COMPARATIVE APPROACHES FOR PLANT GENOMES: UNRAVELING “INTRA” AND “INTER” SPECIES RELATIONSHIPS FROM PRELIMINARY GENE ANNOTATIONS

LUCA AMBROSINO

Università degli Studi di Napoli 'Federico II'

PhD in Computational Biology and Bioinformatics
(Cycle XXVIII)

TUTOR: Doc. Maria Luisa Chiusano

COORDINATOR: Prof. Sergio Cocozza

May 2016

Abstract

Comparative genomics studies the differences and similarities between different species, to transfer biological information from model organisms to newly sequenced genomes, and to understand the evolutionary forces that drive changes in genomic features such as gene sequences, gene order or regulatory sequences.

The detection of ortholog genes between different organisms is a key approach for comparative genomics. For example, gene function prediction is primarily based on the identification of orthologs. On the other hand, the detection of paralog genes is fundamental for gene functions innovation studies. The fast spreading of whole genome sequencing approaches strongly enhanced the need of reliable methods to detect orthologs and paralogs and to understand molecular evolution.

In this thesis, methods for predicting orthology relationships and exploiting the biological knowledge included within sets of paralog genes are shown.

Although the similarity search methods used to identify orthology or paralogy relationships are generally based on the comparison between protein sequences, this analysis can lead to errors due to the lack of a correct and exhaustive definition of such sequences in recently sequenced organisms with a still preliminary annotation. Here we present a methodology that predict orthologs between two species by sequence similarity searches based on mRNA sequences.

Moreover, the features of a web-accessible database on paralog and singleton genes of the model plant *Arabidopsis thaliana*, developed in our lab, are described. Duplicated genes are organized into networks of paralogs, whose

graphical display and analysis enable the investigation of gene families structural relationships and evolution.

Consequently, we applied the developed methodologies to the cross comparison between some economically important plant species, such as tomato, potato and grapevine. The similarities between two distantly related species such as tomato and grapevine, belonging to two different clades, and the distinctive aspects between two closely related members of the family of Solanaceae, potato and tomato, are also highlighted.

Understanding different and common mechanisms that underlie these crop species could provide valuable insights in plant physiology.

To my two girls.

Content

Chapter 1. Introduction

1.1 Comparative genomics in plants

Comparative genomics studies the differences and similarities in genomic features of different species, either to transfer information from well-defined model organisms to those with newly sequenced genomes, or to understand the evolutionary forces that drove changes between species (Xia 2013, Sharma et al. 2014). The reduced costs of sequencing technologies has recently pushed the increase in the number of complete genomes, driving ambitious efforts that aim to the sequencing at species and intra-species level. Therefore, comparative genomics efforts are proving to be more effective considering the key feature of comparative genomics associations, which asserts that the number of matches that can be found among genomes grows as the square of the number of the available genomes (Overbeek et al. 1999, Hanson et al. 2010, Bradbury et al. 2013). The combination of bigger datasets and better tools will further increase the cost-effectiveness of structure and function discoveries via comparative genomics analysis (de Crecy-Lagard and Hanson 2007).

Currently, more than 31000 genomes, including numerous plant genomes, are available in public databases (http://www.ncbi.nlm.nih.gov/genome). The availability of such a huge amount of genomic data has increased the knowledge about gene families evolution and how events like gene duplication, gene loss and gene fusion/fission shaped genome structure and organization (Snel et al. 2000, Snel et al. 2002, Koonin 2005, Dorman 2013). However, due also to limitations in the annotation of recently sequenced genomes, many conserved genes between different species have still no assigned function or share an ambiguous annotation. A major challenge for comparative genomics is the correct prediction of the function for these genes. Although some experimental approaches like microarray analysis, RNA interference, and the yeast two-

hybrid system can be used to experimentally demonstrate the function of a protein, contributing to the expansion of biological knowledge, advances in sequencing technologies have made the rate at which proteins can be experimentally characterized much slower than the rate at which new sequences become available (Gabaldon and Huynen 2004).

1.1.1 *Arabidopsis thaliana* as a model for plants

In the last century, it has been spread the practice of focusing the topic of biological study on a small group of model organisms. This process took place thanks to the stunning development of genetic, molecular and genomic resources that shed light in organisms such as *Saccharomyces cerevisiae*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Mus musculus*, and the plant *Arabidopsis thaliana*. Model organisms were usually chosen based on their small size, short generation time, inbreeding habit and large progeny numbers. Experiments focused on model organisms led to a drastic expansion of biological knowledge involving different areas of science.

Arabidopsis thaliana (Fig. 1), a plant of the Brassicaceae family (Fig. 2), is widely distributed throughout Europe, Asia, and North America.

Figure 1. The *Arabidopsis thaliana* plant.

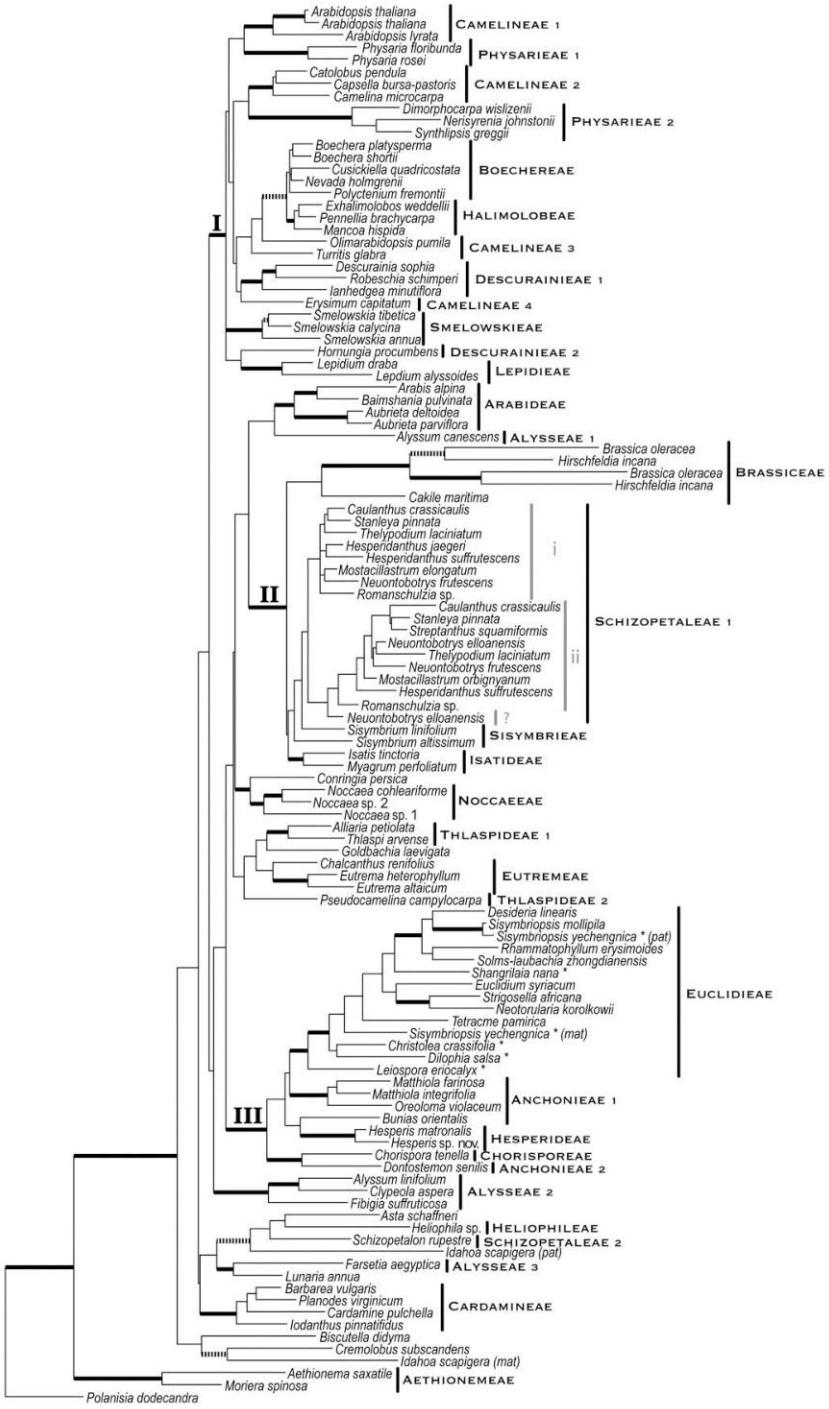


Figure 2. Brassicaceae phylogeny inferred from phytochrome A and ndhF sequence data. Maximum likelihood phylogeny of Brassicaceae phytochrome A showing tribes and lineages (l–lll). Image extracted from (Beilstein et al. 2008)

The entire life cycle, from seed germination to seeds maturation, requires 6 weeks. A mature plant reaches 15 to 20 cm in height, producing approximately 5000 total seeds. The roots have a simple structure, without any symbiotic relationships with nitrogen-fixing bacteria. Natural pathogens of *Arabidopsis thaliana* include different bacteria, fungi, viruses and insects (Meinke et al. 1998).

Arabidopsis genome size is approximately 120 megabases (Mb), organized into 5 chromosomes and containing 33,604 genes (Lamesch et al. 2012). However, several analyses of the genome of *A. thaliana* revealed a high complexity, probably due to at least three events of "Whole Genome Duplications" (Blanc et al. 2000, Vision et al. 2000, Wolfe 2001, Blanc et al. 2003, Blanc and Wolfe 2004, Cui et al. 2006, Jiao et al. 2011, Van de Peer 2011, Jiao et al. 2012), and to other events like translocations, inversions (Ku et al. 2000, Gaut 2001), or chromosome losses (Conner et al. 1998, Johnston et al. 2005, Lysak et al. 2005).

1.1.2 *Solanum lycopersicum:* a reference for Solanaceae

The family of Solanaceae, or nightshade, groups together many fruit and flower species (Knapp 2002, He et al. 2004), some of which with high economic relevance. Many plants of this family, including Tomato (*Solanum lycopersicum*), Potato (*Solanum tuberosum*), Eggplant (*Solanum melongena*) and Pepper (Capsicum spp.), play an important role in the human diet. Moreover, some species of Physalis and Lycium are used both in medicine and in food supply (Wang et al. 2015). Fruits belonging to this family show a pronounced morphological diversity (Knapp 2002), including color, size and shape (Fig. 3).

Figure 3. Fruit morphology in Solanaceae. (1–3), Solanum melongena; (4), Solanum pimpinellifolium; (5–8), Solanum lycopersicum; (9–14), Variants of Capsicum annum; (15), Physalis alkekengi; (16), Physalis floridana; (17–19), Physalis philadelphica. The Chinese lantern in Physalis spp. was opened to show the berry inside. Bar = 1 cm. Image extracted from (Wang et al. 2015)

Some Solanaceae species are widely considered as model organisms for plant genomics and biodiversity studies, most notably Tomato, Tobacco (*Nicotiana tabacum*) and *Petunia hybrida* (Knapp et al. 2004).

The family groups together about 90 genera and 4000 species half of which belonging to the large Solanum genus. This considerable diversity in just one genus that includes both annual and perennial plants from different habitats (The_Tomato_Genome_Consortium 2012) is uncommon in angiosperms, making Solanum interesting both from an evolutionary point of view and for its widespread use in the human diet (Knapp et al. 2004).

Among species belonging to the Solanum genus, Tomato is widely accepted as a model species for Solanaceae, and a reference for studies on fleshy fruit development (Gapper et al. 2013).

Tomato is a highly homozygous diploid, easy to sequenced. Its genome size is 900 megabases (Mb), distributed in 12 chromosomes and containing 34,727 genes. The genome shows a high level of synteny with other economically important Solanaceae species (Fig. 4) (The_Tomato_Genome_Consortium 2012).

In comparison to *Arabidopsis thaliana* genome, Tomato has fewer high-copy, long terminal repeat (LTR) retrotransposons. This confirms previous findings that the Tomato genome, being mostly comprised of low-copy DNA, is unusual among angiosperms (The_Tomato_Genome_Consortium 2012).

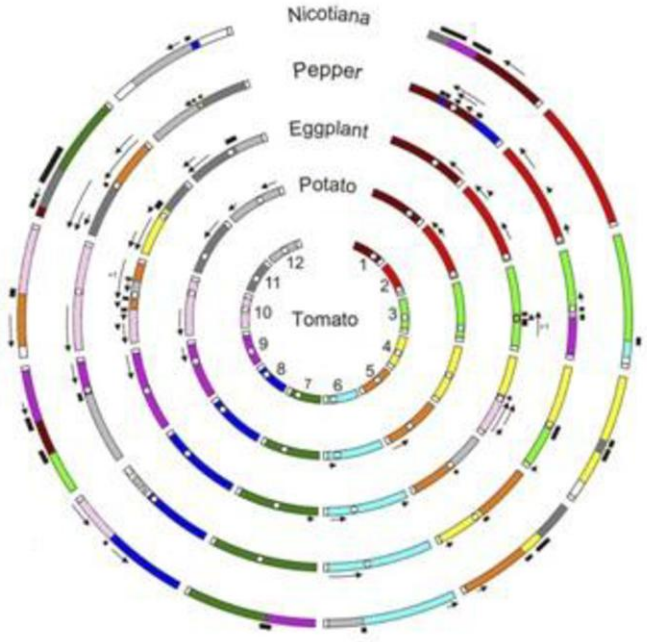


Figure 4. Syntenic relationships in the Solanaceae. Comparative maps of Potato, Eggplant, Pepper and Nicotiana with respect to the Tomato genome. Each Tomato chromosome is represented in a different color and orthologs chromosome segments in other species are shown in the same color. White dots indicate approximate centromere locations. Each black arrow indicates an inversion relative to Tomato and "+1"indicates a minimum of one

inversion. Each black bar beside a chromosome indicates translocation breakpoints relative to Tomato. Picture extracted from (The_Tomato_Genome_Consortium 2012)

1.2 Comparative genomics and homologous genes

Most computational methods for comparative genomics analysis are based on initial similarity searches to detect homology relationships (Coutinho et al. 2015), allowing, among others, the annotation of new genomes based on orthology inference (Moriya et al. 2007) and the estimation of evolutionary rates within gene families (Luz and Vingron 2006). Such comparative studies rely on the analysis of ortholog and paralog genes (Fitch 1970), and consequently on their accurate detection (Trachana et al. 2014).

Orthologs are genes in different species that started diverging from a common ancestor via evolutionary speciation (Fig. 5) (Fitch 1970, Altenhoff and Dessimoz 2012, Chen and Zhang 2012). In comparative genomics, orthologs are used to transfer annotation from characterized genes to loci from newly sequenced genomes. One of the crucial steps of any new genome project is to perform a precise structural and functional annotation, and this is partially reached by defining ortholog relationships with reference gene annotations (Pereira et al. 2014).

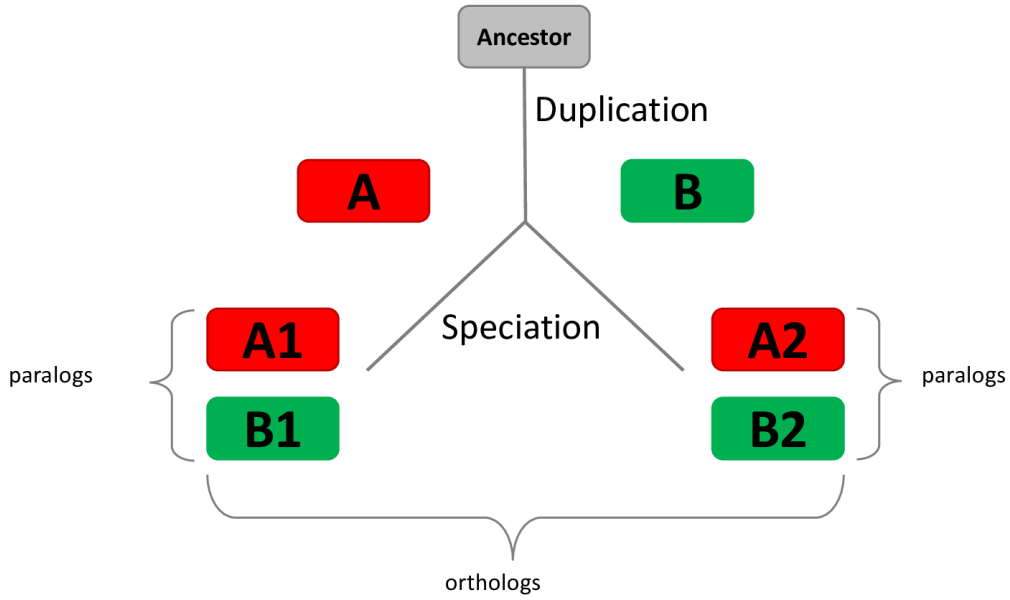


Figure 5. Orthology and paralogy relationships in the evolution of four different genes (A1, A2, A3, A4) that arose from a single common ancestor.

Paralogs are genes in the same species that started diverging via gene duplication (Fig. 5) (Fitch 1970, Altenhoff and Dessimoz 2012). Gene duplication is a fundamental mechanism for creating genetic novelty in organisms by providing new material for gene functions innovation (Long et al. 2003, Magadum et al. 2013). The majority of duplicated genes will vanish over time, while a smaller subset evolves novel or more complex functions (Lynch and Conery 2000). Since plants are particularly susceptible to evolve novel functions via small-scale and large-scale duplication events, the retention of paralogs after gene and genome duplication can act as a driver for their evolution (Rensing 2014). In the model plant *Arabidopsis thaliana*, for instance, paralogs involved in signaling and transcriptional regulation mechanism are more often retained than other genes after "Whole Genome Duplication" events (Blanc and Wolfe 2004, Seoighe and Gehring 2004, Maere et al. 2005).

Overall, based on the "ortholog conjecture" (Altenhoff et al. 2012, Pereira et al. 2014, Rogozin et al. 2014), or standard model of phylogenomics, which claims

that ortholog genes are functionally more similar than paralog genes, protein function changes rapidly after duplication, leading to paralogs with diverged in function, while orthologs tend to have a conserved function. Hence, most interest for orthology is in the context of computational function prediction, while paralogs are commonly used to study gene families organization and function innovation. A coarse approach consists in transferring the functional annotation between one-to-one orthologs. However, scaling the whole evolutionary history of different species into pairwise relationships dares to be an oversimplification. On the contrary, capturing and modeling more evolutionary features as possible, such as gene structures and phylogenetic distances, seems to be the best solution to decipher differences and similarities between different organisms (Altenhoff and Dessimoz 2012).

