function	591821	591964
614453	615574	Similar to At4g30420: WAT1-related protein At4g30420 (Arabidopsis thaliana)	Similar to At4g30420: WAT1-related protein At4g30420 (Arabidopsis thaliana)	600505	601626
616035	616327	Protein of unknown function	Protein of unknown function	602088	602413
616361	617395	Protein of unknown function	Protein of unknown function	602414	603487
620470	621273	Protein of unknown function	Protein of unknown function	606525	607355
621349	622095	Protein of unknown function	Protein of unknown function	607431	608178
622140	622728	Similar to At4g28040: WAT1-related protein At4g28040 (Arabidopsis thaliana)	Similar to At4g28040: WAT1-related protein At4g28040 (Arabidopsis thaliana)	608223	608791

start	end	ChrS_2	ChrR_2	start	end
622503	622668	**PGS1.10**		608490	608731
623960	624472	Protein of unknown function	Protein of unknown function	610374	611273
624646	625223	Protein of unknown function	Protein of unknown function	611313	611948
625263	625898	Protein of unknown function			
626373	627398	Similar to At5g10770: Aspartyl protease family protein At5g10770 (Arabidopsis thaliana)	Similar to At5g10770: Aspartyl protease family protein At5g10770 (Arabidopsis thaliana)	612423	613448
636020	638505	Similar to TET8: Tetraspanin-8 (Arabidopsis thaliana)	Similar to TET8: Tetraspanin-8 (Arabidopsis thaliana)	622325	624796
640326	644587	Similar to TET8: Tetraspanin-8 (Arabidopsis thaliana)	Similar to TET8: Tetraspanin-8 (Arabidopsis thaliana)	626593	630924
645341	645539	**S1-7418SSR**		631678	631876
649161	651986	Similar to COX6B-2: Cytochrome c oxidase subunit 6b-2 (Arabidopsis thaliana)	Similar to COX6B-2: Cytochrome c oxidase subunit 6b-2 (Arabidopsis thaliana)	646026	648688
653333	656370	Similar to FLXL2: Protein FLX-like 2 (Arabidopsis thaliana)	Similar to FLXL2: Protein FLX-like 2 (Arabidopsis thaliana)	651828	653436
659746	661141	Protein of unknown function			
661233	661377	**s1_7442314**		662361	662505
671092	672390	Similar to GAE1: UDP-glucuronate 4-epimerase 1 (Arabidopsis thaliana)	Similar to GAE1: UDP-glucuronate 4-epimerase 1 (Arabidopsis thaliana)	672156	673454
676656	683498	Similar to CDC20-1: Cell division cycle 20.1%2C cofactor of APC complex (Arabidopsis thaliana)	Similar to CDC20-1: Cell division cycle 20.1%2C cofactor of APC complex (Arabidopsis thaliana)	677721	684967
685826	691489	Similar to DDB_G0275467: 5'-nucleotidase domain-containing protein DDB_G0275467 (Dictyostelium discoideum)	Similar to DDB_G0275467: 5'-nucleotidase domain-containing protein DDB_G0275467 (Dictyostelium discoideum)	688185	693864
685811	685969	**S1-7484SSR**		688170	688334
693052	693476	Protein of unknown function	Protein of unknown function	695406	695856
696614	698520	Similar to SNL6: Cinnamoyl-CoA reductase-like SNL6 (Oryza sativa subsp. japonica)	Similar to SNL6: Cinnamoyl-CoA reductase-like SNL6 (Oryza sativa subsp. japonica)	698751	700598
710772	712330	Protein of unknown function	Protein of unknown function	713862	714487
718468	723646	Similar to TTC1: Tetratricopeptide repeat protein 1 (Homo sapiens)	Similar to TTC1: Tetratricopeptide repeat protein 1 (Homo sapiens)	718628	722118

start	end	ChrS_2	ChrR_2	start	end
724884	730608	Similar to RGA3: Putative disease resistance protein RGA3 (Solanum bulbocastanum)	Protein of unknown function	725639	727510
729197	735948	Similar to At5g35200: Putative clathrin assembly protein At5g35200 (Arabidopsis thaliana)	Similar to At5g35200: Putative clathrin assembly protein At5g35200 (Arabidopsis thaliana)	727627	732285
733329	733687	S1-7518SCAR		730264	730612
737533	743754	Similar to AFG1L: AFG1-like ATPase (Homo sapiens)	Similar to AFG1L: AFG1-like ATPase (Homo sapiens)	734420	740808
748414	751096	Similar to AE7: Protein AE7 (Arabidopsis thaliana)	Similar to AE7: Protein AE7 (Arabidopsis thaliana)	744934	747618
752405	754518	Similar to RPL6: 50S ribosomal protein L6%2C chloroplastic (Arabidopsis thaliana)	Similar to RPL6: 50S ribosomal protein L6%2C chloroplastic (Arabidopsis thaliana)	748932	751052
756665	758553	Similar to tmem208: Transmembrane protein 208 (Danio rerio)	Similar to tmem208: Transmembrane protein 208 (Danio rerio)	753202	755090
758929	759084	**s1_7555803**		755466	755621
759478	760215	Similar to FLA20: Putative fasciclin-like arabinogalactan protein 20 (Arabidopsis thaliana)	Similar to FLA20: Putative fasciclin-like arabinogalactan protein 20 (Arabidopsis thaliana)	756017	756754
760858	761451	Similar to FLA20: Putative fasciclin-like arabinogalactan protein 20 (Arabidopsis thaliana)	Similar to FLA20: Putative fasciclin-like arabinogalactan protein 20 (Arabidopsis thaliana)	757406	757996
766371	771358	Similar to ATG18B: Autophagy-related protein 18b (Arabidopsis thaliana)	Similar to ATG18B: Autophagy-related protein 18b (Arabidopsis thaliana)	762599	767942
773594	777934	Similar to At2g23950: Probable LRR receptor-like serine/threonine-protein kinase At2g23950 (Arabidopsis thaliana)	Similar to At2g23950: Probable LRR receptor-like serine/threonine-protein kinase At2g23950 (Arabidopsis thaliana)	770064	774387
784058	784230	**s1_7579835**		780500	780672
789308	790481	Similar to GER2: Putative GDP-L-fucose synthase 2 (Arabidopsis thaliana)	Similar to GER2: Putative GDP-L-fucose synthase 2 (Arabidopsis thaliana)	785750	786921
801807	804479	Similar to MARC1: Mitochondrial amidoxime-reducing component 1 (Homo sapiens)	Similar to MARC1: Mitochondrial amidoxime-reducing component 1 (Homo sapiens)	794641	800917
805776	808481	Similar to PCMP-H81: Pentatricopeptide repeat-containing protein At3g57430%2C chloroplastic (Arabidopsis thaliana)	Similar to PCMP-H81: Pentatricopeptide repeat-containing protein At3g57430%2C chloroplastic (Arabidopsis thaliana)	802213	804918
809197	810001	Similar to RSM27: Mitochondrial 37S ribosomal protein S27 (Saccharomyces cerevisiae (strain ATCC 204508 / S288c))	Similar to RSM27: Mitochondrial 37S ribosomal protein S27 (Saccharomyces cerevisiae (strain ATCC 204508 / S288c))	805634	806438