1.3 Network biology and bioinformatics

Network theory is part of a variety of disciplines, ranging from communications and engineering to medicine and molecular biology (Albert and Barabási 2002, Dorogovtsev and Mendes 2002, Alm and Arkin 2003, Alon 2003, Bray 2003, Newman 2003, Barabasi and Oltvai 2004).

In biology and medicine, for example, the theory of complex networks is involved in application such as drug targets identification (Mason and Verwoerd 2007), function detection of proteins or genes with an unknown annotation (Jeong et al. 2003, Samanta and Liang 2003), or strategy design to treat infective diseases (Eubank et al. 2004). Moreover, the recent rise of the "omics" technologies made available a large amount of information on molecular networks in different organisms (Costanzo et al. 2000, Ito et al. 2001). However, there is the need of a consistent bioinformatics effort to grasp meaningful biological information from the big amount of data coming from expanding high-throughput techniques (Mason and Verwoerd 2007).

Networks are relevant to investigate several aspect in biology. Protein-protein interaction (PPI) networks evince the direct or indirect interactions between the proteins of an organism (Costanzo et al. 2000, Uetz et al. 2000, Ito et al. 2001, Rain et al. 2001, Giot et al. 2003, Li et al. 2004), metabolic networks display biochemical reactions between different compounds (Kanehisa and Goto 2000, Ravasz et al. 2002, Karp et al. 2005), and transcriptional regulatory networks show the regulation activity between different genes (Ihmels et al. 2002, Shen-Orr et al. 2002, Salgado et al. 2006). Thus, the network theory contributes to the representation of such biological relationships and to investigate their key properties (Mason and Verwoerd 2007). The topology of a network may be useful to represent its biological meaning. Often, specific patterns or topologies of a network allow researchers to associate it to specific conditions. Understanding the topology of biological networks, for example, is mandatory to develop effective treatment strategies in severe diseases such as cancer (Vogelstein et al. 2000).

The mathematical discipline that enables the correct study of biological networks is the graph theory (Diestel 2010). Graphs, or networks, can be divided into directed graphs and undirected graphs (Fig. 6).

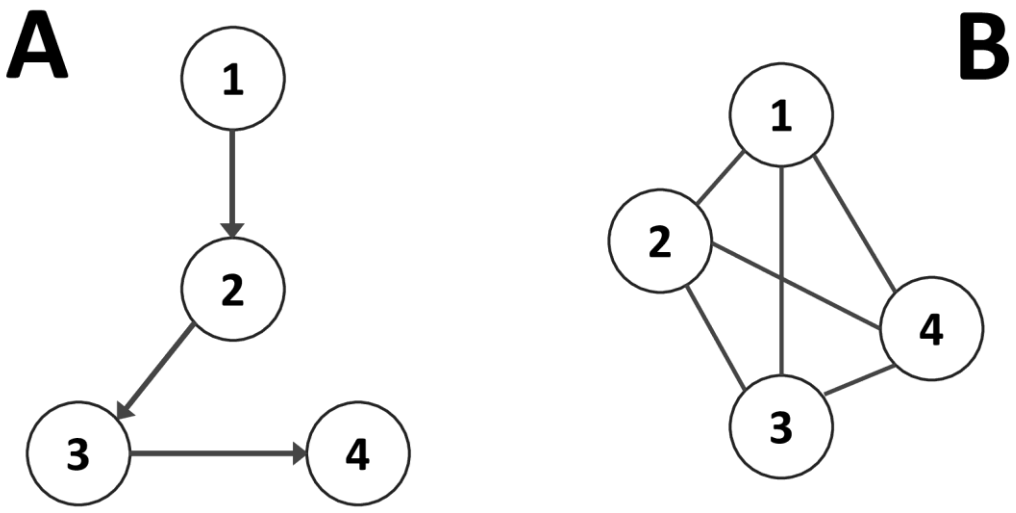


Figure 6. A) Example of a directed graph, comprising four nodes and three edges. **B)** Example of an undirected graph, comprising four nodes and six edges.

A directed graph *G* consists of a set of nodes, *N*(*G*), from 1 to j

$$N(G) = \{n_1,\ldots,n_j\}$$

connected by one or more edges, $E(G) \subseteq N(G)$ X $N(G)$. Each edge $E(m, n)$ connects the starting node n_i to the node n_j (Diestel 2010). The direction of the edges has relevance with the property of the network. Examples of biological networks modelled as directed graphs are metabolic networks or transcriptional regulatory networks (Mason and Verwoerd 2007). In a metabolic network, for example, nodes represent compounds with edges denoting the chemical reactions that converts the substrates into products. As each reaction has a natural direction, such networks are modelled as directed graphs.

An undirected graph, G, also consists of a set of nodes, *N(G)*, and a set of edges set, *E*(*G*). However, in this case the edges do not directionality. The number of nodes in a directed or undirected graph defines the size or order of the graph

(Diestel 2010). Examples of of biological networks modelled as an undirected graph are the PPI networks, in which nodes represent proteins and edges represent physical interactions (Mason and Verwoerd 2007). Two nodes n_i and n_j connected by an edge, in both directed and undirected graph, are adjacent to each other: in graph theory language, n_i and n_j are neighbors (Mason and Verwoerd 2007).

The detection of possible key nodes of a network is a challenge in many application areas, such as communications or management. Several measures and algorithms, called centrality measures, have been developed for ranking the nodes of a network and quantifying their level of importance. A famous example is represented by the PageRank algorithm that enables GOOGLE to find the most relevant web-pages to a specific user query (Mason and Verwoerd 2007). As an example, the centrality measure for a node could be represented by the number of edges connecting that node to other nodes.

A recent trend in research is to apply such centrality measures in order to identify structurally important genes or proteins. Depending on the biological question, it may be crucial to detect central nodes or intermediate nodes that affect the topology of a biological network (Mason and Verwoerd 2007). In particular, researchers are trying to weight the relationship between centrality and essentiality, where a gene or protein is said to be essential for an organism if the organism dies without it. The use of centrality measures to predict essentiality based on network topology has potentially significant applications in many scientific areas (Vogelstein et al. 2000, Jeong et al. 2003).

The fast development of the omics technologies has generated massive amounts of data and the complexity of biological networks increases as data are accumulating (Pavlopoulos et al. 2011). Consequently, bioinformatics is crucial to integrate data arising from different sources and to derive meaningful information from network analysis.

1.4 Aims and scope

The following chapters of this thesis present two different comparative genomics tools and two distinct investigations. In the second chapter, a web-accessible resource on paralog and singleton genes from the model plant *Arabidopsis thaliana* is presented. In particular, I contributed implementing a graphical visualization of the networks of paralog genes within the web pages, to enable a fast exploitation of the information derivable from the graph analysis. The work presented in the second chapter was published in 2016 (Ambrosino et al. 2016). In the third chapter, a transcriptome-based approach to predict orthology relationships is described. The transcriptomes of *Arabidopsis thaliana* and *Sorghum bicolor* were used to test the implemented algorithms and to develop a pipeline able to detect orthologs between two species. A manuscript describing the method showed in the third chapter is under revision at the time of this thesis submission. The fourth chapter describes the comparison between two distantly related species, Tomato and Grapevine, considered as economically important species from asterids and rosids clades, respectively. Understanding different and common mechanisms that underlie these two fleshy fruit species could reveal important knowledge to the plant research. Both orthology and paralogy relationships were detected by a multilevel approach using sequences from genes, transcripts and proteins, and parallel functional analysis were conducted to fulfill the aim of the work. The fifth chapter describes the comparison between Tomato and Potato, both belonging to the family of Solanaceae. Also in this case a multi-level approach was applied in order to detect orthology and paralogy relationships between the two Solanaceae species. The obtained results from this comparison highligths common features and peculiar aspects between these two crop species. The

analyses performed in the fourth and fifth chapters are the subjects of two manuscripts in preparation at the time of this thesis submission.

The general aim of this thesis is:

1) the development of tools useful in comparative genomics strategies, such as the investigation of the paralog and single copy genes of the reference species *Arabidopsis thaliana* through a network-based approach(Chapter 2), or the identification of orthology relationships starting from mRNA sequences (Chapter 3);

2) the application of such methods and tools to the comparison between some economically important species, such Tomato, Potato and Grapevine (Chapter 4 and 5).

Chapter 2. A database of paralog and singleton genes from the reference plant *Arabidopsis thaliana*

2.1 Introduction

Arabidopsis thaliana, belonging to the family of *Brassicaceae*, was the first plant to be completely sequenced in 2000, being a reference species for plants thanks to its short generation time, the small size that limited the requirement for growth facilities, the prolific seed production through self-pollination, and its small diploid genome (The_Arabidopsis_Genome_Initiative 2000, Koornneef and Meinke 2010). Since its first release, the Arabidopsis genome has been thoroughly investigated, posing the basis for a deeper understanding of plant development and environmental responses by enabling a better assessment of the structure and functionality of plant genomes (Meinke et al. 1998, The_Arabidopsis_Genome_Initiative 2000, Somerville and Koornneef 2002, Bevan and Walsh 2005). However, deeper analyses of the genome of *A. thaliana* also revealed a high complexity due to several events of whole genome duplications, the occurrence of large-scale duplications and deep reshuffling (Simillion et al. 2002). In particular, these studies showed evidence of at least three rounds of whole genome duplications (Blanc et al. 2000, Vision et al. 2000, Wolfe 2001, Blanc et al. 2003, Blanc and Wolfe 2004, Cui et al. 2006, Jiao et al. 2011, Van de Peer 2011, Jiao et al. 2012). Moreover, the high frequency of gene reduction, i.e. gene loss after each duplication event, diploidization, translocations and inversions (Ku et al. 2000, Gaut 2001), and probable chromosome losses (Conner et al. 1998, Johnston et al. 2005, Lysak et al. 2005), further contributed to reshuffle the retained portions of the genome.

Assuming that gene duplications play a key role in the origin of novel gene functions (Hughes 2005, Flagel and Wendel 2009, Kaessmann 2010, Magadum et al. 2013, Rensing 2014), this issue has often been considered for its relevance

in understanding gene functionality, from their expression (He and Zhang 2005) to the complexity of their regulatory networks (Teichmann and Babu 2004). However, a clear assessment of the duplicated gene content in a genome, accompanied by a reliable description of those genes that are in single copy in a species, is also necessary to support functional and evolutionary analysis (Sangiovanni et al. 2013).

One of the goals, immediately after the release of the Arabidopsis genome, was the definition of all the gene structures and functions of the model plant (Somerville and Dangl 2000). However, the genome of Arabidopsis still contains thousands of protein coding genes with an unknown or incomplete annotation (Frishman 2007, Hanson et al. 2010). Consequently, a poor functional annotation of plant genomes limits the predictive power of comparative genomics analyses.

Although several collections of ortholog genes are today available (O'Brien et al. 2005, Chen et al. 2006, Rouard et al. 2011, Van Bel et al. 2012, Flicek et al. 2013, NCBI_Resource_Coordinators 2013, Waterhouse et al. 2013, Powell et al. 2014), only one reference web accessible database, EPGD (Eukaryotic Paralog Group Database) (Ding et al. 2008), is exclusively dedicated to paralogs in 26 available eukaryotic genomes. Indeed, EPGD is a web resource for integrating and displaying eukaryotic paralog information, in terms of paralog families and intragenome segmental duplications. However, paralogs at intra genome level can be also accessed from some of the collections worldwide available which are related to orthologs, such as Ensembl Compara (Flicek et al. 2013) which include both animal and plant, NCBI Homologene (NCBI_Resource_Coordinators 2013) and Plaza (Proost et al. 2009), which is exclusively dedicated to plant genomes. These databases, however, when showing clusters of paralogs, refer to a list of orthologs or paralogs of a reference gene, without providing an overview of the relationships in the cluster.

In this chapter *pATsi* is described, namely a database in which the entire

collection of protein coding genes of *A. thaliana* is organized in different sets of paralogs and singleton genes identified thanks to a dedicated pipeline described in (Sangiovanni et al. 2013). The paralog genes are here presented in the form of networks of paralogs, accessible also by a graphical approach, with the aim of clearly describing those genes that share direct paralogy relationships in a network. Moreover, gene association by similarity is assigned using two different cutoffs. This allows to provide some more insights on the structural and evolutionary relationships among the genes.

A detailed classification of the genes not classified within the networks of paralogs is also provided in this database, useful to define a reference for similar investigations and to support functional and evolutionary studies on the *A. thaliana* genome.

2.2 Database construction and content

2.2.1 Data source

The entire *Arabidopsis thaliana* genome, intergenic regions and gene family information (TAIR9 release) were downloaded from the TAIR (The Arabidopsis Information Resource) web server (Lamesch et al. 2012). The non-redundant collection obtained from transcription factor families databases (Yilmaz et al. 2011, Zhang et al. 2011) was used to enrich the list of gene families. *A. thaliana* Expressed Sequence Tag (EST) sequences were downloaded from GenBank (release of 8 April 2010).

2.2.2 Collection of paralogs and singletons

In order to identify Arabidopsis paralog and singleton genes, a suitable pipeline was implemented and applied (Sangiovanni et al. 2013). The analysis was based on an all-against-all protein sequence similarity search performed with

BLASTp software (Altschul et al. 1990), using two different cutoffs settings: a more stringent expected value ($E<=10^{-10}$) to select the similarities with greater specifity, and a less stringent one ($E<=10^{-5}$) (Rubin et al. 2000, He and Zhang 2005). Then the two collections were filtered applying the Rost's Formula (Rost 1999, Moreno-Hagelsieb and Latimer 2008), to discriminate significant similarity relationships from less reliable sequence similarities. In this way, 22522 and 21843 structurally related genes organized in two different datasets were obtained, a more stringent dataset (dataset A), with higher similarity levels between the genes, and a less stringent one (dataset B), including more genes, sharing lower similarities respectively. Genes associated by structural similarities in the two datasets were considered as paralogs.

Several filtering steps were also used to identify genes without significant similarity with other protein-coding genes or with any other nucleotide sequence similarity with any region of the entire genome sequence, permitting to collect genes that could be reliably classified as "true singleton". All the genes were therefore classified considering several distinct features based on the pipeline described in (Sangiovanni et al. 2013). All the classes available and the gene numbers associated to each class are summarized in Table 1.

CLASSIFICATION	GENE NUMBER	ANALYSIS
Non-protein coding genes	6070	miscRNAs, tRNAs, rRNAs, ncRNAs, pseudogenes, transposons and unknown genes
Paralogs classified into networks	22522	All-against-all BLASTp $E \leq 10^{-5}$
Unassigned genes due to the Rost's formula	405	Filtering with Rost's formula
Unassigned genes due to the masking filter	213	All-against-all BLASTp $E \leq 10^{-5}$ without masking filter
Unassigned genes due to loose protein similarity	440	All-against-all BLASTp $E \leq 10^{-3}$ of protein coding genes
Unassigned genes due to the ORF annotation error	2	Transcripts BLASTx $E \leq 10^{-5}$ versus proteins for ORF validation
Unassigned genes due to similarities with non-protein coding genes	178	Full genes BLASTn $E \leq 10^{-5}$ versus non-protein coding genes
Unassigned genes due to similarities with intergenic regions	0	Full genes BLASTn $E \leq 10^{-5}$ versus intergenic regions
Singletons not confirmed by ESTs (no EST trace)	24	Transcripts BLASTn (free E-value cutoff) versus ESTs
Singletons not confirmed by ESTs (discarded by e-value cutoff)	688	Filtering of BLASTn results by $E \leq 10^{-5}$
Singletons not confirmed by ESTs (discarded by coverage and identity requirements)	201	Filtering of BLASTn versus EST results by coverage and identity
Singletons not confirmed by ESTs	100	Filtering by Delta >= 20 (EST length >= 20 nt than the transcript)
Singletons confirmed by ESTs	9	0 < Delta < 20 (EST length greater than transcript but less than 20 nt)
Singleton confirmed by ESTs	2387	Delta <=0 (Transcript longer than the EST)

Table 1. A summary of the classes of genes classified in *pATsi*. The analysis performed to obtain genes in each class are also reported (Sangiovanni et al. 2013).

2.2.3 Networks construction

Paralog genes were organized into two different sets of networks, depending on the E-value cut-off. The less stringent cut-off ($E<=10^{-5}$) led to a set of 2754 networks including 22522 paralog genes, while the more stringent cut-off ($E<=10^{-10}$) led to a set of 3017 networks containing 21843 paralogs. Each gene is connected by at least one paralogy relationship, visually represented by an edge, to at least another gene in the same network. The networks have various sizes depending on the gene content, ranging from 2 to 6834 genes, this last number reflecting the maximum number of genes in a network, and corresponding to the biggest network (Fig. 7) obtained with the less stringent threshold ($E<=10^{-5}$).

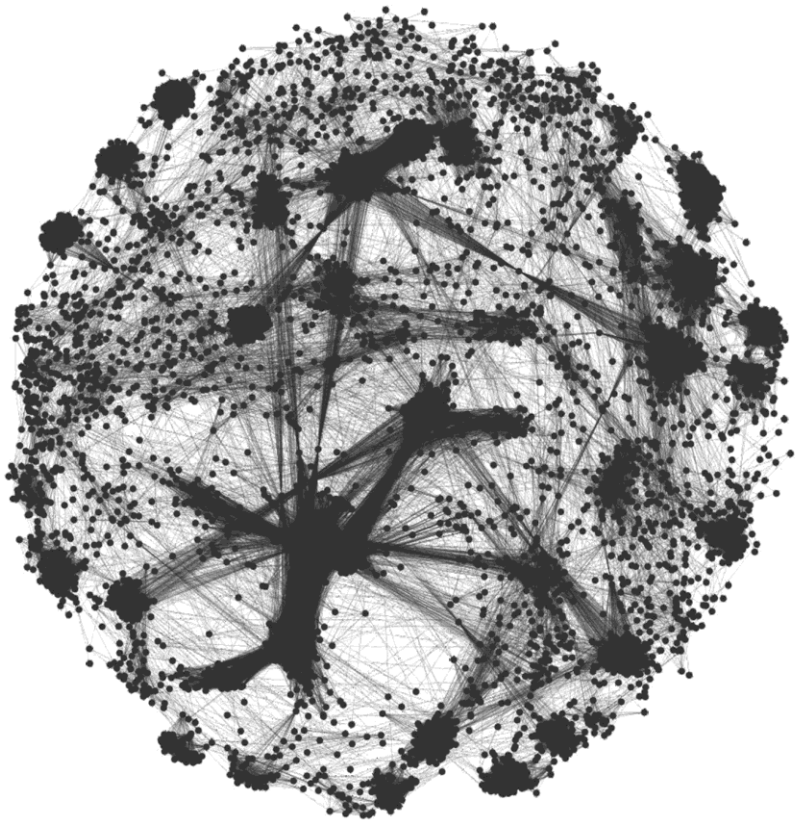


Figure 7. View of the largest network of paralogs, consisting of 6,834 genes. Each dot in dark grey represents a single gene, and each line in light grey represents a paralogy relationship between two genes.