start	end	ChrS_2	ChrR_2	start	end
810509	811067	Protein of unknown function	Protein of unknown function	806946	807503
822546	823445	Similar to GGP5: Gamma-glutamyl peptidase 5 (Arabidopsis thaliana)	Similar to GGP5: Gamma-glutamyl peptidase 5 (Arabidopsis thaliana)	818994	819893
825153	831699	Similar to ASHR3: Histone-lysine N-methyltransferase ASHR3 (Arabidopsis thaliana)	Similar to ASHR3: Histone-lysine N-methyltransferase ASHR3 (Arabidopsis thaliana)	821601	828305
838748	840863	Similar to At1g16060: AP2-like ethylene-responsive transcription factor At1g16060 (Arabidopsis thaliana)	Similar to At1g16060: AP2-like ethylene-responsive transcription factor At1g16060 (Arabidopsis thaliana)	835354	837469
845772	847512	Similar to HHP2: Heptahelical transmembrane protein 2 (Arabidopsis thaliana)	Similar to HHP2: Heptahelical transmembrane protein 2 (Arabidopsis thaliana)	842377	844117
848361	859890	Similar to TSS: Protein TSS (Arabidopsis thaliana)	Similar to TSS: Protein TSS (Arabidopsis thaliana)	844966	856478
866640	867138	Protein of unknown function	Protein of unknown function	863227	863724
867329	870456	Similar to NUP43: Nuclear pore complex protein NUP43 (Arabidopsis thaliana)	Similar to NUP43: Nuclear pore complex protein NUP43 (Arabidopsis thaliana)	863915	867041
871293	873337	Similar to RPL35: 50S ribosomal protein L35%2C chloroplastic (Arabidopsis thaliana)	Similar to RPL35: 50S ribosomal protein L35%2C chloroplastic (Arabidopsis thaliana)	867878	869922
874699	874779	**S1-7700SSR**		871283	871363
877669	881362	Protein of unknown function	Protein of unknown function	874259	877952
885756	894727	Similar to ATG11: Autophagy-related protein 11 (Arabidopsis thaliana)	Similar to ATG11: Autophagy-related protein 11 (Arabidopsis thaliana)	882345	891315
901969	911275	Similar to CASTOR: Ion channel CASTOR (Lotus japonicus)	Similar to CASTOR: Ion channel CASTOR (Lotus japonicus)	898557	907863
909116	909316	**S1-7745SSR**		905704	905904
921460	921866	Protein of unknown function	Protein of unknown function	918093	918499
934841	944506	Similar to Cysteine synthase%2C chloroplastic/chromoplastic (Solanum tuberosum)	Similar to Cysteine synthase%2C chloroplastic/chromoplastic (Solanum tuberosum)	931473	941138
952236	953567	Similar to SDR1: (+)-neomenthol dehydrogenase (Arabidopsis thaliana)	Similar to SDR1: (+)-neomenthol dehydrogenase (Arabidopsis thaliana)	948868	950199
955311	958446	Similar to Os05g0567100: Aspartic proteinase oryzasin-1 (Oryza sativa subsp. japonica)	Similar to Os05g0567100: Aspartic proteinase oryzasin-1 (Oryza sativa subsp. japonica)	951943	955078
962837	968207	Similar to At3g62120: Proline--tRNA ligase%2C cytoplasmic (Arabidopsis thaliana)	Similar to At3g62120: Proline--tRNA ligase%2C cytoplasmic (Arabidopsis thaliana)	959469	964839

start	end	ChrS_2	ChrR_2	start	end
969230	970722	Similar to DGAT3: Diacylglycerol O-acyltransferase 3%2C cytosolic (Arabidopsis thaliana)	Similar to DGAT3: Diacylglycerol O-acyltransferase 3%2C cytosolic (Arabidopsis thaliana)	965862	967354
972204	975757	Similar to CHUP1: Protein CHUP1%2C chloroplastic (Arabidopsis thaliana)	Similar to CHUP1: Protein CHUP1%2C chloroplastic (Arabidopsis thaliana)	968836	972389
980644	1005077	Similar to ABCB11: ABC transporter B family member 11 (Arabidopsis thaliana)	Similar to ABCB11: ABC transporter B family member 11 (Arabidopsis thaliana)	977276	1001709
1012186	1013103	Similar to nt5c2: Cytosolic purine 5'-nucleotidase (Dictyostelium discoideum)	Similar to nt5c2: Cytosolic purine 5'-nucleotidase (Dictyostelium discoideum)	1008818	1009735
1021857	1027342	Similar to ABCB4: ABC transporter B family member 4 (Arabidopsis thaliana)	Similar to ABCB4: ABC transporter B family member 4 (Arabidopsis thaliana)	1018489	1023974
1039014	1044547	Similar to ABCB11: ABC transporter B family member 11 (Arabidopsis thaliana)	Similar to ABCB11: ABC transporter B family member 11 (Arabidopsis thaliana)	1035646	1041179
1068994	1069978	Protein of unknown function	Protein of unknown function	1065626	1066610
1073259	1089039	Similar to ABCB4: ABC transporter B family member 4 (Arabidopsis thaliana)	Similar to ABCB4: ABC transporter B family member 4 (Arabidopsis thaliana)	1069891	1085671
1096824	1098042	Protein of unknown function	Protein of unknown function	1093456	1094674
1103801	1109249	Similar to ABCB4: ABC transporter B family member 4 (Arabidopsis thaliana)	Similar to ABCB4: ABC transporter B family member 4 (Arabidopsis thaliana)	1100431	1105879
1118182	1119183	Protein of unknown function	Protein of unknown function	1114812	1115813
1124342	1126708	Similar to TGD4: Protein TRIGALACTOSYLDIACYLGLYCEROL 4%2C chloroplastic (Arabidopsis thaliana)	Similar to TGD4: Protein TRIGALACTOSYLDIACYLGLYCEROL 4%2C chloroplastic (Arabidopsis thaliana)	1120947	1123247
1126842	1129113	Similar to truA1: tRNA pseudouridine synthase A 1 (Protochlamydia amoebophila (strain UWE25))	Similar to truA1: tRNA pseudouridine synthase A 1 (Protochlamydia amoebophila (strain UWE25))	1123453	1125723
1130040	1130248	**S1-7982SSR**		1126650	1126858
1130779	1133343	Protein of unknown function	Protein of unknown function	1127389	1129953
1131299	1131479	**S1_7983920**		1127909	1128089
1135079	1136631	Similar to Late embryogenesis abundant protein D-29 (Gossypium hirsutum)	Similar to Late embryogenesis abundant protein D-29 (Gossypium hirsutum)	1131680	1133241
1146188	1150609	Similar to CIPK1: CBL-interacting serine/threonine-protein kinase 1 (Arabidopsis thaliana)	Similar to CIPK1: CBL-interacting serine/threonine-protein kinase 1 (Arabidopsis thaliana)	1141229	1145650
			Protein of unknown function	1145953	1146486

start	end	ChrS_2	ChrR_2	start	end
1149067	1154372	Similar to CRSH: Probable GTP diphosphokinase CRSH%2C chloroplastic (Arabidopsis thaliana)	Similar to CRSH: Probable GTP diphosphokinase CRSH%2C chloroplastic (Arabidopsis thaliana)	1146968	1149422
1154983	1159483	Similar to RPL3B: 50S ribosomal protein L3-2%2C chloroplastic (Arabidopsis thaliana)	Similar to RPL3B: 50S ribosomal protein L3-2%2C chloroplastic (Arabidopsis thaliana)	1150023	1154523
1161539	1163509	Similar to PME28: Putative pectinesterase/pectinesterase inhibitor 28 (Arabidopsis thaliana)	Similar to PME28: Putative pectinesterase/pectinesterase inhibitor 28 (Arabidopsis thaliana)	1156572	1158542
1164638	1174692	Similar to PCMP-H61: Pentatricopeptide repeat-containing protein At5g66520 (Arabidopsis thaliana)	Similar to PCMP-H61: Pentatricopeptide repeat-containing protein At5g66520 (Arabidopsis thaliana)	1159682	1170025
1175967	1178974	Similar to PDK: [Pyruvate dehydrogenase (acetyl-transferring)] kinase%2C mitochondrial (Arabidopsis thaliana)	Similar to PDK: [Pyruvate dehydrogenase (acetyl-transferring)] kinase%2C mitochondrial (Arabidopsis thaliana)	1170500	1173508
1177839	1177998	**s1_8042406**		1172373	1172532
1184140	1188661	Similar to NPSN13: Novel plant SNARE 13 (Arabidopsis thaliana)	Similar to NPSN13: Novel plant SNARE 13 (Arabidopsis thaliana)	1178660	1183193
1189319	1191028	Similar to Slx1b: Structure-specific endonuclease subunit SLX1 (Mus musculus)	Similar to Slx1b: Structure-specific endonuclease subunit SLX1 (Mus musculus)	1183859	1185748
1191566	1192624	Similar to NAC071: NAC domain-containing protein 71 (Arabidopsis thaliana)	Similar to NAC071: NAC domain-containing protein 71 (Arabidopsis thaliana)	1186082	1187140
1194146	1207102	Similar to ZRANB3: DNA annealing helicase and endonuclease ZRANB3 (Bos taurus)			
			Similar to smarcal1: SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily A-like protein 1 (Danio rerio)	1188993	1201785
1205603	1205479	**s1_7786880**		1200209	1200333
1208240	1209689	Protein of unknown function	Protein of unknown function	1202864	1204313
1210055	1211113	Protein of unknown function	Protein of unknown function	1204679	1205737
1212737	1216222	Similar to At3g17430: Probable sugar phosphate/phosphate translocator At3g17430 (Arabidopsis thaliana)	Similar to At3g17430: Probable sugar phosphate/phosphate translocator At3g17430 (Arabidopsis thaliana)	1207798	1211265
1213891	1214174	**S1-8060SCAR**		1208952	1209235
1226587	1226687	**PGS1.21**		1220669	1220761