The use of a more stringent threshold defines less relationships between the genes, resulting in a larger number of networks in comparison with the one obtained with a less stringent threshold, and networks that may contain less genes. The reason for this behavior is that a less stringent cut off (dataset A) includes among paralogs also genes that are excluded by the more stringent one (dataset B). To keep trace of the relationships between network organization at different cutoffs, the network naming is assigned as follows. The networks at the less stringent threshold were named as NETxGy_z. The letter x indicates a number assigned when sorting the total amount of networks by decreasing network size, y indicates the network size (i.e. the number of included genes) and z is the number of networks or singletons in which the network is split when the more stringent cutoff is applied.

Results from the two cutoffs, both considered significant for similar approaches (Rubin et al. 2000, He and Zhang 2005), are here provided as they can be useful for an approximate estimation of conserved or variable network organizations at these settings.

2.2.4 Database development

The relational database described in this chapter was designed using MySql v.5.5.31 and InnoDb storage engine. In order to improve the efficiency and to decrease the execution time of the queries, all tables are indexed using normal BTree indexing based on individual and multiple index keys.

2.3 Database usage

pATsi can be accessed at http://cab.unina.it/athparalog/main.html (Fig. 8-A).

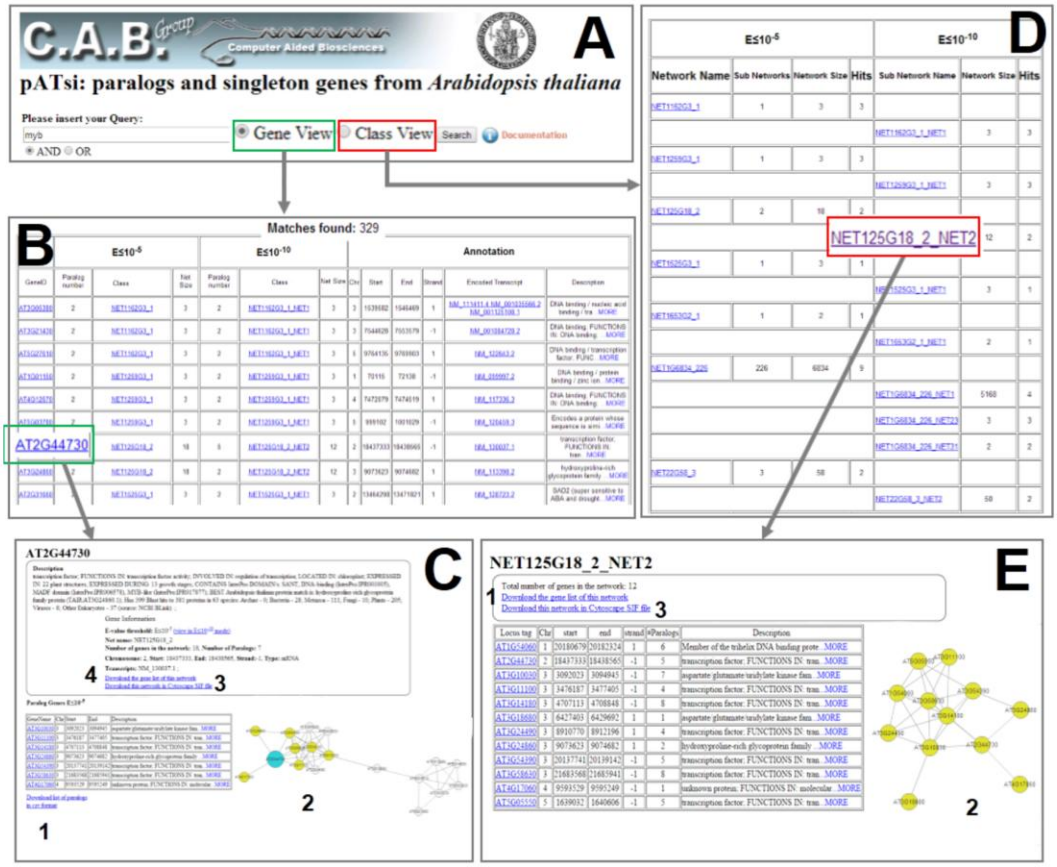


Figure 8. Possible queries workflow in *pATsi* web interface. (**A**) Main page of the *pATsi* database browser; for each query, the user can switch from the gene view (bordered in green) to the class view (bordered in red). (**B**) List of genes associated with a query. (**C**) Gene information page. In (C2) a network graph is shown; each circle is a GeneID, with the light blue-circled one representing the selected GeneID and the yellow-circled one(s) representing the paralog(s) of the selected GeneID; gray lines represent paralogies between genes. (**D**) List of networks associated with a query. (**E**) Network information page. Image extracted from (Ambrosino et al. 2016).

All *Arabidopsis thaliana* genes and networks can be browsed or searched using key words. The query for searching specific key words in *pATsi* can accept a gene locus ID, a network name, or every string to search the annotation content. Two different views are provided: the gene view and the class view.

2.3.1 The gene view

The gene view (Fig. 8-A in green) results in the list of loci associated to the query (Fig. 8-B). Each row of the list contains the following information:

* *Gene ID* represents the official TAIR classification for *A. thaliana*. Clicking on the Gene ID it is possible to browse the Gene ID related page, containing information about the gene investigated and the two networks organization in which the gene may be included. In case the gene is a singleton, no network organization is shown.
* *Paralog number* shows the number of paralogs of the gene at the two different cut-offs, or zero in case of singletons or non-mRNA genes.
* *Class:* if the gene has paralogs at one of the e-value cutoffs, the network name is shown; for genes without paralogs, the name of the class is reported. In case of unassigned genes, the classification field contains a brief explanation of the reason that led to that specific category.
* *Net size:* this field shows the number of all paralog genes contained into the network, zero if the corresponding gene is a *singleton* or a non-mRNA gene.
* *Chr:* the chromosome on which the locus maps.
* *Start/End:* starting/ending position of the locus on the chromosome.
* *Strand:* direction of the locus transcription.
* *Encoded transcript:* the RefSeq or the encoded transcript. Each RefSeq has a link to GenBank.
* *Description:* the TAIR functional annotation (Lamesch et al. 2012) for each of the RefSeq.

By clicking on the GeneID in the results table, the user will be redirected to a new page (Fig. 8-C). In the topmost part of the page, the GeneID, several details about the gene annotation and possible network information are reported. In the bottom left part of the page, the list of paralogs of the selected gene is shown. Each gene is also crosslinked to its specific description in the database. The list

of paralogs can be downloaded (Fig. 8-C1). In the bottom center part of the page, an interactive network graph is displayed (Fig. 8-C2) using CytoscapeWeb, a web-based network visualization tool (Lopes et al. 2010). Users are enabled to interact with the displayed network by selecting nodes and edges, and modeling the network view accordingly. For each network, it is possible to download a file (Fig. 8-C3) which can be easily imported into Cytoscape, for onsite visualization or for managing more complex networks (Shannon et al. 2003). The list of genes that are included in the displayed network can be downloaded too (Fig. 8-C4).

2.3.2 The class view

Selecting the class view (Fig. 8-A in red), the query process organizes the genes associated to the query into classes, separating the genes not included in the network from those included in networks. The resulting page also provides the list of networks associated to the query (Fig. 8-D) indicating in each row of the list of networks the following information:

- *Network:* the name assigned to the network at the lowest cutoff ($E \leq 10^{-5}$). Clicking on the network name it is possible to browse the Network ID related page, containing information about the network investigated and the genes included in it. For each network, a list of one or more sub networks is also shown.
- *Sub networks* shows the number of sub networks or singletons in which the network is split when the cutoff of $E \leq 10^{-10}$ is applied.
- *Network Size:* i.e. the number of genes included in the network.
- *Hits:* the number of matching genes with the user query.

By clicking on the network name in the resulting table, the user will be redirected to a new page (Fig. 8-E). In the topmost part of the page, the network

name and the number of genes are shown. In the left part of the page, the list of genes of the selected network is reported. It is also possible to download the list of genes (Fig. 8-E1). In the right part of the page, the network graph is displayed (Fig. 8-E2). The file of the network in .sif format can be downloaded too (Fig. 8-E3).

As mentioned above, networks here presented are classified according to two different thresholds. The use of a more stringent threshold defines a lower number of paralogy relationships between genes, hence obtaining a larger number of networks in comparison with the ones obtained with the less stringent threshold. This is explained considering the effects of the less stringent cut off, that permits to include in a network also genes that are otherwise excluded when the more stringent threshold is used (Fig. 9).

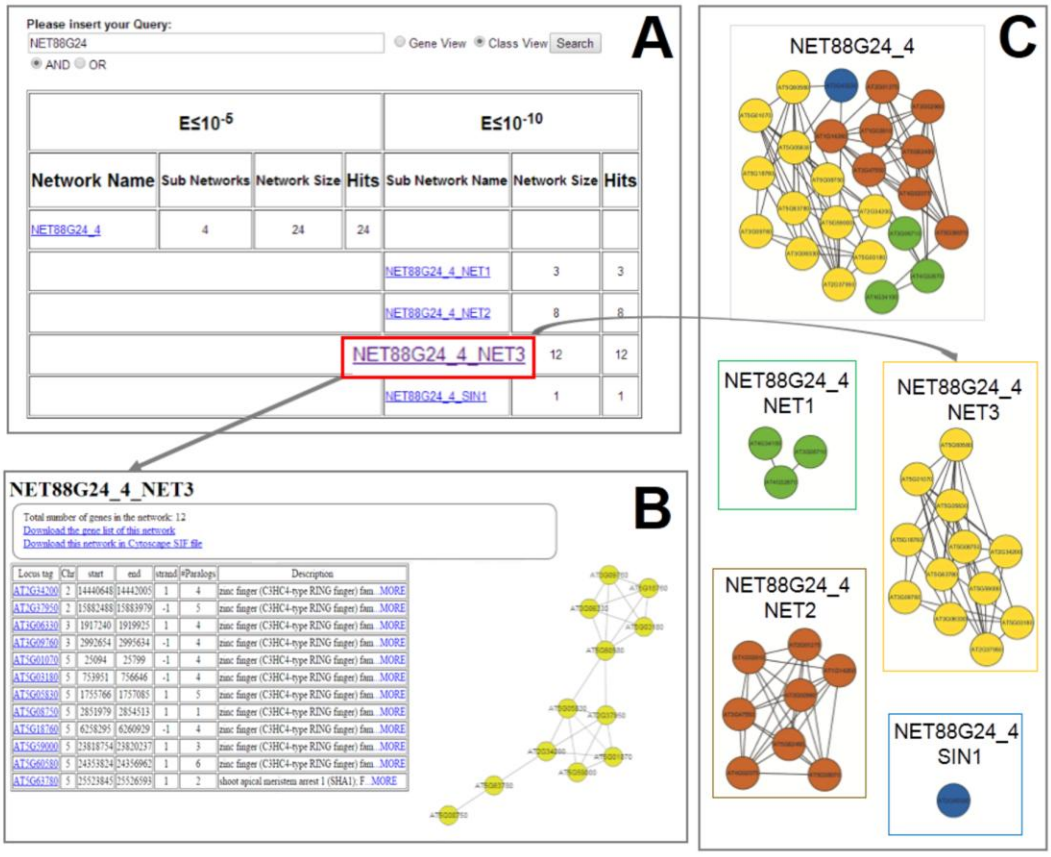


Figure 9. Network organization. (**A**) List of subnetworks associated with a network query. (**B**) Network information page. (**C**) Graphic representation of a network of 24 genes (NET88G24_4) splitted into three subnetworks (NET88G24_4 NET1-NET2-NET3) and one singleton (NET88G24_4 SIN1). Image extracted from (Ambrosino et al. 2016).

2.3.3 A case study: the acetyltransferases family

In order to test *pATsi*, we queried the database with the "serine acetyltransferase" keyword, obtaining six matches in the gene view mode, i.e. one non protein-coding gene and five protein-coding gene grouped in one network, NET253G11_2, together with other six genes (Fig. 10).

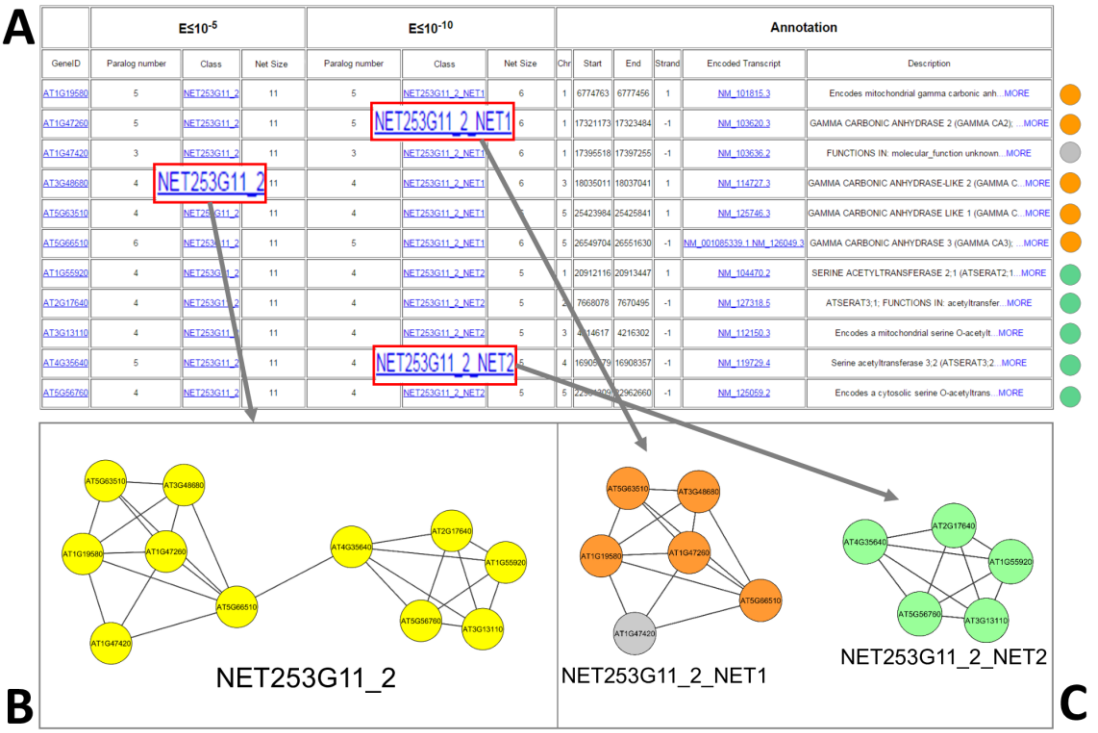


Figure 10. Example query. (**A**) List of genes associated with the NET253G11_2 network. (**B**) Graph view of NET253G11_2_NET1 and NET253G11_2_NET2 subnetworks. Orange circles represent genes annotated as gamma carbonic anhydrases; yellow circles represent genes annotated as serine acetyltransferases; gray circle represents a gene with an unknown function. Image extracted from (Ambrosino et al. 2016).

The family of serine acetyltransferases catalyze the limiting reaction in cysteine biosynthesis (Nguyen et al. 2012, Yi et al. 2013, Tavares et al. 2015). These enzymes are of great interest to the scientific community, because of their active role in creating nutritionally essential sulphur amino acids, which largely contribute to a healthier diet for humans and animals (Tabe et al. 2010). Analyzing the identified network (NET253G11_2), we noticed that, switching to a more stringent cut-off ($E \leq 10^{-10}$) two splitted sub networks were obtained. The first one (NET253G11_2_NET1) is formed exactly by the five serine acetyltransferase enzymes previously detected, and the other (NET253G11_2_NET2) is formed by five genes annotated as gamma carbonic anhydrases and one genes with an unknown function (Fig.10-C). The function

of the unknown gene can be inferred from its paralogs in the same sub network, namely the gamma carbonic anhydrase enzymes. Moreover, serine acetyltransferases and carbonic anhydrases belong to the trimeric LpxA-like superfamily, a set of enzymes with trimeric repeats of hexapeptide motifs. Therefore, setting a less stringent threshold, different enzymes were grouped within the same network based on their common origin from LpxA-like superfamily. Setting a more stringent threshold, instead, allowed the discrimination of different enzymatic families into different networks. Therefore, the use of such networks helps the investigation of gene family organization and splitting when different thresholds are applied (Sangiovanni et al. 2013).

2.4 Conclusions

The collection here described is useful for an efficient exploitation of the Arabidopsis gene content, contributing to the identification of structurally related genes, to their functional assignment, and to the classification of singleton genes within the genome (Sangiovanni et al. 2013). Additionally, *pATsi* provides a novel approach to classify protein-coding genes from *A. thaliana*, based on similarity defined both at intragene level or comparing with all the rest of the genome regions, focusing on paralogs to build gene networks and singleton genes for an appropriate classification.

There are several collections today available for paralog gene classification in plants (Kinsella et al. 2011, Van Bel et al. 2012) also including data from *A. thaliana* genes. However, the different collections of paralog genes from these species are not easily comparable and no one represent a reference for related works. Moreover, some studies also aimed to identify singleton genes from *A. thaliana* (Duarte et al. 2010), but no dataset is provided for related efforts. Nevertheless, no reference collection is today available to fully access both the genes having at least one paralog and being singletons in *A. thaliana*.

One of the novelty of this database is to provide immediate access to gene classes and possible relationships, and therefore it represents a resource for gene family analysis, comparative genomics, and to support the unraveling of the complexity of the Arabidopsis genome. Moreover, this piece of information is based on a fully reproducible methodology (Sangiovanni et al. 2013), with the aim to provide a common reference and common frameworks in associated efforts. This is essential also to trace and compare different scientific results.

This database provides a permanent resource for studies that need a reference collection for gene family classification and comparative analysis. Please refer to the recently published article for details on this work (Ambrosino et al. 2016).

Chapter 3. *Transcriptologs*, a transcriptome-based approach to predict orthologs

3.1 Introduction

The detection of ortholog genes is a relevant issue in many research areas. Indeed, finding orthologs between species is useful for structure, functional and evolutionary inferences (Dessimoz et al. 2005, O'Brien et al. 2005, Chen et al. 2006, Hulsen et al. 2006, Kuzniar et al. 2008, Proost et al. 2009, Altenhoff et al. 2011, Rouard et al. 2011, Dessimoz et al. 2012, Flicek et al. 2013, Waterhouse et al. 2013, Huerta-Cepas et al. 2014, Powell et al. 2014, Schreiber et al. 2014). Ortholog genes, i.e. two gene copies in two species derived from a common ancestor that diverged after a speciation event, are usually investigated for a wide range of applications in comparative genomics, phylogenetic analysis, function prediction and annotations of newly sequenced genomes (Altenhoff and Dessimoz 2009, Dessimoz et al. 2012). In particular, it is common to preliminarily exploit orthology relationships for transferring functional information from genes in well-defined model organisms to still uncharacterized genes in newly sequenced genomes (Sonnhammer and Koonin 2002, Koonin 2005, Dolinski and Botstein 2007) paving the way to understand speciation and gain loss of gene functionalities, highlighting peculiarities or conservation among species.

The increasing number of fully sequenced genomes further pushed the flourishing of computational methods to detect orthologs among species (Tatusov et al. 1997, Koonin 2005, Alexeyenko et al. 2006, Chen et al. 2006, Gabaldon 2008, Altenhoff and Dessimoz 2009, Altenhoff et al. 2011, Kristensen et al. 2011, Flicek et al. 2013, Waterhouse et al. 2013, Huerta-Cepas et al. 2014, Powell et al. 2014, Schreiber et al. 2014). Currently, most of the

approaches for inferring orthology can be grouped mainly into graph-based methods, which define orthologs based on sequence similarity, and tree-based methods, which classify all the splits of a given gene tree as duplication or speciation, according to the phylogeny of the analyzed species (Gabaldon 2008, Kuzniar et al. 2008, Kristensen et al. 2011). Graph-based methods include two steps. i) pairs of ortholog genes are detected, and as a consequence, graphs with nodes representing genes and edges representing relationships are defined; ii) clusters of ortholog genes are defined based on the structure of the graphs. The simplest graph-construction approach identifies orthologs between genes in pair of genomes (Altenhoff and Dessimoz 2012). The key assumption is that the ortholog genes are those among homolog genes with the minimum divergence or the maximum similarity. Therefore, estimating the evolutionary relationships by sequence similarity measures, this basic approach consists in the detection of all the genes in two different genomes that are reciprocally the best hit of each other (Tatusov et al. 1997, Huynen and Bork 1998, Hughes 2005), i.e. those with the highest similarity or the minimum distance, according to the measure established. This widespread approach is generally defined as the search for the Bidirectional Best Hits (BBH), and it establishes that genes x_i and y_i, from species X and Y, are the best putative orthologs if x_i is the best hit of y_i, and y_i is the best hit of x_i, in all versus all similarity searches (Overbeek et al. 1999). The detection of BBHs between genes from two genomes is computationally efficient because each gene collection can be scanned independently, and sequence alignments can be computed by efficient approaches, based on dynamic programming (Smith and Waterman 1981) or heuristic algorithms, such as the BLAST set of programs (Moreno-Hagelsieb and Latimer 2008). However, the BBHs detection process has some limits. Primarily, some genes in a species can have more than one ortholog in another species. This happens whenever a gene is duplicated after a speciation event while the ortholog counterpart in the other genome remains in single copy, namely a singleton gene (Sangiovanni et al. 2013). Remm et al. (Remm et al.

2001) refer to these duplicated genes after a speciation event as in-paralogs, developing a dedicated algorithm for their detection called Inparanoid. A different approach for detecting the in-paralogs consists in the implementation of a score tolerance threshold or a confidence range around the BBHs to expand the notion of the best hit into groups of best hits, in order to identify one-to-many or many-to-many orthologs (Dessimoz et al. 2005, Fulton et al. 2006).

Graph-based methods can work with only two-species at a time, and in particular these algorithms are not effective for large evolutionary distances (Huynen and Bork 1998), since low sequence similarities may not be detected at all. On the other hand, tree-based methods can work on more species and provide more information than pairs or groups of orthologs, like evolutionary distances, the order of duplication and speciation events. However, these methods are computationally much more expensive than graph-based algorithms (Kristensen et al. 2011, Altenhoff and Dessimoz 2012). Moreover, especially when phylogenetic trees include large numbers of genes and genomes, they may also be less reliable, in particular when large evolutionary distances occurs. BBH detecting algorithms, instead, are much more faster and easy to automate when based on heuristic approaches (Kristensen et al. 2011). Since there isn't a widely accepted standard set of orthologs, a statistical approach was carried out to compare several methods for ortholog detection (Hulsen et al. 2006, Chen et al. 2007, Gabaldon 2008, Kuzniar et al. 2008, Altenhoff and Dessimoz 2009). By these measures, no single method achieved optimal performance. Overall, many BBH algorithms reach high sensitivity at the cost of specificity, while the tree-based methods showed the opposite trend. At short-evolutionary distances, instead, graph-based methods and tree-based methods produce similar sets of orthologs (Kristensen et al. 2011). A recent study (Altenhoff and Dessimoz 2009), however, showed that sometimes more complex tree reconstruction/reconciliation approaches are outperformed by pairwise comparison approaches like BBH. This suggests that tree reconciliation, although it is more powerful in theory, is not rigorously the best

method in practice. This probably explains why many people prefer to use simple BBH implementations rather than a more complex orthology method (Altenhoff and Dessimoz 2009, Kristensen et al. 2011).

Despite the similarity search able to identify orthology or paralogy relationships is generally based on a comparison of protein sequences, this type of analysis can lead to errors due to the lack of a correct and exhaustive definition of such sequences in recently sequenced species with still a preliminary annotation. In Trachana et al. (Trachana et al. 2011), genome annotation emerged as the largest single influencer of the quality of orthology detection procedures, affecting up to 30% of the performance of these methods. Therefore, trying to overcome the limitations due to the quality of protein sequences predictions, which are typical in recently sequenced genomes but still affect also stable and more established annotated ones, we developed a method for the detection of orthologs that uses transcriptomic references instead of proteomic ones. Moreover, the proposed approach allows to exploit the information content of a nucleotide sequence that is three times higher than the corresponding protein code.

3.2 Material and methods

3.2.1 Data sets

Trascriptome and proteome collections for *Arabidopsis thaliana* (release TAIR 10) (The_Arabidopsis_Genome_Initiative 2000) and *Sorghum bicolor* (release JGI 1.4) (Paterson et al. 2009) were downloaded from the TAIR (The_Arabidopsis_Information_Resource) and from the JGI genome source websites (Joint_Genome_Institute), respectively. Moreover, we downloaded the ortholog collections between *A. thaliana* and *S. bicolor* publicly available in the Ensemble Plant BioMart (Flicek et al. 2013) and PLAZA (Proost et al. 2009) dedicated resources.

3.2.2 Similarity detection

An all-against-all sequence similarity search of the two protein and mRNA collections was performed using the BLASTp and tBLASTx programs of the BLAST package (Camacho et al. 2009), respectively. Parameters fixed for the comparisons are Expect-value (E-value, E) cut-off at 10^{-3} and max_target_seqs at 100. Moreover, we performed also an all-against-all sequence similarity search using the BLASTp program, setting a less stringent Expect-value (E-value, E) cut-off of 1000 to validate and compare the results from reference ortholog databases.

3.2.3 Algorithm description

In order to identify BBHs and expanded BBHs (eBBH) based on transcript collections, we developed *Transcriptologs*, a dedicated method consisting of two procedures, namely *alignment_reconstruction* (Fig. 11) and *BBH* (Fig. 12), implemented by the Python programming language (v3.3.3).

Algorithm 1 alignment_reconstruction

```
1:  procedure alignment_recostruction (Species1_vs_Species2.txt)
2:      for all Species1 query x_i in Species1-Species2 alignments a_l do
3:          for all Species2 subject y_j do
4:              store alignment fragments f_m corresponding to best score alignment
5:              store strand s_h corresponding to best score alignment
6:              for all other alignment fragments f_n do
7:                  store strand s_k corresponding to f_n strand
8:                  if strand s_h = s_k and alignment fragment f_n do not overlap f_m do
9:                      store alignment fragments f_n
10:                 end if
11:             end for
12:             compute new alignment a_g with stored alignment fragments f_m, f_n
13:                 new_alignments_1_vs_2[x_i ][y_j] ← a_g
14:         end for
15:     end for
16:     return new_alignments_1_vs_2
17: end procedure
```

Figure 11. Pseudo code of the alignment recostruction algorithm we developed.

Algorithm 2 BBH

```
1:  procedure BBH (new_alignments_1_vs_2, new_alignments_2_vs_1)
2:      for all Species1 query x_i in new_alignments_1_vs_2 a_f do
3:          compute Species2 subject y_j with the best score hit
4:          compute Species2 subject y_k with the score hit in a range around the best score
5:      end for
6:      for all Species2 query y_n in new_alignments_2_vs_1 a_g do
7:          compute Arabidopsis subject x_m with the best score hit
8:          compute Arabidopsis subject x_o with the score hit in a range around the best score
9:      end for
10:     for all best score hits x_i , y_j do
11:         for all best score hits x_m , y_n do
12:             if x_i=x_m and y_j=y_n do
13:                 BBH[x_i ][y_j] ← x_i , y_j
14:             end if
15:         end for
16:     end for
17:     for all score hits x_i , y_k do
18:         for all best score hits x_o, y_n do
19:             if x_i=x_o and y_k=y_n do
20:                 eBBH[x_i ][y_k] ← x_i y_k
21:             end if
20:         end for
21:     end for
22:     return BBH, eBBH
23: end procedure
```

Figure 12. Pseudo code of BBH algorithm we developed.

The method considers the two resulting files from the reciprocal t-BLASTx transcript similarity searches (e.g. Species1_vs_Species2.txt and Species2_vs_Species1.txt). The tBLASTx results may include possible different alignments between a query sequence x_i and a subject sequence y_j from the set of sequences X and Y of the two species under comparison, each

alignment defined by different fragments f_m all belonging to the same frame. In order to define more extended alignments, we designed a dedicated procedure (alignment_reconstruction, Fig. 11) that selects the alignment fragments corresponding to the best scored alignment, and then adds other fragments from alignments from different reading frames on the same strand s_h, if present. The fragments are added exclusively if they do not overlap regions already considered in the procedure of alignment reconstruction (Fig. 13).

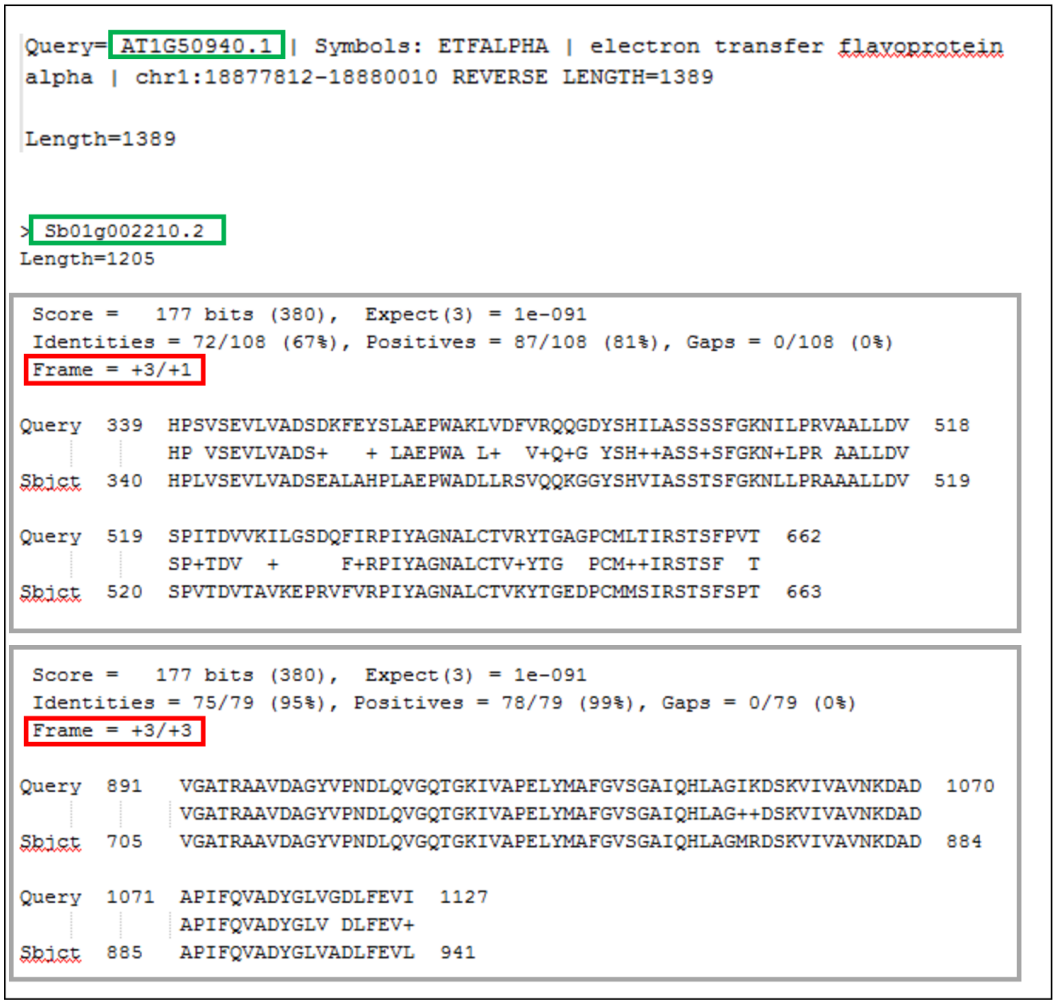


Figure 13. Improvement example of the total alignment length. If we have to align two sequences AT1G50940.1 and Sb01g002210.2 (highlighted in green), the tBLASTx program provides different alignment fragments (highlighted in grey), each one corresponding to a given reading frame (highlighted in red) of the two sequences. In this example the algorithm we designed is able to rebuild an entire alignment using an alignment fragment with a reading frame of +3/+1 and an alignment fragment with a reading frame of +2/+3, since they do not share overlapping segments of the aligned sequences.

The score of the extended final alignment is defined as the sum of the scores of the single alignment fragments added during the reconstruction.

When selecting reciprocal hits, we also implemented the possibility to set a tolerance threshold around the score associated to the BBH to define eBBHs. This permits to define other sequences y_k which are similar, in an established range, to the query sequence x_i. Therefore, the method can detect the best hit that is bidirectional, but also other bidirectional hits with score in preferred ranges from the best one (Fig. 12).

3.3 Results and discussion

3.3.1 Comparison of reference databases

In order to measure the stringency level of orthology relationships based on our sequence similarity searches compared to the ones that are available in public collections of orthologs, we performed an initial all-against-all homology search between *Arabidopsis thaliana* and *Sorghum bicolor* setting a very high E-value cut-off at 1000. We compared the homology relationships detected by the BLASTp program with the orthologs collections available in Ensembl Plant BioMart (Flicek et al. 2013) and in PLAZA (Proost et al. 2009). The results are summarized in Figure 14.

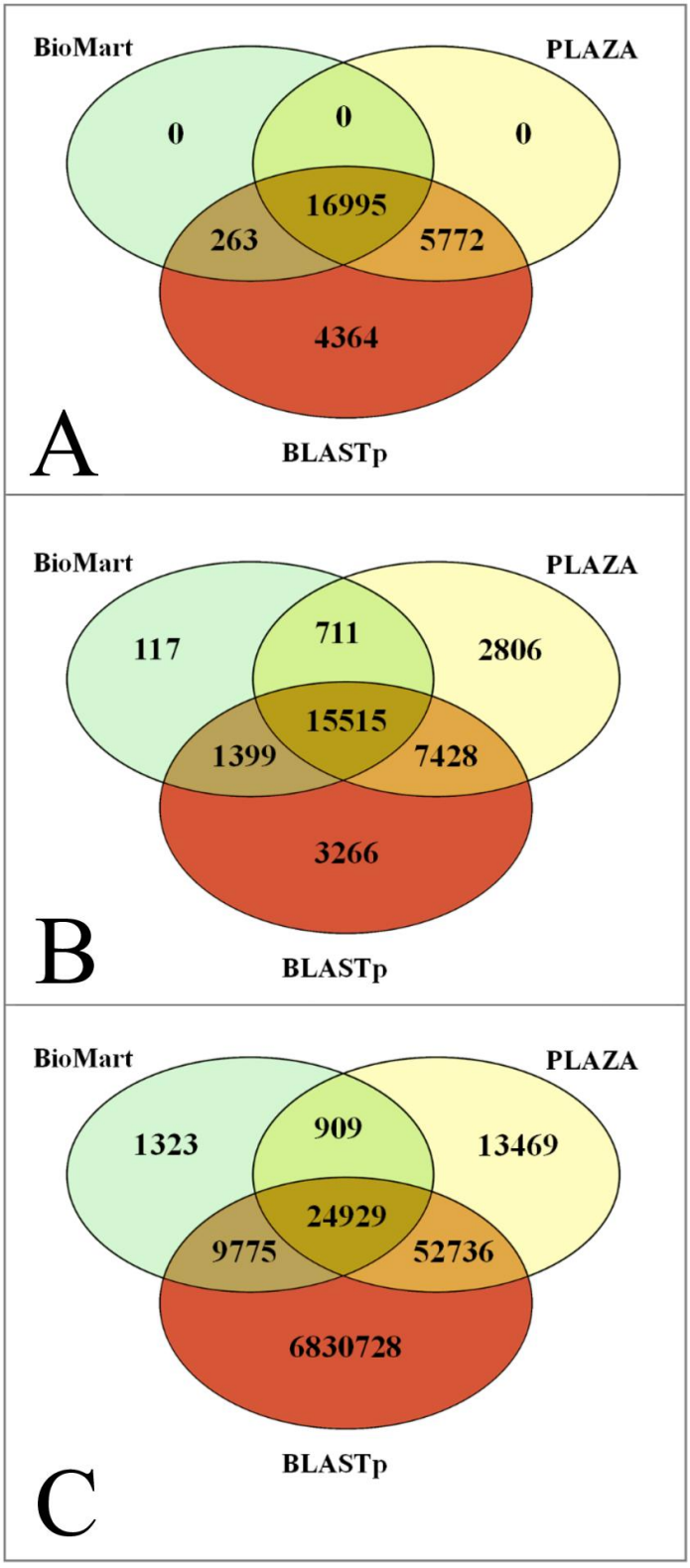


Figure 14. Comparison of results detected by BioMart, PLAZA and an in house BLASTp analysis. A) Venn diagram showing the number of Arabidopsis genes that have a relationship with a Sorghum counterpart. **B)** Venn diagram showing the number of Sorghum genes that have a relationship with an Arabidopsis gene. **C)** Venn diagram showing the number of relationships between Arabidopsis and Sorghum.