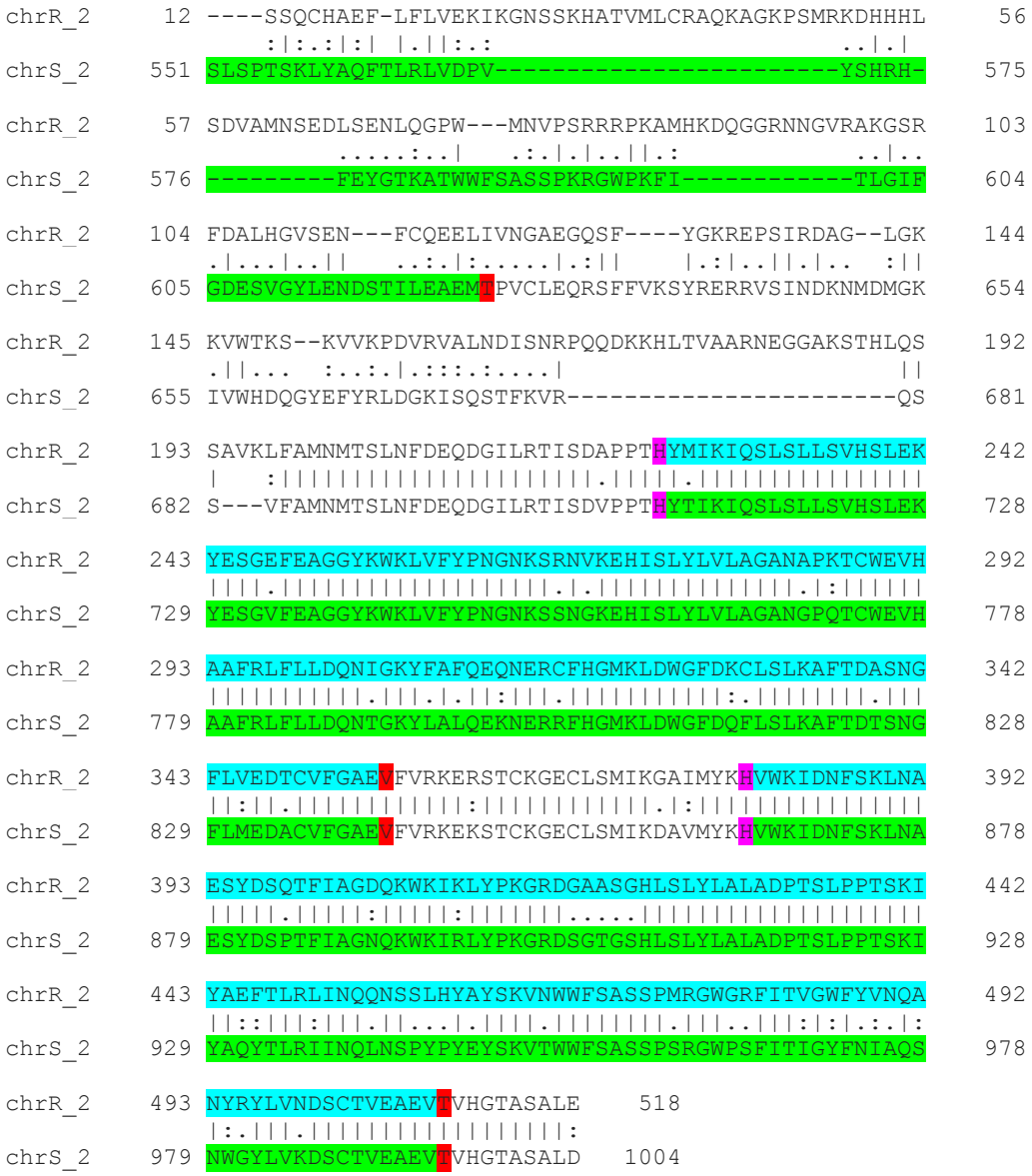
start	end	ChrS_2	ChrR_2	start	end
1226923	1229393	Similar to At5g18500: Probable receptor-like protein kinase At5g18500 (Arabidopsis thaliana)	Similar to At5g18500: Probable receptor-like protein kinase At5g18500 (Arabidopsis thaliana)	1220997	1223467
1230285	1234174	Similar to PTI1: Pto-interacting protein 1 (Solanum lycopersicum)	Similar to PTI1: Pto-interacting protein 1 (Solanum lycopersicum)	1226262	1227132
1240912	1254719	Similar to SPL1: Squamosa promoter-binding-like protein 1 (Arabidopsis thaliana)	Similar to SPL1: Squamosa promoter-binding-like protein 1 (Arabidopsis thaliana)	1234864	1248671
1261425	1263177	Protein of unknown function	Protein of unknown function	1255377	1257129
1261608	1261884	**S1-8109SSR**		1255560	1255836
1274528	1276849	Similar to EMB2750: Pentatricopeptide repeat-containing protein At3g06430%2C chloroplastic (Arabidopsis thaliana)	Similar to EMB2750: Pentatricopeptide repeat-containing protein At3g06430%2C chloroplastic (Arabidopsis thaliana)	1261520	1263356
1278010	1285801	Similar to PDR3: Pleiotropic drug resistance protein 3 (Nicotiana tabacum)	Similar to PDR3: Pleiotropic drug resistance protein 3 (Nicotiana tabacum)	1264938	1272745
1287526	1294482	Similar to PDR3: Pleiotropic drug resistance protein 3 (Nicotiana tabacum)	Similar to PDR3: Pleiotropic drug resistance protein 3 (Nicotiana tabacum)	1274477	1281431
1291453	1291625	**s1_8132703**		1278404	1278574
1297550	1299707	Similar to METK5: S-adenosylmethionine synthase 5 (Vitis vinifera)	Similar to METK5: S-adenosylmethionine synthase 5 (Vitis vinifera)	1284674	1285855
1304574	1306456	Similar to ATG8I: Autophagy-related protein 8i (Arabidopsis thaliana)	Similar to ATG8I: Autophagy-related protein 8i (Arabidopsis thaliana)	1293226	1295113
1306781	1318698	Similar to At3g58210: MATH domain and coiled-coil domain-containing protein At3g58210 (Arabidopsis thaliana)	Similar to At3g58210: MATH domain and coiled-coil domain-containing protein At3g58210 (Arabidopsis thaliana)	1295626	1304726
			Protein of unknown function	1305114	1305750
			Similar to At3g58210: MATH domain and coiled-coil domain-containing protein At3g58210 (Arabidopsis thaliana)	1306469	1311005
1315675	1315790	**ZP002**		1307945	1308055
1319766	1320352	Similar to At3g58210: MATH domain and coiled-coil domain-containing protein At3g58210 (Arabidopsis thaliana)	Protein of unknown function	1312087	1312665
1321122	1321207	**MA067**		1313447	1313734
1330290	1331326	Similar to UBP13: Ubiquitin carboxyl-terminal hydrolase 13 (Arabidopsis thaliana)	Similar to UBP13: Ubiquitin carboxyl-terminal hydrolase 13 (Arabidopsis thaliana)	1314280	1317229