Considering the Arabidopsis genes that have a homolog counterpart in Sorghum, the BLASTp analysis includes all the genes detected also by BioMart and PLAZA (Fig. 14-A). If we consider, instead, the Sorghum genes that have a homolog counterpart in Arabidopsis (Fig. 14-B), although there is a significant number of genes that are in common among the three collections (16995), there are 117 genes detected only by BioMart, 2806 genes detected only by PLAZA and 909 genes detected by both. We obtained a similar behavior also when considering directly the homology relationships (Fig. 14-C); in fact there are 1323 relationships detected only by BioMart, 13469 detected only by PLAZA and 909 detected by both. Moreover, due to the high E-value used in our preliminary analysis, it comes out a huge number of relationships detected only by BLASTp (6830728, Fig. 14-C). If we filter out from these only highly significant matches with an E-value cut-off of 10^{-100}, again we obtain a very large number (65996 relationships).We decided to set a looser E-value cut-off in order to include in our analysis as many relationships, available in other collections, as possible. Interestingly we could not include all the similarity relationships available in other collections of orthologs, even by using in our analysis a less stringent and not so reliable threshold. Moreover, we noticed the presence of a huge amount of homology relationships associated to high E-values that we were not able to find in other public collections of orthologs. We can conclude that orthologs collections available in open access databases are quite heterogeneous between them, and often the provided quality standards of the detected orthologs are not so high.

3.3.2 Orthology inference

Transcriptologs results were compared to protein-based sequence similarity searches performing all-against-all independent analyses. Protein sequences (BLASTp) and translated mRNAs (tBLASTx) sequences were both analyzed setting an E-value cutoff at 10^{-3}.

We considered translated nucleotide since the protein similarity scoring is more sensitive than the nucleotide based one. Moreover, the results could be appropriately compared with results from classical protein based approaches. In addition, this approach would also assess similarity between two sequences in presence of frame shifts due to sequencing errors, annotation limits or true evolutionary divergence.

For each detected pair of query-subject hit, the tBLASTx provides a list of alignment fragments grouped by frame, corresponding to different alignments with an associated score. The alignment reconstruction algorithm (Fig. 11) attempts to reconstruct the most extended alignment between the two mRNAs. Indeed, the algorithm collects all the fragments with the same reading frame originated from the BLAST best score alignment. Then, it adds fragments coming from different reading frames, as long as they are on the same strand and they do not overlap the already collected ones. The new alignments and their scores, defined by the sum of the scores of the contributing fragments, are the final result of the alignment reconstruction algorithm.

In the example test we considered, 82721 tBLASTx resulting alignments out of 1181628 total matches (Table 2) were reconstructed adding at least one alignment fragment among those included in the tBLASTx original output. The improved algorithm led to an increase in: a) the average score values of about 54 units compared to the original tBLASTx output; b) the average number of alignment fragments forming the final complete alignment; c) the average number of identity matches; d) the average alignment length (Table 1).

	NORMAL ALGORITHM	MODIFIED ALGORITHM	Δ (modified algorithm – normal)
Score	200.02	254.73	+54.71
N. of Fragments	2.60	4.00	+1.40
Identity	112.05	142.83	+30.78
Alignment Length	167.77	217.36	+49.59

Table 2. Comparison of results from tBLASTx and *Transcriptologs*. Mean values of the score, number of fragments, number of identities matches and alignment length, related to the alignments that were refined by our implementation, are shown.

Subsequently, BBHs between *A. thaliana* and *S. bicolor* were detected using results from the protein and the transcript based reciprocal BLAST results, respectively. In details, 11284 BBHs were detected by using protein sequences, while 11235 BBHs were detected by using mRNA sequences, with 8674 common results (Fig. 15-B).

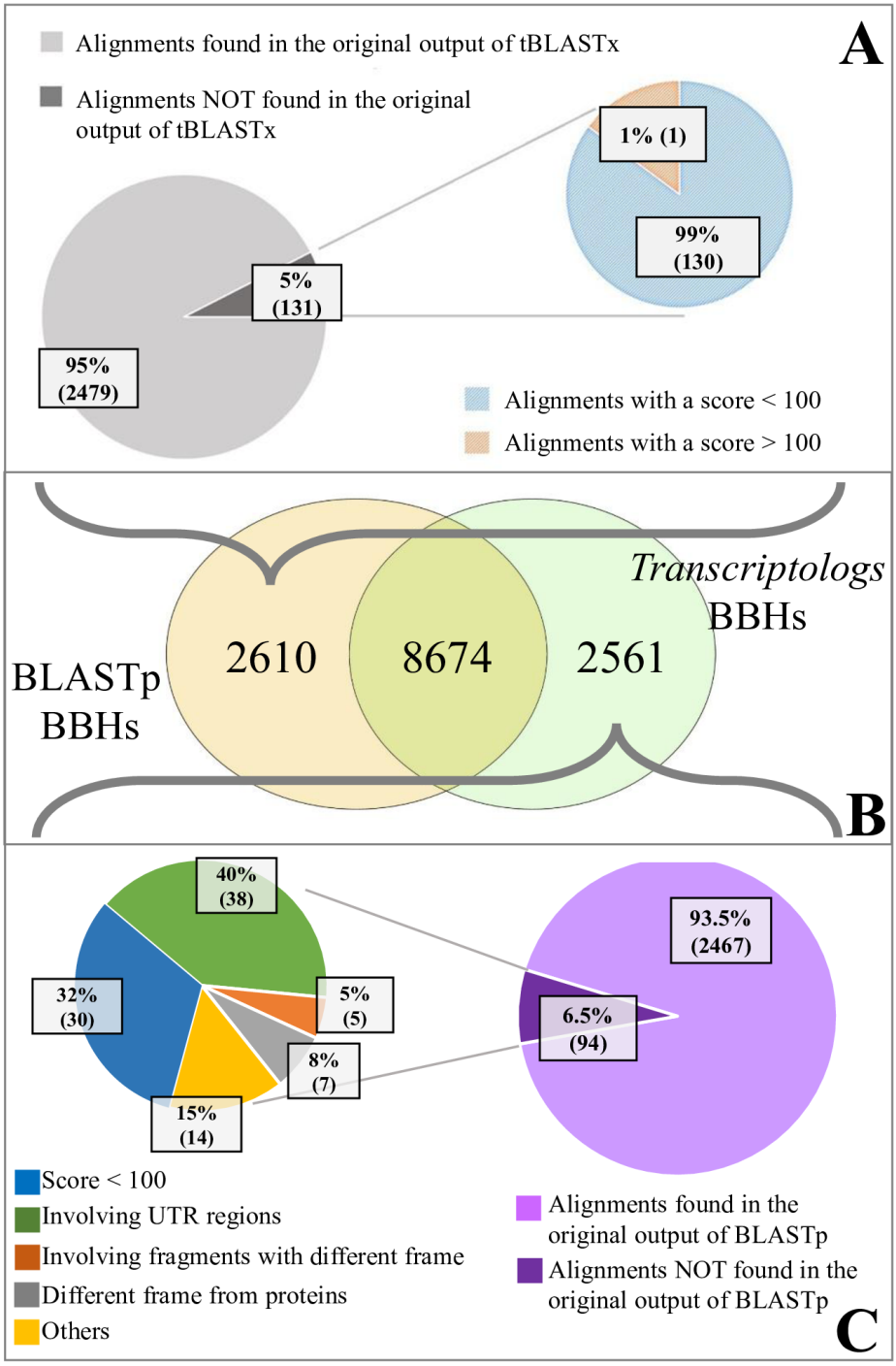


Figure 15. Comparison between *Transcriptologs* and BLASTp analyses. A) Pie charts showing some features of BBHs detected only by using protein sequences. **B)** Venn diagram showing differences and similarities in the number of BBHs detected using protein sequences and transcript sequences. **C)** Pie charts showing some features of BBHs detected only by using transcript sequences. In the pie chart on the left: the number of alignments that involve UTR regions is shown in green; the number of alignments obtained from at least two fragments having different reading frame between them is shown in orange; the number of alignments with a different reading frame in comparison to the predicted proteins is shown in gray; the number of alignments with a similarity score less than 100 is shown in blue; the remaining number of alignments is shown in yellow.

Moreover, 2610 BBHs were exclusively detected by the protein based analysis, while 2561 BBHs were exclusively from transcript sequences (Fig. 15-B). Figure 16 shows the distribution of the scores and E-values of the alignments of these two specific BBHs datasets. We evaluated the quality of the resulting alignments by considering the score and the E-value of each alignment. Since the score is a numerical value used to assess the biological relevance of a finding, while the E-value associated to a score express the probability to obtain by chance that score, the lower the E-value the more the alignment is significant. Figure 10-A/C shows that the scores of tBLASTx BBHs, though generally comparable with those of BLASTp BBHs, reached higher figures (in the upper right of fig. 16-A). A similar behavior was confirmed by the E-values distribution (Fig. 16-B/D), where the number of less significant E-values of some of the BLASTp BBHs was larger (Fig. 16-B/D).

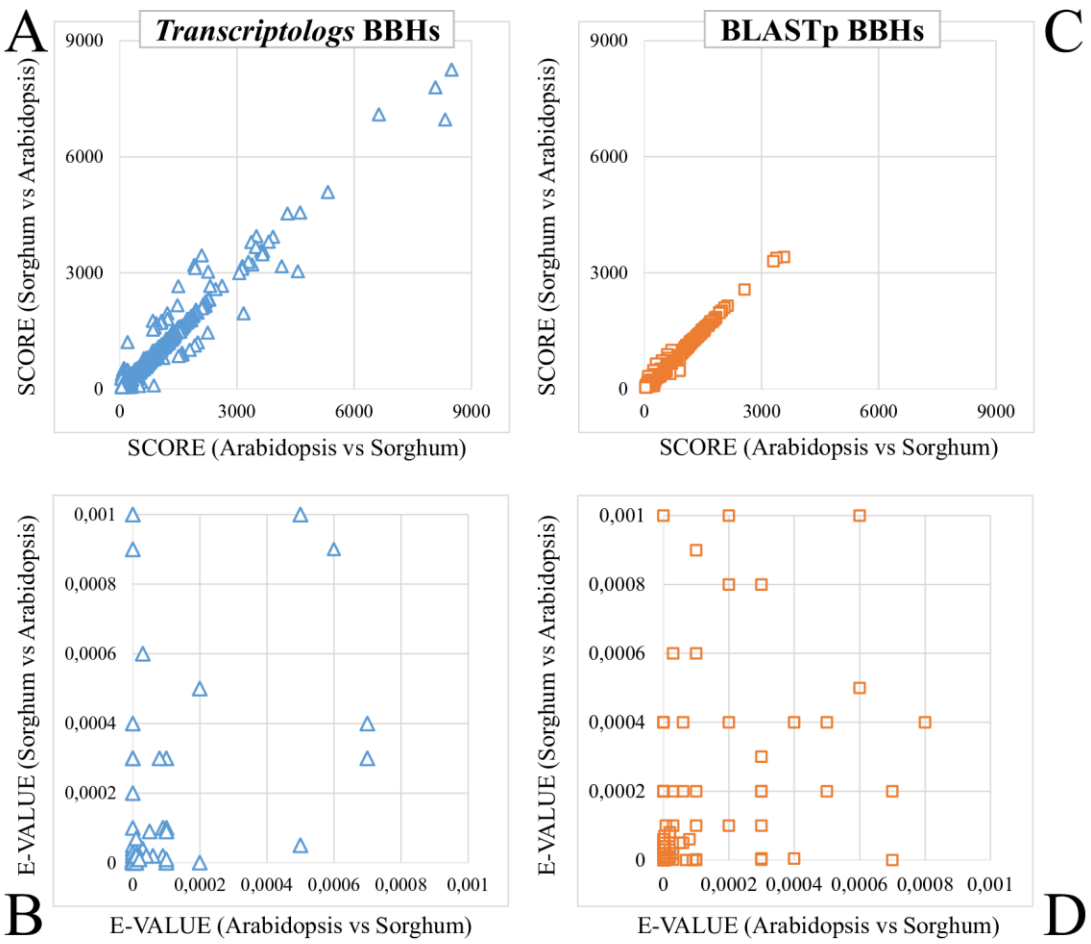


Figure 16. Comparison between *Transcriptologs* and protein BBHs. A) Distribution of the BBH scores detected only by using transcript sequences. **B)** Distribution of the BBH E-values detected only by using transcript sequences. **C)** Distribution of the BBH scores detected only by using protein sequences. **D)** Distribution of the BBH E-values detected only by using protein sequences.

Then, among the BBHs exclusively detected by the BLASTp (2610 matches) and by the tBLASTx (2561 matches) methods, we considered the cases in which the same Arabidopsis gene found a different Sorghum ortholog when considering the transcript based comparison or the protein based comparison (Fig. 17-A), and vice versa (Fig. 17-B).

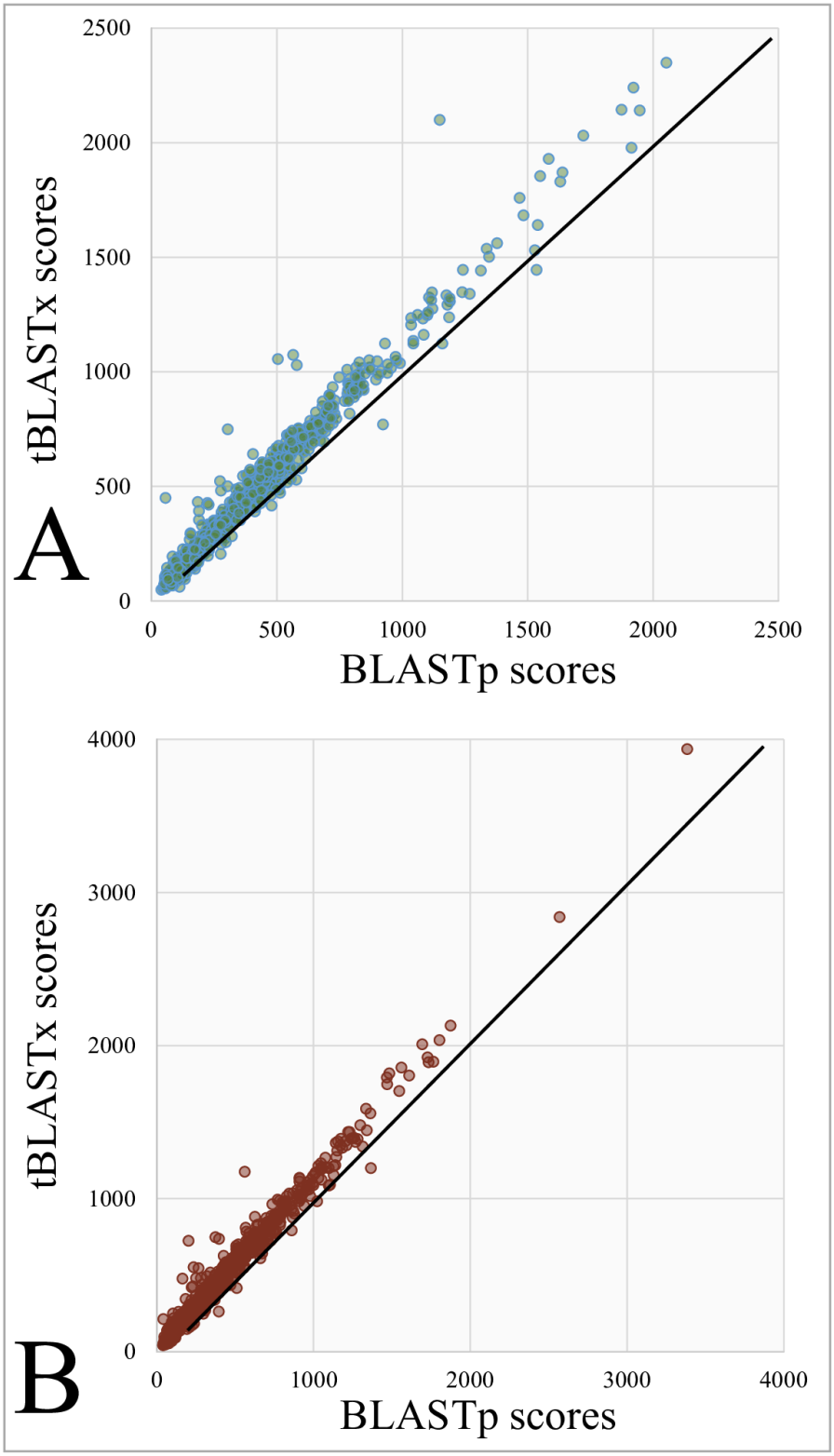


Figure 17. Comparison between *Transcriptologs* and protein BBHs. A) Distribution of the BBH scores detected exclusively by using transcript and protein sequences, involving the same *Arabidopsis thaliana* gene. **B)** Distribution of the BBH scores detected exclusively by using transcript and protein sequences, involving the same *Sorghum bicolor* gene.

Plotting the score distribution of the corresponding alignments based on the two different approaches, we noticed higher scores of the similarities detected by tBLASTx considering the two species (Fig. 17). This highlighted that transcript based comparisons detect alignments with a higher score when compared to the ones obtained from protein sequences, finding more appropriate associations.

In details, most of the protein BBHs (2479 on 2610) were detected also by the tBLASTx analysis before the selection of the Bidirectional Best Hits (Fig. 15-A/B). Indeed, they just were not selected among the best reciprocal hits. However, 131 relationships within these BBHs exclusively detected starting from protein sequences were not found at all by the similarity search based on transcript sequences (Fig. 15-A). Interestingly, only in 1 case of these the score resulted higher than 100, this highlighting the general minor relevance of the alignment associated to the protein based approach.

Next, considering details of the transcript BBHs, most of them (2467 on 2561) were detected also by the BLASTp analysis (Fig. 15-B/C). Again we noticed a small group of relationships (94) that were not found at all by the similarity search based on protein sequences (Fig. 15-C). Among them the 78 percent (64 of 94 matches) had a score higher than 100. In order to further investigate the reasons of the lack of relationships detected at protein sequence level, we deeper analyzed these 64 matches. We observed that: a) 38 alignments were expanded including UTR regions of the transcript sequences; b) 12 alignments involved reading frames not corresponding to the annotated protein coding regions. Specifically, 7 alignments involved alternative reading frames when compared with the expected protein coding ones, while 5 alignments were reconstructed with fragments from different reading frames (Fig. 15-C). These results highlight that the BBHs based on transcript sequences showed an overall better quality assessed by their scores. Moreover, it is also evident that the transcript based approach is more sensitive in detecting alignments relevant for the

identification of best orthologs, that otherwise would have been missed when based on a protein comparison approach.

3.4 Conclusion

Transcriptologs is a method for the identification of orthology relationships exploiting transcriptome sequences. The orthologs are detected by BBHs defined by revisited tBLASTx results.

As a case study, we tested *Transcriptologs* to define orthologs between *Arabidopsis thaliana* and *Sorghum bicolor*, since reference annotations are available for both species, together with ortholog collections from several external resources.

The method was implemented for a straightforward exploitation of transcript sequences, because of their direct sequencing and their higher reliability. Indeed, large scale definition of transcript sequences is today easily achievable thanks to classical (EST sequencing) and novel (RNAseq) technologies in comparison to proteome sequencing. Moreover, the revisiting of tBLASTx output performed by *Transcriptologs* overcomes possible limits in the definition of the correct coding frame. The method also exploits a wider region for the detection of similarities, including the UTRs. Therefore, as here demonstrated, it has a higher sensitivity in the detection of the BBHs.

Although classical approaches and publicly available collections are based on protein sequence similarity searches, we first showed the heterogeneity of the current collections today available, including results that are often incomparable. Then, considering the E-value of an alignment between two sequences as a surrogate of the alignment quality, we showed that orthology relationships available from these collections are not comparable neither can be interpreted based on a BLAST similarity search.

We compared the results from protein based BBHs and transcript based BBHs, and we investigated on the main differences between them. We highlighted that similarity searches at transcript level can lead to different results when compared to protein based analyses. In particular, considering the quality of the alignments, we assert that orthologs detected using transcriptomic data have higher scoring, taking advantages of reconstructed alignments that are expanded along the transcripts, including also regions with different coding frame. This approach may overcomes sequencing errors and possible limits in the detection of similarities that could be hidden at protein level.

This method may integrate classical approaches, since it confirms results from previous orthologs collections based on protein sequences and it can highlight new relationships thanks to the exploitation of a higher information content.

Moreover, *Transcriptologs* can support a widespread analytical approach such as the ortholog detection exploiting more accessible and reliable data such as those from transcript sequences.

Chapter 4. A multilevel comparison between distantly related species Tomato and Grapevine

4.1 Introduction

Numerous economically important crop species, such as Tomato, Potato, Pepper, Tobacco and other annual plants, belongs to the Solanaceae, one of the families of the Asterid clade of dicotyledonous plants. In particular, *Solanum lycopersicum* (Tomato) is considered the model organism for Solanaceae and, specifically, for fleshy fruit species.

Vitis vinifera, a perennial plant, is another economically important species, consumed as fruit or for wine production, which belongs to the Vitaceae family. Recent phylogenetic studies classified the Vitaceae family as the earliest diverging lineage of the Rosid clade (Jansen et al. 2006), making it an excellent model for the Rosids in comparative genomics studies. Although its small genome size of about 475 Mb (Lodhi and Reisch 1995), a high number of chromosomes (19) suggested an ancestral polyploidy event of *V. vinifera* genome (Lewis 1979). However, a more recent analysis of the Grapevine genome indicated the absence of both recent and ancient duplication events to the genome organization of Grapevine as well as to all of the Rosid species (Jaillon et al. 2007).

Asterids and Rosids approximately diverged from their last common ancestor 125 million years ago (Wikstrom et al. 2001). A lot of chromosomal rearrangements and a consistent genome reorganization should occur in such a long divergence period. A comparative genomics work on *Solanum lycopersicum*, *Coffea canephora* and *Vitis vinifera* (Guyot et al. 2012) detected the presence of significant synteny fragmented into relatively small blocks of about 4 Mb between the asterid and rosid clades, despite the divergent

chromosomal histories between Tomato and Grapevine. The highlighted synteny is particularly interesting considering the differences in the number of chromosomes *n* and in the genome size *x* between Tomato (*n*=12, *x*=12,965 Mb) and Grapevine (*n*=19, *x*=19,475 Mb).

To date, no other studies report on the comparison between these two economically important species. In this chapter, the comparison between *S. lycopersicum* (iTAG 2.3 annotation version) and *V. vinifera* (V1 annotation version) is described, exploiting both orthology and paralogy relationships to infer about common or peculiar aspects of both plants species.

4.2 Material and methods

4.2.1 Data sets

Genes, transcripts and proteins collections for *Solanum lycopersicum* (The_Tomato_Genome_Consortium 2012) (release iTAG 2.3, 34727 sequences) and *Vitis vinifera* (Grimplet et al. 2012) (release CRIBI V1, 29971 sequences) were downloaded from the Sol Genomics Network website (Fernandez-Pozo et al. 2015) and from CRIBI website (CRIBI), in .fasta format.

4.2.2 Orthology prediction

All-against-all sequence similarity searches between *S. lycopersicum* and *V. vinifera* genes, mRNAs and proteins collections were performed using the BLASTn, tBLASTx and BLASTp programs of the BLAST package (Camacho et al. 2009), respectively. All the similarity searches were carried out setting an Expect-value (E-value, E) cut-off to 10^{-3} and max_target_seqs parameter to 500. In order to identify orthologs between *Solanum lycopersicum* and *Vitis vinifera*, we used a dedicated program developed with Python programming language (v3.3.3) that takes in input the results of the performed similarity searches. In

order to define more extended alignments, a Python algorithm described in Chapter 3 was developed (alignment_reconstruction, Fig. 11). Moreover, this technique is based on the Bidirectional Best Hit (BBH) approach (Tatusov et al. 1997, Huynen and Bork 1998, Hughes 2005), relying on the assumption that genes x_i and y_i, from species X and Y, are the best putative orthologs if x_i is the best hit of y_i, and y_i is the best hit of x_i, in all-against-all similarity searches (Overbeek et al. 1999).

4.2.3 Paralogy prediction

For each organism separately, all-against-all genes, mRNAs and proteins similarity searches were performed using the BLASTn, tBLASTx and BLASTp programs, respectively. All the similarity searches were carried out setting two different E-value cut-off to 10^{-50} and 10^{-3}, and max_target_seqs parameter to 500.

4.2.4 Networks construction and species-specific genes identification

The network construction process took as input the BBHs and the paralogs (E-value 10^{-50}) detected before. This procedure extracted all the connected components into different separated undirected graphs, each node representing a gene, an mRNA or a protein and each edge representing an orthology or paralogy relationship. The species-specific genes identification was performed filtering out all the genes, mRNAs and proteins that share at least one orthology or paralogy relation, even at a loose E-value threshold (10^{-3}), from the complete *S. lycopersicum* and *V. vinifera* gene collections.

4.2.5 Protein domains prediction

An InterProScan (version 5.14-53.0) analysis (Jones et al. 2014) was performed on the entire protein collections of both Tomato and Grapevine, activating the "iprlookup" parameter (in order to obtain information about InterPro domains) and setting the output format to .tsv. This software, downloadable at https://www.ebi.ac.uk/interpro/download.html, allows sequences to be scanned against the InterPro database (Mitchell et al. 2015), a reference collection for protein domains.

4.2.6 Metabolic pathways and enzyme classification

A sequence similarity search between the complete *S. lycopersicum* mRNA collection and the entire Swiss-Prot protein collection was performed using the tBLASTn program of the BLAST package (Camacho et al. 2009), setting an E-value cut-off to 10^{-3} and "max_target_seqs" parameter to 500. Only the alignments with at least the 90% of identity and the 90% of coverage were retained for subsequent analyses. Among them, the tomato mRNAs that matched a Swiss-Prot protein associated to an Enzyme Commission number (EC number) were identified. This same procedure was applied also to detect the *V. vinifera* genes associated to a valid EC number.

The metabolic pathways associated to the detected enzymes were identified using the KEGG Database (Kanehisa and Goto 2000).

4.3 Results and discussion

4.3.1 Inter-species relations

In order to provide a comprehensive overview of the cross comparison between *S. lycopersicum* and *V. vinifera*, we performed all-against-all similarity searches using gene versus gene (BLASTn), translated mRNA versus translated mRNA

(tBLASTx) and protein versus protein (BLASTp) searches. This multilevel analysis was performed setting an E-value threshold of 10^{-3}.

Subsequently, orthologs between *S. lycopersicum* and *V. vinifera* were detected by Bidirectional Best Hit approach using data from genes, transcripts and proteins similarity searches performed before. In details, 13359 BBH relationships were detected by using gene sequences, 13366 BBH relationships were detected by using mRNA sequences, and 13358 ones were detected by using protein sequences (Fig. 18-A).

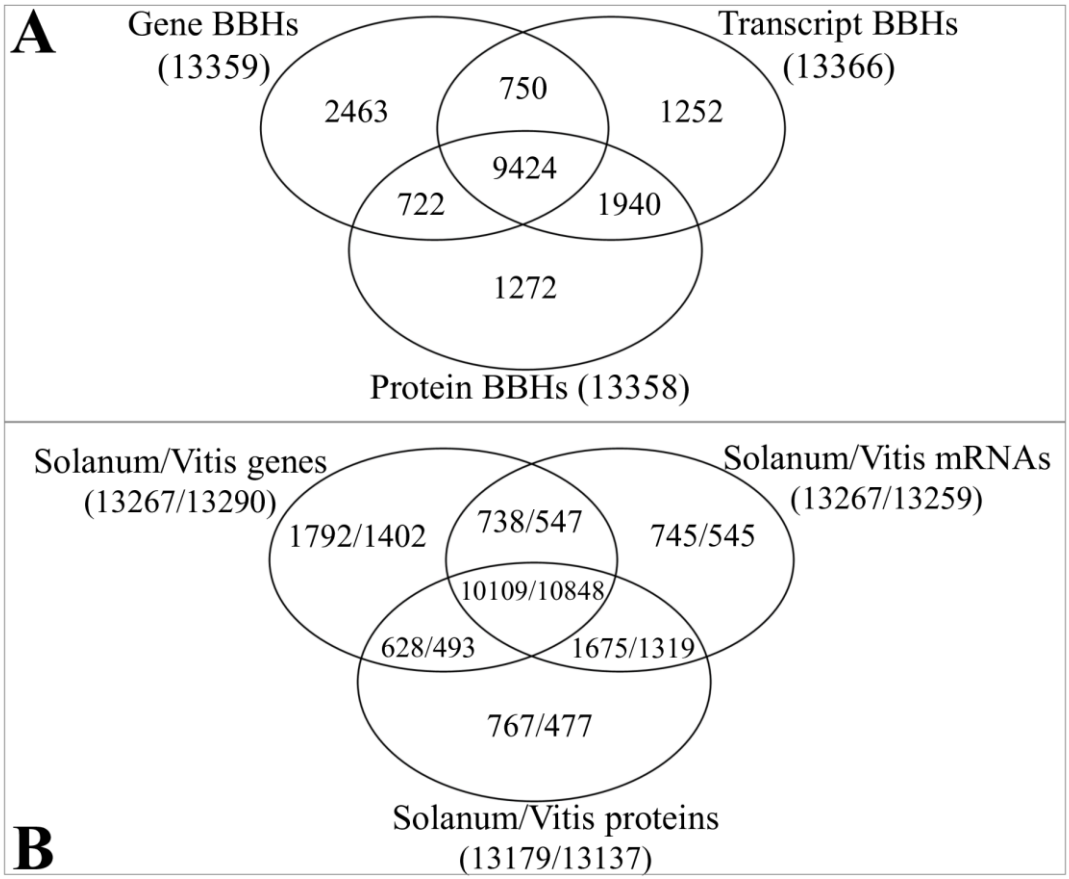
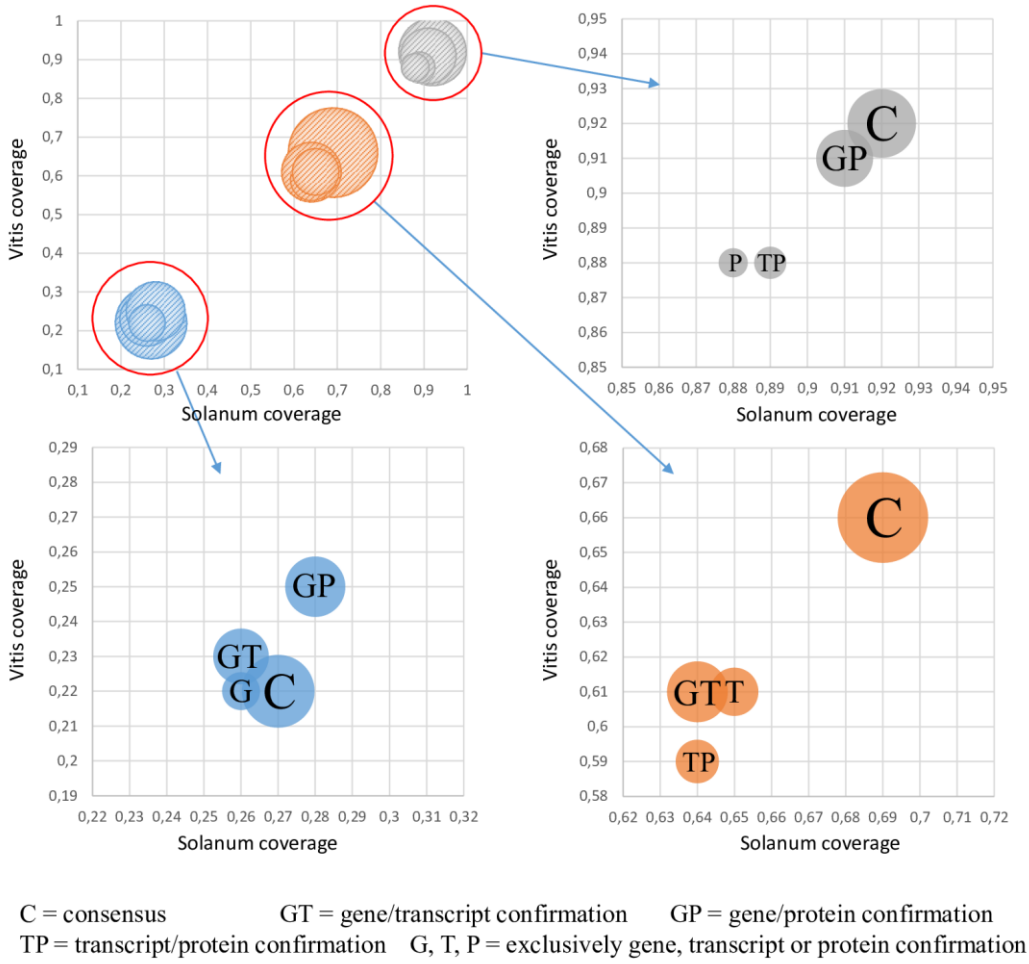


Figure 18. Comparison between genes, mRNAs and proteins similarity searches. A) Venn diagram showing differences and similarities in the number of BBHs detected using genes, mRNAs and protein sequences. **B)** Venn diagram showing the number of *S. lycopersicum* genes that have a relationship with a *V. vinifera* counterpart, and vice versa.

In these three different BBHs collections, despite the differences in the structure between genes, mRNAs and proteins, we observed a high number (9424) of matches in common. As shown in figure 19, the average score of the BBHs that are in common is higher than the ones of the remaining BBHs. This shows that the three different methods to detect a strong reliable core of orthologs between two different species.



C = consensus GT = gene/transcript confirmation GP = gene/protein confirmation
TP = transcript/protein confirmation G, T, P = exclusively gene, transcript or protein confirmation

Figure 19. In blue) Groups of BBHs detected between gene sequences. **In orange)** groups of BBHs detected between transcript sequences. **In gray)** groups of BBHs detected between protein sequences. Each circle represent a group of BBHs. The diameter of each circle is proportional to the BBH score average. The consensus groups together the BBHs that are common to all three different methods.

Moreover, considering the *S. lycopersicum* genes that have a relation with a *V. vinifera* counterpart, although there is a significant number of genes that are in common among the three collections (10109), there are 1792 genes detected exclusively by genes similarity search, 745 detected exclusively by transcripts similarity search and 767 detected exclusively by proteins similarity search (Fig. 18-B). Similar numbers were found when we considered the *V. vinifera genes* that have a relation with a *S. lycopersicum* counterpart. Also in this case, though we observed a significant number of genes that are in common among the three collections (10848), we found 1792 genes detected exclusively by genes similarity search, 745 detected exclusively by transcripts similarity search and 767 detected exclusively by proteins similarity search (Fig. 18-B). Therefore, a general overview at the results coming from the performed BBH analysis revealed the presence of 16454 *S. lycopersicum* genes that are orthologs of 15631 *V. vinifera* genes (Fig. 20). Among them, a robust and reliable core of BBH (10109 *S. lycopersicum* genes and 10848 *V. vinifera* genes), defined as the consensus of three different methods accordingly to a novel multi-level approach, was identified.