start	end	ChrS_2	ChrR_2	start	end
1339639	1347374	Similar to UBP12: Ubiquitin carboxyl-terminal hydrolase 12 (Arabidopsis thaliana)	Similar to UBP12: Ubiquitin carboxyl-terminal hydrolase 12 (Arabidopsis thaliana)	1317863	1331925
1348497	1349955	Similar to PSMG4: Proteasome assembly chaperone 4 (Homo sapiens)	Similar to PSMG4: Proteasome assembly chaperone 4 (Homo sapiens)	1333386	1334449
1351151	1358216	Similar to GFS12: Protein GFS12 (Arabidopsis thaliana)	Similar to GFS12: Protein GFS12 (Arabidopsis thaliana)	1335701	1342742
1359522	1362252	Similar to ACLA-3: ATP-citrate synthase alpha chain protein 3 (Oryza sativa subsp. japonica)	Similar to ACLA-3: ATP-citrate synthase alpha chain protein 3 (Oryza sativa subsp. japonica)	1344049	1346782
1360103	1360169	**PGS1-24**		1344632	1344694
1364384	1365469	Similar to NAC071: NAC domain-containing protein 71 (Arabidopsis thaliana)	Similar to NAC071: NAC domain-containing protein 71 (Arabidopsis thaliana)	1350553	1351611
1371430	1372823	Protein of unknown function			
			Similar to SLX1: Structure-specific endonuclease subunit SLX1 (Cryptococcus neoformans var. neoformans serotype D (strain JEC21 / ATCC MYA-565))	1352163	1353885
1374610	1376367	Similar to MED21: Mediator of RNA polymerase II transcription subunit 21 (Arabidopsis thaliana)	Similar to MED21: Mediator of RNA polymerase II transcription subunit 21 (Arabidopsis thaliana)	1360497	1364062
1375823	1377529	Similar to METTL13: Methyltransferase-like protein 13 (Bos taurus)	Similar to METTL13: Methyltransferase-like protein 13 (Bos taurus)	1363518	1365231
1379702	1387111	Similar to CMTA1: Calmodulin-binding transcription activator 1 (Arabidopsis thaliana)	Similar to CMTA1: Calmodulin-binding transcription activator 1 (Arabidopsis thaliana)	1367497	1374906
1388005	1399532	Similar to KIN12C: Kinesin-like protein KIN-12C (Arabidopsis thaliana)	Similar to KIN12C: Kinesin-like protein KIN-12C (Arabidopsis thaliana)	1375802	1387336
1406092	1407285	Protein of unknown function	Protein of unknown function	1393895	1395089
1411138	1412063	Similar to 1-Cys peroxiredoxin (Medicago truncatula)	Similar to 1-Cys peroxiredoxin (Medicago truncatula)	1398969	1399894
			Similar to At3g59020: Importin beta-like SAD2 homolog (Arabidopsis thaliana)	1402495	1409798
			Similar to ARF1: ADP-ribosylation factor 1 (Salix bakko)	1410072	1411352

Protein sequence similar to MATH domain and coiled – coil domain containing protein At3g58210 annotated in resistant and susceptible haplotype. MATH domains within the protein sequence are highlighted in blu for the resistant haplotype and in green for the susceptible one.

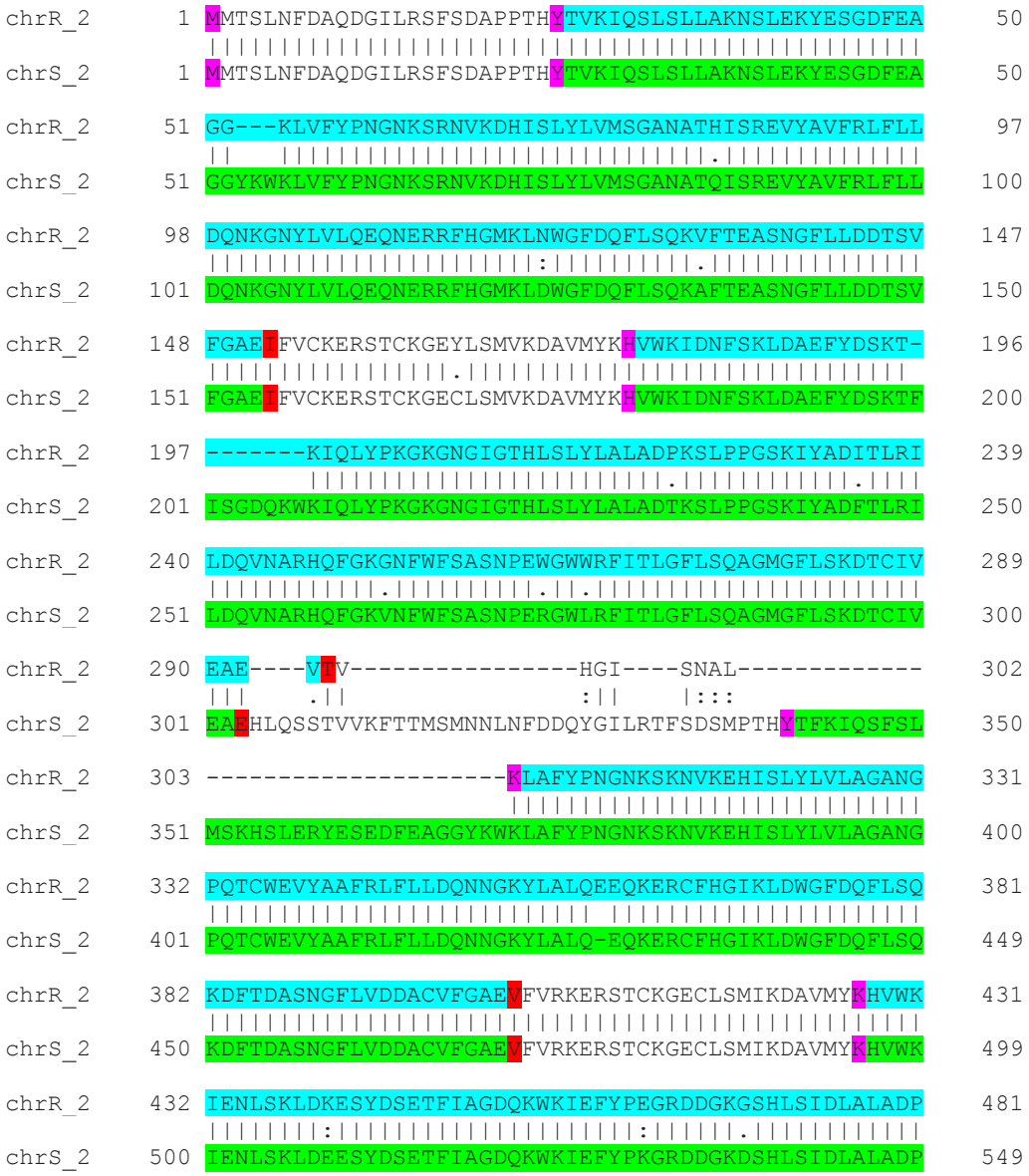
Line 1: chrR_2 - Similar to MATH domain and coiled – coil domain containing protein At3g58210 (518 aa).
Line 2: chrS_2 - Similar to MATH domain and coiled – coil domain containing protein At3g58210 (1004 aa).

```
chrR_2      1 --------------------------------------------------      0
chrS_2      1 MMTSLNFDAQDGILRSFSDAPPTHYTVKIQSLSLLAKNSLEKYESGDFEA     50
chrR_2      1 --------------------------------------------------      0
chrS_2     51 GGYKWKLVFYPNGNKSRNVKDHISLYLVMSGANATQISREVYAVFRLFLL    100
chrR_2      1 --------------------------------------------------      0
chrS_2    101 DQNKGNYLVLQEQNERRFHGMKLDWGFDQFLSQKAFTEASNGFLLDDTSV    150
chrR_2      1 --------------------------------------------------      0
chrS_2    151 FGAETFVCKERSTCKGECLSMVKDAVMYKHVWKIDNFSKLDAEFYDSKTF    200
chrR_2      1 --------------------------------------------------      0
chrS_2    201 ISGDQKWKIQLYPKGKGNGIGTHLSLYLALADTKSLPPGSKIYADFTLRI    250
chrR_2      1 --------------------------------------------------      0
chrS_2    251 LDQVNARHQFGKVNFWFSASNPERGWLRFITLGFLSQAGMGFLSKDTCIV    300
chrR_2      1 --------------------------------------------------      0
chrS_2    301 EAEHLQSSTVVKFTTMSMNNLNFDDQYGILRTFSDSMPTHYTFKIQSFSL    350
chrR_2      1 --------------------------------------------------      0
chrS_2    351 MSKHSLERYESEDFEAGGYKWKLAFYPNGNKSKNVKEHISLYLVLAGANG    400
chrR_2      1 --------------------------------------------------      0
chrS_2    401 PQTCWEVYAAFRLFLLDQNNGKYLALQEQKERCFHGIKLDWGFDQFLSQK    450
chrR_2      1 --------------------------------------------------      0
chrS_2    451 DFTDASNGFLVDDACVFGAEVFVRKERSTCKGECLSMIKDAVMYKHVWKI    500
chrR_2      1 ------------------------------MDMGK-----IKNALT---     11
                                             .|.||     |..||.
chrS_2    501 ENLSKLDEESYDSETFIAGDQKWKIEFYPKGRDDGKDSHLSIDLALADPT    550
```