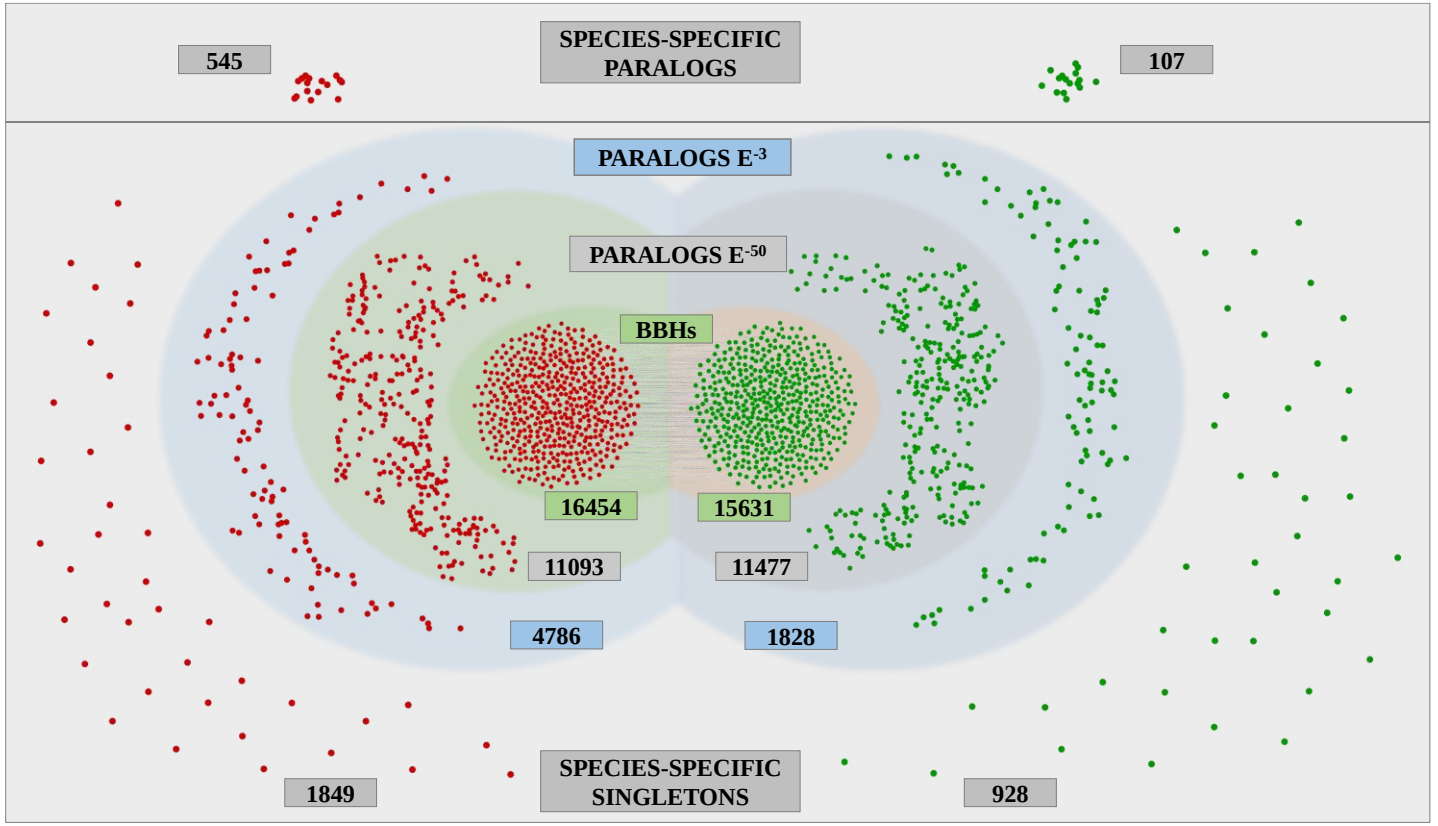


Figure 20. General overview of the cross comparison between *S. lycopersicum* and *V. vinifera*. Tomato and Grapevine genes are represented in red and in green, respectively. BBHs are shown on the orange background; paralogs detected with the stringent E-value threshold (10^{-50}) are shown on the green background; paralogs detected with the loose E-value threshold (10^{-3}) are shown on the blue background; species-specific paralogs and single-copy genes (singletons) are shown on the light gray background.

4.3.2 Intra-species relations

S. lycopersicum and *V. vinifera* paralogs were detected by all-against-all sequence similarity searches using gene, mRNA and protein sequences, respectively. Accordingly to the work of Rosenfeld et al. (Rosenfeld and DeSalle 2012), we set a stringent E-value threshold to E^{-50} in order to maximize the number and the accuracy of the gene families. With the aim of identifying expansions or reductions in the number of genes of related gene families of *S. lycopersicum* and *V. vinifera*, we identified duplicated genes starting exclusively from ortholog pairs. By this approach we expanded the ortholog collection of 11093 paralogs in Tomato and 11477 paralogs in Grapevine (Fig. 20), grouped together into 3601 networks (Fig. 21). In detail, we identify 2143 two-genes networks formed by a S. lycopersicum gene and a V. vinifera gene connected by an orthology relation, 1356 networks formed by a number of genes between 3 and 9, and 102 networks having a number of genes higher or equal to 10 (Fig. 21-A).

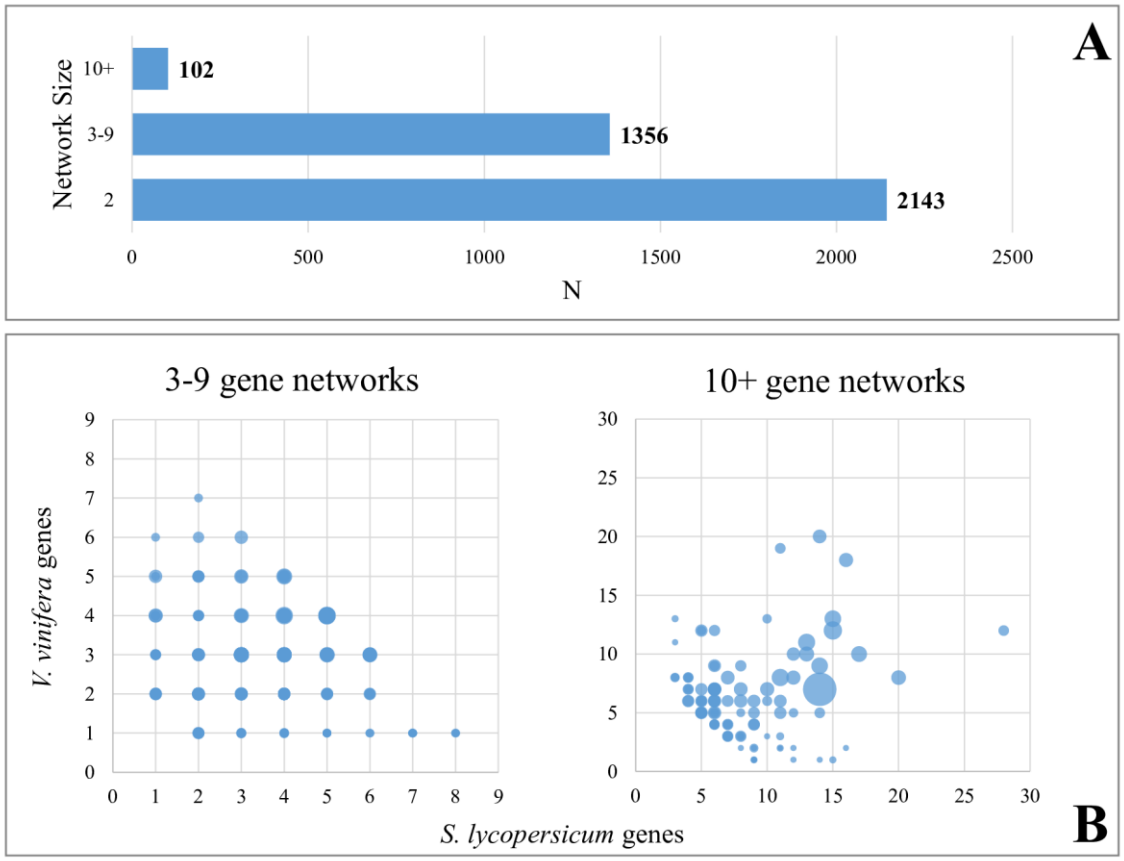


Figure 21. Ortholog/paralog networks detected with a stringent E-value threshold (10^{-50}). A) Bar chart showing the number of networks classified according to their size. **B)** Scatter plots showing the distribution of the networks based on the number of *S. lycopersicum* and *V. vinifera* genes included within them. The diameter of the circles is proportional to the number of BBHs inside each network.

In Fig. 21-B we provided an overview of the networks distribution based on their size and on the number of BBHs connecting Tomato and Grapevine genes. In these graphic representations, if we focus on the networks distributed along a hypothetical bisector that splits the charts, we are looking at the gene families that did not undergo significant changes in the number of members between the two plant species. Moreover, in order to identify the networks that have expansions or reductions in the number of genes of the related gene families of *S. lycopersicum* and *V. vinifera*, we have to look at the networks that are far from a hypothetical bisector that splits the charts. Moving towards the Cartesian axes of these charts (Fig. 21-B), the level of the expansion in the size of a gene family of a species compared to the other will increase. Furthermore, networks with larger number of orthologs are represented by circles with larger diameters. In this way, based on the number of BBHs within each network, it is possible to infer the most conserved gene families between Tomato and Grapevine.

4.3.3 Species-specific genes

The species-specific genes were detected filtering out from the complete *S. lycopersicum* and *V. vinifera* gene collections, all the genes, mRNAs and proteins that have at least one ortholog counterpart, or that share at least on paralogy relation detected starting exclusively from an ortholog pair, even at a loose E-value threshold (10^{-3}). Differences between networks detected at E-value 10^{-50} and the ones detected at E-value 10^{-3} are summarized in Table 3. It is interesting to note, for both the E-value thresholds, the presence of a large network that contains most of the nodes representing *Solanum lycopersicum* and *Vitis vinifera* genes.

	E^{-3}	E^{-50}
Total nodes	61269	54655
Total edges	3699964	1354314
Tomato nodes	32333	27547
Grapevine nodes	28936	27108
Orthology edges	17823	17823
Paralogy edges	3682141	1336491
Total networks	641	3601
Total 2-genes networks	385	2143
Total 3-9 genes networks	243	1356
Total 10+ genes networks	12	102
"Big network" nodes	59306	43236
"Big network" edges	3695231	1328306
"Big network" tomato nodes	31312	21456
"Big network" grapevine nodes	27994	21780

Table 3. Summary statistics for the network detected by using both E-value thresholds.

By this approach, we detected 514 species-specific paralogs and 1849 species-specific single-copy (singleton) genes in Tomato, and 107 species-specific paralogs and 928 species-specific singleton genes in Grapevine (Fig. 20).

4.3.4 Protein domains classification

Protein domains were predicted for both species by InterProScan software (Jones et al. 2014), providing a functional overview for Tomato and Grapevine. The more frequent domains in terms of occurrence in both species are the P-loop containing nucleoside triphosphate hydrolase domain and different types

of kinase domains (Tab. 4). It’s interesting to note how some domains have much more occurrences in a species compared to the other, such as the Aminotransferase-like plant mobile domain (112 occurences in Tomato and 10 occurrences in Grapevine) or the PGG domain (32 occurences in Tomato and 137 occurences in Grapevine) (Tab. 4).

InterPro ID	Description	N (Tom)	N (Gra)
IPR027417	P-loop containing nucleoside triphosphate hydrolase	1445	1362
IPR011009	Protein kinase-like domain	1227	1413
IPR000719	Protein kinase domain	1149	1309
IPR002290	Serine/threonine/dual specificity protein kinase, catalytic domain	848	966
IPR008271	Serine/threonine-protein kinase, active site	842	965
IPR013083	Zinc finger, RING/FYVE/PHD-type	682	520
IPR017441	Protein kinase, ATP binding site	675	789
IPR011990	Tetratricopeptide-like helical domain	575	641
IPR013320	Concanavalin A-like lectin/glucanase domain	528	747
IPR009057	Homeodomain-like	476	392
IPR002885	Pentatricopeptide repeat	474	566
IPR016040	NAD(P)-binding domain	446	484
IPR001841	Zinc finger, RING-type	435	323
IPR001245	Serine-threonine/tyrosine-protein kinase catalytic domain	419	506
IPR029058	Alpha/Beta hydrolase fold	405	382
IPR001611	Leucine-rich repeat	385	537
IPR016024	Armadillo-type fold	369	353
IPR001810	F-box domain	336	193
IPR012677	Nucleotide-binding alpha-beta plait domain	313	253
IPR015943	WD40/YVTN repeat-like-containing domain	308	288
IPR029063	S-adenosyl-L-methionine-dependent methyltransferase	294	311
IPR001005	SANT/Myb domain	294	271
IPR003593	AAA+ ATPase domain	289	316
IPR017986	WD40-repeat-containing domain	287	277
IPR011989	Armadillo-like helical	282	294
IPR000504	RNA recognition motif domain	277	225
IPR013210	Leucine-rich repeat-containing N-terminal, plant-type	271	234
IPR020846	Major facilitator superfamily domain	269	210
IPR012336	Thioredoxin-like fold	265	265
IPR001680	WD40 repeat	265	250
IPR001128	Cytochrome P450	256	396
IPR002182	NB-ARC	250	357
IPR017930	Myb domain	245	216
IPR012337	Ribonuclease H-like domain	237	86
IPR002401	Cytochrome P450, E-class, group I	236	338
IPR003591	Leucine-rich repeat, typical subtype	235	349
IPR017972	Cytochrome P450, conserved site	228	293
IPR017853	Glycoside hydrolase superfamily	212	270
IPR027443	Isopenicillin N synthase-like	199	165
IPR011992	EF-hand domain pair	198	183
IPR016177	DNA-binding domain	186	147
IPR005123	Oxoglutarate/iron-dependent dioxygenase	185	168
IPR011598	Myc-type, basic helix-loop-helix (bHLH) domain	180	139
IPR002048	EF-hand domain	177	160
IPR011991	Winged helix-turn-helix DNA-binding domain	174	161
IPR013781	Glycoside hydrolase, catalytic domain	173	221
IPR001471	AP2/ERF domain	173	136
IPR002213	UDP-glucuronosyl/UDP-glucosyltransferase	172	224
IPR012340	Nucleic acid-binding, OB-fold	171	131
IPR012334	Pectin lyase fold	164	143
IPR011050	Pectin lyase fold/virulence factor	163	143
IPR018247	EF-Hand 1, calcium-binding site	162	151
IPR014001	Helicase superfamily 1/2, ATP-binding domain	162	133
IPR007087	Zinc finger, C2H2	161	115
IPR003653	Ulp1 protease family, C-terminal catalytic domain	161	45
IPR023214	HAD-like domain	160	166
IPR001650	Helicase, C-terminal	156	128
IPR019775	WD40 repeat, conserved site	150	154
IPR007527	Zinc finger, SWIM-type	148	30
IPR003439	ABC transporter-like	144	168
IPR026992	Non-haem dioxygenase N-terminal domain	143	138
IPR029044	Nucleotide-diphospho-sugar transferases	142	159
IPR009072	Histone-fold	133	84

IPR006564	Zinc finger, PMZ-type	132	26
IPR015424	Pyridoxal phosphate-dependent transferase	131	121
IPR013026	Tetratricopeptide repeat-containing domain	129	120
IPR013830	SGNH hydrolase-type esterase domain	128	100
IPR020683	Ankyrin repeat-containing domain	126	225
IPR003959	ATPase, AAA-type, core	126	119
IPR015421	Pyridoxal phosphate-dependent transferase, major region, subdomain 1	125	114
IPR000008	C2 domain	125	101
IPR010255	Haem peroxidase	125	96
IPR014729	Rossmann-like alpha/beta/alpha sandwich fold	123	120
IPR008972	Cupredoxin	122	161
IPR005225	Small GTP-binding protein domain	121	102
IPR019734	Tetratricopeptide repeat	120	121
IPR029071	Ubiquitin-related domain	119	94
IPR002016	Haem peroxidase, plant/fungal/bacterial	119	91
IPR001878	Zinc finger, CCHC-type	119	67
IPR018289	MULE transposase domain	119	27
IPR014710	RmlC-like jelly roll fold	118	145
IPR011011	Zinc finger, FYVE/PHD-type	116	89
IPR002110	Ankyrin repeat	115	180
IPR013785	Aldolase-type TIM barrel	113	105
IPR001623	DnaJ domain	113	97
IPR015300	DNA-binding pseudobarrel domain	113	53
IPR023213	Chloramphenicol acetyltransferase-like domain	112	91
IPR019557	Aminotransferase-like, plant mobile domain	112	10
IPR016140	Bifunctional inhibitor/plant lipid transfer protein/seed storage helical domain	111	70
IPR001932	PPM-type phosphatase domain	110	87
IPR001356	Homeobox domain	109	83
IPR023393	START-like domain	108	77
IPR006447	Myb domain, plants	108	71
IPR000823	Plant peroxidase	107	81
IPR003676	Small auxin-up RNA	106	77
IPR002100	Transcription factor, MADS-box	105	61
IPR021109	Aspartic peptidase domain	103	77
IPR003340	B3 DNA binding domain	103	51
IPR003441	NAC domain	102	74
IPR002347	Glucose/ribitol dehydrogenase	101	100
IPR002198	Short-chain dehydrogenase/reductase SDR	101	95
IPR003480	Transferase	101	82
IPR017451	F-box associated interaction domain	101	19
IPR020472	G-protein beta WD-40 repeat	100	100
IPR019793	Peroxidases heam-ligand binding site	100	81
IPR001480	Bulb-type lectin domain	96	119
IPR010987	Glutathione S-transferase, C-terminal-like	95	114
IPR011051	RmlC-like cupin domain	82	119
IPR004045	Glutathione S-transferase, N-terminal	77	104
IPR008949	Isoprenoid synthase domain	69	114
IPR008930	Terpenoid cyclases/protein prenyltransferase alpha-alpha toroid	53	139
IPR016039	Thiolase-like	50	105
IPR026961	PGG domain	32	137

Table 4. Summary of protein domains common to both species detected by scanning the IntertPro database. All the domains with at least 100 occurrences in Tomato or Grapevine are listed.

If we focus on the proteins domains exclusively present in one species rather than the other, we notice different helicaces domains (DNA helicase domains, DNA helicase Pif1-like domains, Helitron helicase-like domains) are present exclusively in Tomato and not in Grapevine (Tab. 5).

InterPro ID	Description	N
IPR015410	Domain of unknown function DUF1985	101
IPR010285	DNA helicase Pif1-like	60
IPR025312	Domain of unknown function DUF4216	43
IPR002648	Adenylate dimethylallyltransferase	24
IPR003840	DNA helicase	20
IPR021929	Late blight resistance protein R1	17
IPR025476	Helitron helicase-like domain	17
IPR025398	Domain of unknown function DUF4371	16
IPR028919	Viral movement protein	11
IPR000114	Ribosomal protein L16	10
IPR003871	Domain of unknown function DUF223	10
IPR006912	Harbinger transposase-derived protein	10
IPR009632	Fruit-specific protein	10
IPR000310	Orn/Lys/Arg decarboxylase, major domain	8
IPR001268	NADH:ubiquinone oxidoreductase, 30kDa subunit	8
IPR005798	Cytochrome b/b6, C-terminal	8
IPR019645	Uncharacterised protein family Ycf15	8
IPR000515	ABC transporter type 1, transmembrane domain MetI-like	7
IPR004231	Carboxypeptidase A inhibitor-like	7
IPR020798	Ribosomal protein L16, conserved site	7
IPR006706	Extensin domain	6
IPR025452	Domain of unknown function DUF4218	6
IPR001457	NADH:ubiquinone/plastoquinone oxidoreductase, chain 6	5
IPR001516	NADH-Ubiquinone oxidoreductase (complex I), chain 5 N-terminal	5
IPR007836	Ribosomal protein L41	5
IPR012942	Sensitivity To Red Light Reduced-like, SRR1	5
IPR016213	Polyphenol oxidase	5
IPR017452	GPCR, rhodopsin-like, 7TM	5
IPR001463	Sodium:alanine symporter	4
IPR001705	Ribosomal protein L33	4
IPR016439	Ceramide synthase component Lag1/Lac1	4
IPR017443	Ribulose bisphosphate carboxylase, large subunit, ferrodoxin-like N-terminal	4
IPR022546	Uncharacterised protein family Ycf68	4
IPR029480	Transposase-associated domain	4

Table 5. Summary of protein domains exclusively detected in Tomato. All the domains with at least 4 occurrences are shown.