chrR_2 12 ----SSQCHAEF-LFLVEKIKGNSSKHATVMLCRAQKAGKPSMRKDHHHL 56
chrS_2 551 SLSPTSKLYAQFTLRLVDPV----------YSHRH- 575
chrR_2 57 SDVAMNSEDLSENLQGPW---MNVPSRRRPKAMHKDQGGRNNGVRAKGSR 103
chrS_2 576 ----------FEYGTKATWWFSASSPKRGWPKFI----------TLGIF 604
chrR_2 104 FDALHGVSEN---FCQEELIVNGAEGQSF----YGKREPSIRDAG--LGK 144
chrS_2 605 GDESVGYLENDSTILEAEMTPVCLEQRSFFVKSYRERRVSINDKNMDMGK 654
chrR_2 145 KVWTKS--KVVKPDVRVALNDISNRPQQDKKHLTVAARNEGGAKSTHLQS 192
chrS_2 655 IVWHDQGYEFYRLDGKISQSTFKVR----------QS 681
chrR_2 193 SAVKLFAMNMTSLNFDEQDGILRTISDAPPTHYMIKIQSLSLLSVHSLEK 242
chrS_2 682 S---VFAMNMTSLNFDEQDGILRTISDVPPTHYTIKIQSLSLLSVHSLEK 728
chrR_2 243 YESGEFEAGGYKWKLVFYPNGNKSRNVKEHISLYLVLAGANAPKTCWEVH 292
chrS_2 729 YESGVFEAGGYKWKLVFYPNGNKSSNGKEHISLYLVLAGANGPQTCWEVH 778
chrR_2 293 AAFRLFLLDQNIGKYFAFQEQNERCFHGMKLDWGFDKCLSLKAFTDASNG 342
chrS_2 779 AAFRLFLLDQNTGKYLALQEKNERRFHGMKLDWGFDQFLSLKAFTDTSNG 828
chrR_2 343 FLVEDTCVFGAEVFVRKERSTCKGECLSMIKGAIMYKHVWKIDNFSKLNA 392
chrS_2 829 FLMEDACVFGAEVFVRKEKSTCKGECLSMIKDAVMYKHVWKIDNFSKLNA 878
chrR_2 393 ESYDSQTFIAGDQKWKIKLYPKGRDGAASGHLSLYLALADPTSLPPTSKI 442
chrS_2 879 ESYDSPTFIAGNQKWKIRLYPKGRDSGTGSHLSLYLALADPTSLPPTSKI 928
chrR_2 443 YAEFTLRLINQQNSSLHYAYSKVNWWFSASSPMRGWGRFITVGWFYVNQA 492
chrS_2 929 YAQYTLRIINQLNSPYPYEYSKVTWWFSASSPSRGWPSFITIGYFNIAQS 978
chrR_2 493 NYRYLVNDSCTVEAEVTVHGTASALE 518
chrS_2 979 NWGYLVKDSCTVEAEVTVHGTASALD 1004

Protein sequence similar to MATH domain and coiled – coil domain containing protein At3g58210 annotated in resistant and susceptible haplotype. MATH domains within the protein sequence are highlighted in blu for the resistant haplotype and in green for the susceptible one.

Line 1: chrR_2 - Similar to MATH domain and coiled – coil domain containing protein At3g58210 (565 aa).
Line 2: chrS_2 - Similar to MATH domain and coiled – coil domain containing protein At3g58210 (1004 aa).



```
chrR_2    482 TSLSPTSKLYAQFTLRLVDPVYSSRHFEYGAKATWWFSASSPKRGWPKFI    531
              ||||||||||||||||||||||.||||||.||||||||||||||||||||
chrS_2    550 TSLSPTSKLYAQFTLRLVDPVYSHRHFEYGTKATWWFSASSPKRGWPKFI    599

chrR_2    532 TLGHFSDKTLGYLENDSTIVEAEVTVLGTASALD----------------    565
              |||.|.|:::|||||||||:|||:|.:    .|:
chrS_2    600 TLGIFGDESVGYLENDSTILEAEMTPV----CLEQRSFFVKSYRERRVSI    645

chrR_2    566 --------------------------------------------------    565

chrS_2    646 NDKNMDMGKIVWHDQGYEFYRLDGKISQSTFKVRQSSVFAMNMTSLNFDE    695

chrR_2    566 --------------------------------------------------    565

chrS_2    696 QDGILRTISDVPPTHYTIKIQSLSLLSVHSLEKYESGVFEAGGYKWKLVF    745

chrR_2    566 --------------------------------------------------    565

chrS_2    746 YPNGNKSSNGKEHISLYLVLAGANGPQTCWEVHAAFRLFLLDQNTGKYLA    795

chrR_2    566 --------------------------------------------------    565

chrS_2    796 LQEKNERRFHGMKLDWGFDQFLSLKAFTDTSNGFLMEDACVFGAEVFVRK    845

chrR_2    566 --------------------------------------------------    565

chrS_2    846 EKSTCKGECLSMIKDAVMYKHVWKIDNFSKLNAESYDSPTFIAGNQKWKI    895

chrR_2    566 --------------------------------------------------    565

chrS_2    896 RLYPKGRDSGTGSHLSLYLALADPTSLPPTSKIYAQYTLRIINQLNSPYP    945

chrR_2    566 --------------------------------------------------    565

chrS_2    946 YEYSKVTWWFSASSPSRGWPSFITIGYFNIAQSNWGYLVKDSCTVEAEVT    995

chrR_2    566 ---------    565

chrS_2    996 VHGTASALD   1004
```

Comparison of the sequence of the protein of unknown function annotated in the resistan haplotype and the protein similar to MATH domain and coiled – coil domain containing protein At3g58210 annotated the susceptible haplotype. MATH domains within the protein sequence is highlighted in green for the susceptible one.

Line 1: chrR_2 – Protein of unknown function (96 aa)

Line 2: chrS_2 - Similar to MATH domain and coiled – coil domain containing protein At3g58210 (124 aa).

```
chrR_2       1 MASSGMINICWNR--SIFLGKAKLSIHVTFCQQRH---------HFS---      36
               ||||||:|||||.  :|.:.|.:     .|..:::          |.|
chrS_2       1 MASSGMVNICWNSKDTILVRKGQ-----NFRDKKYVGKWQWKGTHVSVFL      45

chrR_2      37 -QERPE--FQGQEIC----------------PGYTWFNALTPAWGRQTFI      67
                ...||  ..|.::.                .||||||||||||||.||
chrS_2      46 RLANPEKLSPGSQLLTEYTLRIVDQLNAKHKTGYTWFNALTPAWGRQAFI      95

chrR_2      68 KLGTFKMSDRGYLVNNARGRGRGHCTWNC     96
               ||||.||||:||||||||||.|||||.|||
chrS_2      96 KLGTSKMSDQGYLVNNARGRSRGHCPWNC    124
```

Protein sequence similar to UBP13: Ubiquitin carboxyl-terminal hydrolase 13 (*Arabidopsis thaliana*) annotated in resistant and susceptible haplotype. MATH domains within the protein sequence are highlighted in blu for the resistant haplotype and in green for the susceptible one.

Line 1: chrR_2 – Similar to UBP13: Ubiquitin carboxyl-terminal hydrolase 13 (*Arabidopsis thaliana*) – 270 aa

Line 2: chrS_2 - Similar to UBP13: Ubiquitin carboxyl-terminal hydrolase 13 (*Arabidopsis thaliana*) – 244 aa

```
chrR_2       1 MSLIFDQDGLSRSLSNSPPTHYTLTIESFSMLTENSVDT-YESG---ELF     46
                                             |...||:.| :|..   .||
chrS_2       1 ------------------------------MAEANSLQTGWEVSVDFRLF     20

chrR_2      47 LLDQNKGIYLVLQDANMNKMCLHGAMLQVGFDRVIPLNAFSVASNGYLID     96
               ||||||||||||||||||||||||||||||||||||||||||||||||||
chrS_2      21 LLDQNKGIYLVLQDANMNKMCLHGAMLQVGFDRVIPLNAFSVASNGYLID     70

chrR_2      97 DTCVFGAEVFVCKERRAGKAECLSRIKKAFMNKHCWKIESFSTLLFQCLQ    146
               ||||||||||||||||||||||||||||||||||||||||||||..||||
chrS_2      71 DTCVFGAEVFVCKERRAGKAECLSRIKKAFMNKHCWKIESFSTLKSQCLQ    120

chrR_2     147 SELFTAGGQKWKIELYPKGDGDGENTHVSVYLSLLANPEKLSPGSQLLTE    196
               ||||||||||||||||||.||||||||||||||||||||||||||||||
chrS_2     121 SELFTAGGQKWKIELYPKGDDDGENTHVSVYLSLLANPEKLSPGSQLLTE    170

chrR_2     197 CTVRIVDQLNGKDKSRELNHAWFSASSSSWGWPCFIKLDSFKMLDNGYLV    246
               |||||||||:|||||||||:||||||||:||||||||||||||||||||
chrS_2     171 CTVRIVDQLDGKDKSRELNYAWFSASSSTWGWPCFIKLDSFKMLDNGYLV    220

chrR_2     247 KNTCLVEAEVTVHGIAKALEPTDD    270
               ||||||||||||||||||||||||
chrS_2     221 KNTCLVEAEVTVHGIAKALEPTDD    244
```

Protein sequence similar to UBP13: Ubiquitin carboxyl-terminal hydrolase 12 (*Arabidopsis thaliana*) annotated in resistant and susceptible haplotype. MATH domains within the protein sequence are highlighted in blu for the resistant haplotype and in green for the susceptible one.

Line 1: chrR_2 – Similar to UBP12: Ubiquitin carboxyl-terminal hydrolase 13 (*Arabidopsis thaliana*) – 605 aa

Line 2: chrS_2 - Similar to UBP12: Ubiquitin carboxyl-terminal hydrolase 13 (*Arabidopsis thaliana*) – 729 aa

```
chrR_2      1 MATLNLKFEPDAAESFSSLERHSAGRYESGQFDAGGYKWKLVVYPNGYKQ     50
              ||||||.||||||||||||||||||||||||||||||||||||||||||||
chrS_2      1 MATLNRKFEPDAAESFSSLERHSAGRYESGQFDAGGYKWKLVVYPNGYKQ     50

chrR_2     51 KNVDDHISVYLEMAGADLLQTGWEVFVDFRLFLLDQNKGIYLVLQDANLN    100
              ||||||||||||||||.||||||||||||||||||||||||||||||||||
chrS_2     51 KNVDDHISVYLEMAGADSLQTGWEVFVDFRLFLLDQNKGIYLVLQDANLN    100

chrR_2    101 KMCLHGAMFEVGFDRVIPLNAFTDSSNGYLINDTCVFGAEVFVCKERRAG    150
              ||||||||||||||||||||||||||||||||||||||||||||||||||||
chrS_2    101 KMCLHGAMFEVGFDRVIPLNAFTDSSNGYLINDTCVFGAEVFVCKERRAG    150

chrR_2    151 KAERLYTINSAMYKHPWKVYIPLKFRPELLESKPFFAGGQTWEIRLYPKG    200
              ||||||||||||||||||||||||||||||||||||||||||:|||||||
chrS_2    151 KAERLYTINSAMYKHPWKVYIPLKFRPELLESKPFFAGGQTWKIRLYPKG    200

chrR_2    201 YDKGKDTHVSVYLKLANPEPASKILTEFTLRIVDQLNGKHFFCKGCEWFC    250
              ||||||||||||||||||||||||||||||||||||||||||||||||||||
chrS_2    201 YDKGKDTHVSVYLKLANPEPASKILTEFTLRIVDQLNGKHFFCKGCEWFC    250

chrR_2    251 ALRPSF---GFSRLIAFDILKQLDKG----VSTPFSYTPPTHYTLKIESF    293
              ||||||   ||     |...:.|..|    ||||||||||||||||||||
chrS_2    251 ALRPSFVRQGF-----FGEEQLLSGGRGHRVSTPFSYTPPTHYTLKIESF    295

chrR_2    294 SLLKKHSADRFESGEFDAGGYKWKLVVYPNGYEKKNVEDHISVYLEMAGA    343
              ||||||||||||||||||||||||||||||||||||||||||||||||||||
chrS_2    296 SLLKKHSADRFESGEFDAGGYKWKLVVYPNGYEKKNVEDHISVYLEMAGA    345

chrR_2    344 ESLETGWEVFVDFRLFLLDQNKGIYLVLQDANLKKMCLHVAMLEVGFDRV    393
              ||||||||||||||||||||||||||||||||||||||||||||||||||||
chrS_2    346 ESLETGWEVFVDFRLFLLDQNKGIYLVLQDANLKKMCLHVAMLEVGFDRV    395

chrR_2    394 IPLKAFADASNGYLIDDTCVFGAEVFVCKERRAGKAECLPRINNAVIVSE    443
              ||||||||||||||||||||||||||||||||||||||||||||||||||||
chrS_2    396 IPLKAFADASNGYLIDDTCVFGAEVFVCKERRAGKAECLPRINNAVIVSE    445

chrR_2    444 ENNDFMNKHVWKIEEFSKLKPEPLESKPFNAGGQTWKIQLFPKGDSHGKD    493
              ||||||||||||||||||||||||||||||||||||||||||||||||||||
chrS_2    446 ENNDFMNKHVWKIEEFSKLKPEPLESKPFNAGGQTWKIQLFPKGDSHGKD    495

chrR_2    494 THVSLYLTLANPEKLSTAPKILAQFTLRIVDQLNAKHFFRHDSNCFRASS    543
              ||||||||||||||||||||||||||||||||||||||||||||||||||||
chrS_2    496 THVSLYLTLANPEKLSTAPKILAQFTLRIVDQLNAKHFFRHDSNCFRASS    545
```

```
chrR_2    544 PSWGWSNFIMLGFFKERDK-----------GYLVMNTCVVEAEDVQAML    581
              ||||||||||||||||||             .|.:             .:
chrS_2    546 PSWGWSNFIMLGFFKERDKEVSRPFSDSPPTHYFL-------------KI    582

chrR_2    582 PKLTLMEQLMVKMLEKMDNNTKAI--------------------------    605
              ...:|:::.....|..:.:....
chrS_2    583 ESFSLLKKYSADRFESGEFDAGGYKWKLVVYPNGYKKKNVEDHISVYLEM    632

chrR_2    606 -------------------------------------------------    605

chrS_2    633 AGAESLQTDANLKKMCLHAAMLEVGFDRVIPLKAFADASNGYLIDDTCVF    682

chrR_2    606 -------------------------------------------------    605

chrS_2    683 GAEVFVCKERRAGKAECLPRINNAVIVSKENYDFLHKKENNDFVNKHVWK    732

chrR_2    606 ------------------------    605

chrS_2    733 IEQFSKLTPERLESKPLNAGGQTW    756
```

Tab. 6 - List of software used in this work. On the left, the activity carried out with each software is reported.

Description	Software	
Quality assessment of Illumina reads	bbduk2	https://github.com/BioInfoTools/BBMap/blob/master/sh/bbduk2.sh
	cutadapt	https://cutadapt.readthedocs.io/en/stable/
	ERNE-FILTER	erne-soruceforge.net
BAC clones assembly	CLC Genomics Workbench v3	https://www.qiagenbioinformatics.com/products/clc-genomics-workbench
	iAssembler-v1.3.2	http://bioinfo.bti.cornell.edu/tool/iAssembler/
Sequence alignment	Dotter tool	http://www.sanger.ac.uk/ science/tools/seqtools
	GEvo comparative sequence alignment	https://genomevolution.org/coge/GEvo.pl
	Nucleotide-Nucleotide BLAST 2.2.27+	ftp://ftp.ncbi.nlm.nih.gov/blast/
	NUCleotide MUMer	http://mummer.sourceforge.net/
Reads alignment	BWA mem	http://bio-bwa.sourceforge.net/
	BLASR	https://github.com/PacificBiosciences/blasr
Files index	BWA index	http://bio-bwa.sourceforge.net/
	Samtools	http://samtools.sourceforge.net/
Assembly of pacbio reads	Canu software	https://github.com/marbl/canu
SNPs/indels calling	Unified Genotyper of GATK	https://software.broadinstitute.org/gatk/download/
Gene prediction annotation	MAKER pipeline	http://www.yandell-lab.org/software/maker.html
	RepeatMasker	http://www.repeatmasker.org/
	Exonerate	https://www.ebi.ac.uk/about/vertebrate-genomics/software/exonerate
	SNAP	https://github.com/KorfLab/SNAP
	Augustus	http://augustus.gobics.de/

ACKNOWLEDGMENTS

I thank all the partner of the national research project PRIN-VIRES for the opportunity offered to work on this project.

I thank Rachele Falchi, Nicoletta Felice, Giusi Zaina and Alessandro Spadotto for the support and advice during the laboratory activity; Irena Jurman and Federica Cattonaro for Illumina sequencing; Simone Scalabrin for assistance in *de novo* assembly and the valuable bioinformatics support; prof. Guido Cipriani and all my collegues and friends Serena Foria, Catalina Pinto, Corinne Monte, Alice Colussi, Nicola Zorzin for advice and support.

I also thank prof. Dick de Ridder for the time I spent at bioinformatics group at Wageningen University in Netherlands.

I would especially like to thank my supervisor, prof. Raffaele Testolin, for his support, advice and guidance during this work.