Similarly, we noticed a Leucine-rich repeat and some Aerolysin or Agglutinin domains exclusively detected in the Grapevine protein collection (Tab 6).

InterPro ID	Description	N
IPR011713	Leucine-rich repeat 3	15
IPR025314	Domain of unknown function DUF4219	5
IPR005830	Aerolysin	4
IPR008998	Agglutinin domain	4
IPR023307	Aerolysin-like toxin, beta complex domain	4

Table 6. Summary of protein domains exclusively detected in Grapevine. All the domains with at least 4 occurrences are shown.

4.3.5 Metabolic pathways and enzyme classification

With the aim of highlighting common or distinctive metabolic features of the compared species, we performed sequence similarity searches between *S. lycopersicum*, *V. vinifera* and the entire Swiss-Prot protein collections, identifying the enzyme-coding genes of both plants. The represented Venn diagram of the detected enzymatic classes (Fig. 22) shows that 168 and 38 of them were detected exclusively in Tomato and Grapevine, respectively. The most represented enzymatic class exclusively detected in Tomato belongs to the oxidoreductases and transferases (Annex A), while in Grapevine belongs to the transferases and lyases (Annex B).

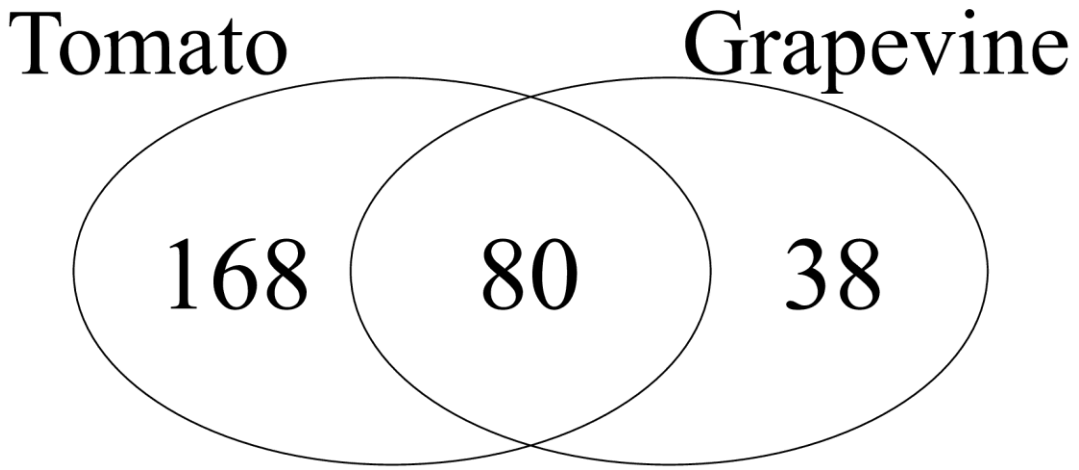


Figure 22. Venn diagram showing peculiar or common enzymatic classes detected in Tomato and Grapevine.

We investigate the KEGG database (Kanehisa and Goto 2000) to find the metabolic pathways in which the previously detected enzymatic classes are involved. In this way, we were able to detect pathways that involve enzymatic classes exclusively detected in Tomato (Tab 7), assuming the existence of some metabolic pathways preferentially activated in Tomato rather than in Grapevine. We were not able to detect, however, specific metabolic pathways in Grapevine.

KEGG_ID	PATHWAY DESCRIPTION
ec00332	Carbapenem biosynthesis
ec00642	Ethylbenzene degradation
ec00760	Nicotinate and nicotinamide metabolism
ec00052	Galactose metabolism
ec00073	Cutin, suberine and wax biosynthesis
ec00100	Steroid biosynthesis
ec00140	Steroid hormone biosynthesis
ec00231	Puromycin biosynthesis
ec00281	Geraniol degradation
ec00310	Lysine degradation
ec00340	Histidine metabolism
ec00351	DDT degradation
ec00361	Chlorocyclohexane and chlorobenzene degradation
ec00362	Benzoate degradation
ec00363	Bisphenol degradation
ec00364	Fluorobenzoate degradation
ec00365	Furfural degradation
ec00400	Phenylalanine, tyrosine and tryptophan biosynthesis
ec00401	Novobiocin biosynthesis
ec00430	Taurine and hypotaurine metabolism
ec00511	Other glycan degradation
ec00523	Polyketide sugar unit biosynthesis
ec00524	Butirosin and neomycin biosynthesis
ec00531	Glycosaminoglycan degradation
ec00564	Glycerophospholipid metabolism
ec00565	Ether lipid metabolism
ec00591	Linoleic acid metabolism
ec00622	Xylene degradation
ec00623	Toluene degradation
ec00627	Aminobenzoate degradation
ec00633	Nitrotoluene degradation
ec00643	Styrene degradation
ec00660	C5-Branched dibasic acid metabolism
ec00730	Thiamine metabolism
ec00780	Biotin metabolism
ec00790	Folate biosynthesis
ec00860	Porphyrin and chlorophyll metabolism
ec00903	Limonene and pinene degradation
ec00904	Diterpenoid biosynthesis
ec00905	Brassinosteroid biosynthesis
ec00930	Caprolactam degradation
ec00950	Isoquinoline alkaloid biosynthesis
ec00960	Tropane, piperidine and pyridine alkaloid biosynthesis
ec00966	Glucosinolate biosynthesis
ec00981	Insect hormone biosynthesis
ec01040	Biosynthesis of unsaturated fatty acids
ec01056	Biosynthesis of type II polyketide backbone

Table 7. Summary of metabolic pathways exclusively detected in Tomato.

4.4 Conclusion

Although Tomato and Grapevine are phylogenetically distant species, we showed the presence of a strong core of orthologs genes, detected according to a multi-level procedure with three different methods.

Moreover, networks of ortholog/paralog genes were built between the compared species, to investigate about the organization and the evolution of gene families in different organisms. By this approach, we detected gene families of one species that undergoes an expansion/reduction in the number of their elements when compared to the other species. The analysis of such networks allows also identifying cases in which genes belonging to a given gene family are closely related by orthology/paralogy relationships to gene with an unknown function or incomplete annotation, enabling the transfer of the information relating to the gene family.

Species-specific genes of Tomato and Grapevine, with no shared sequence similarity with the other species, were also detected.

The protein domains common to both species, as the ones exclusively detected in Tomato and Grapevine, were predicted.

Finally, the common and the distinctive enzymatic classes and the related metabolic pathway were predicted for the two compared species.

In this chapter we showed how different organism are related between them at genomic level, detecting those conserved genes that preside similar functions and regulative mechanisms, and identifying those genes that give to each organism its peculiar features.

Chapter 5. Homologies prediction between Tomato and Potato highlights unique features and common aspects in the family of *Solanaceae*

5.1 Introduction

The Solanaceae is a large family of more than 3000 species, including tuber or fruit-bearing vegetables such as Tomato (*Solanum lycopersicum*), Potato (*Solanum tuberosum*), pepper (*Capsicum annuum*) and eggplant/aubergine (*Solanum melongena*) (Knapp 2002, Sesso et al. 2003) This family is economically very important, and is the most valuable in terms of vegetable crops (Foolad 2007). The Solanaceae show a considerable adaptability to different climatic conditions, showing a remarkable phenotypic diversity between the species belonging to this family (Knapp 2002).

The tuber crop Potato is the fourth most important food crop in the world (after wheat, maize and rice) (Knapp 2002). The tubers represents for the human diet a fundamental source of starch, vitamins and antioxidants (Burlingame et al. 2009). Biodiversity of Potato is quite deep, with more than 4000 known varieties, many of which belonging to the *Solanum tuberosum* species (Burlingame et al. 2009).

Potato genome has an estimated size of 844 Mb split over 12 chromosomes (Xu et al. 2011). The Potato Genome Sequencing Consortium (PGSC) has sequenced two diploid Potato genotypes: the heterozygous diploid *S. tuberosum* Group Tuberosum genetics line RH89-039-16 (RH) and the doubled monoploid *S. tuberosum* Group Phureja clone DM1–3 (DM) (Watanabe 2015). 39,031 protein-coding genes were predicted using an assembly of the 86% of the whole genome of the doubled monoploid Potato clone (Xu et al. 2011).

Although the importance of tubers is universally recognized, the evolutionary and developmental mechanisms that underlie their initiation and growth are still unclear (Xu et al. 2011). Therefore, comparative genomics strategies, which take into account close related species of *Solanum tuberosum*, may help to unravel hidden features about the tubers.

Among species belonging to the Solanum genus, *Solanum lycopersicum* is widely accepted as a reference, and is closely related to *Solanum tuberosum*. Tomato and Potato diverged approximately 12 million years ago (Moniz de Sa and Drouin 1996). The genetic colinearity, namely the arrangement of genes on chromosomes in the same orther, between Potato and Tomato chromosomes was demonstrated by different comparative analyses (Bonierbale et al. 1988, Tanksley et al. 1992). The main structural difference between the genomes of Potato and Tomato consists in five chromosomal rearrangements involving only a single break near the centromeres (Paterson et al. 2000). Overall, Tomato and Potato have a difference in the gene copy number in rearranged segments of about 7%. This difference in the copy number is compatible with the observation that Tomato has a slightly larger genome size compared to Potato (Peters et al. 2012). Moreover, discovery of an inversion associated with chromosome 6 suggests that the Potato and Tomato genomes may contain significantly more structural rearrangements than those previously reported by the earlier comparative analyses (Iovene et al. 2008). Finally, a comparative genomic study associated to the release of the Tomato genome peaked to nine the number of large inversions between Potato and Tomato (The_Tomato_Genome_Consortium 2012). The same study predicted the existence of 18,320 orthologs pairs between Tomato and Potato.

In this chapter, in order to give a more comprehensive definition of the sequence-based homology relationships between *S. lycopersicum* and *S. tuberosum*, including a paralogy detection analysis in addition to the orthologs

definition, the multilevel comparison that integrates gene, mRNA and protein similarity searches is described.

5.2 Material and methods

5.2.1 Data sets

Genes, mRNAs and proteins collections from *Solanum lycopersicum* (The_Tomato_Genome_Consortium 2012) (release iTAG 2.3, 34727 sequences) and *Solanum tuberosum* Group Phureja (Xu et al. 2011) (release PGSC 4.03, 39028 representative sequences) were downloaded from the Sol Genomics Network website (Fernandez-Pozo et al. 2015) and from Spud DB website (Hirsch et al. 2014).

The UniProtKB reviewed (Swiss-Prot) protein collection was downloaded from the Uniprot database (Uniprot_consortium 2015), available at http://www.uniprot.org/uniprot/.

The list of all enzymes and the related pathways was obtained from the KEGG database (Kanehisa and Goto 2000), available at http://www.genome.jp/kegg/.

5.2.2 Orthology prediction

All-against-all sequence similarity searches between complete *S. lycopersicum* and *S. tuberosum* genes, mRNAs and proteins collections were performed using the BLASTn, tBLASTx and BLASTp programs of the BLAST package (Camacho et al. 2009), respectively. All the similarity searches were carried out setting an E-value cut-off to 10^{-3} and max_target_seqs parameter to 500. In order to identify orthologs between *Solanum lycopersicum* and *Solanum tuberosum*, we used a dedicated program developed with Python programming language (v3.3.3) that takes in input the results of the performed similarity searches. In order to define more extended alignments, I developed a Python

algorithm described in Chapter 3 (alignment_reconstruction, Fig. 11). Moreover, we implemented the pipeline with the Bidirectional Best Hit (BBH) approach (Tatusov et al. 1997, Huynen and Bork 1998, Hughes 2005), which is based on the assumption that genes x_i and y_i, from species X and Y, are the best putative orthologs if x_i is the best hit of y_i, and y_i is the best hit of x_i, in all-against-all similarity searches (Overbeek et al. 1999).

5.2.3 Paralogy prediction

For each organism separately, all-against-all genes, mRNAs and proteins similarity searches were performed using the BLASTn, tBLASTx and BLASTp programs, respectively. All the similarity searches were carried out setting two different E-value cut-off to 10^{-50} and 10^{-3}, and max_target_seqs parameter to 500.

5.2.4 Networks construction and species-specific genes identification

The network construction process considered as input the BBHs and the paralogs, detected with an E-value cutoff of 10^{-50}, previously identified. All the connected components are organized into separated undirected graph, each node representing a gene, mRNA or protein and each edge representing an orthology or paralogy relationship. The species-specific genes prediction was performed filtering out all the genes, mRNAs and proteins that share at least one homology relationship, even at a loose E-value threshold (10^{-3}), from the complete *S. lycopersicum* and *S. tuberosum* gene collections.

5.2.5 Protein domains prediction

An InterProScan (version 5.14-53.0) analysis (Jones et al. 2014) was performed on the entire protein collections of Tomato and Potato, activating the

"iprlookup" parameter and setting the output format to .tsv. This software, downloadable at https://www.ebi.ac.uk/interpro/download.html, allows sequences to be scanned against the InterPro database (Mitchell et al. 2015), a reference collection for protein domains.

5.2.6 Metabolic pathways and enzyme classification

Sequence similarity searches between the entire Swiss-Prot protein collection and the complete Tomato and Potato mRNA collection, respectively, was performed using the tBLASTn program of the BLAST program (Camacho et al. 2009), setting an E-value cut-off to 10^{-3} and max_target_seqs parameter to 500. The alignments with at least the 90% of identity and the 90% of coverage were retained for subsequent analyses. Among them, the Tomato and Potato mRNAs that matched a Swiss-Prot protein associated to an Enzyme Commission number (EC number) were identified.

The metabolic pathways associated to the detected enzymes were identified using the KEGG Database (Kanehisa and Goto 2000).

5.3 Results and discussion

5.3.1 Inter-species relations

With the aim of providing a comprehensive overview of the comparison between the related species of *S. lycopersicum* and *S. tuberosum,* we performed all-against-all similarity searches using gene versus gene (BLASTn), translated mRNA versus translated mRNA (tBLASTx) and protein versus protein (BLASTp) searches. This integrated analysis was performed setting an E-value threshold of 10^{-3}.

Then, orthologs between *S. lycopersicum* and *S. tuberosum* were detected by the Bidirectional Best Hit methodology using data from the genes, transcripts

and proteins similarity searches performed before. In detail, 21015 BBHs were detected by using gene sequences, 19683 BBHs were detected by using mRNA sequences, and 19550 BBHs were detected by using protein sequences (Fig. 23).

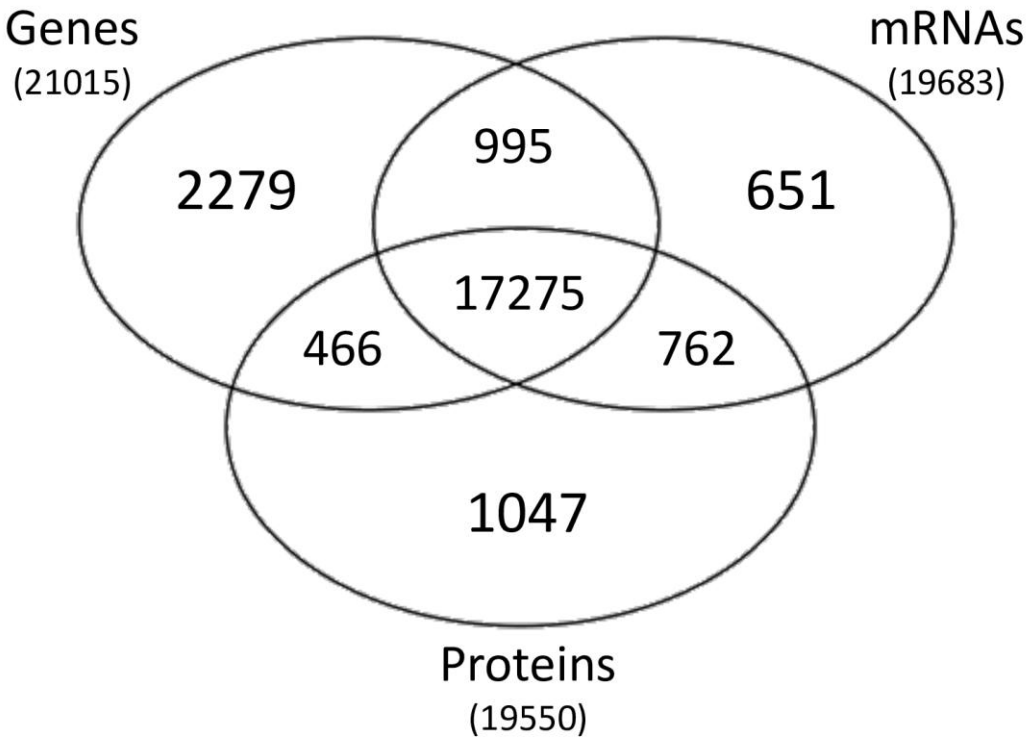


Figure 23. Venn diagram showing differences and similarities in the number of BBHs detected using genes, mRNAs and protein sequences.

We can notice a large number of BBHs (17275) that are in common between the three different analyses, corresponding to more than 82% of the total number of relationships detected by each different method. Concerning the differences between the three different methods, it is evident that the number of relationships detected exclusively by using gene sequences (2279) is much larger than the ones detected by using mRNA (651) or protein (1047) sequences. This emphasizes that the intronic regions included within the gene sequences probably are less conserved between two different species in comparison to mRNA or protein sequences.

Then, considering the Tomato genes that have a relationship with a Potato counterpart, although there is a large number of genes that are in common among the three collections (17801), there are 1804 genes detected exclusively by genes similarity search, 322 detected exclusively by transcripts similarity search and 443 detected exclusively by proteins similarity search (Fig. 24).

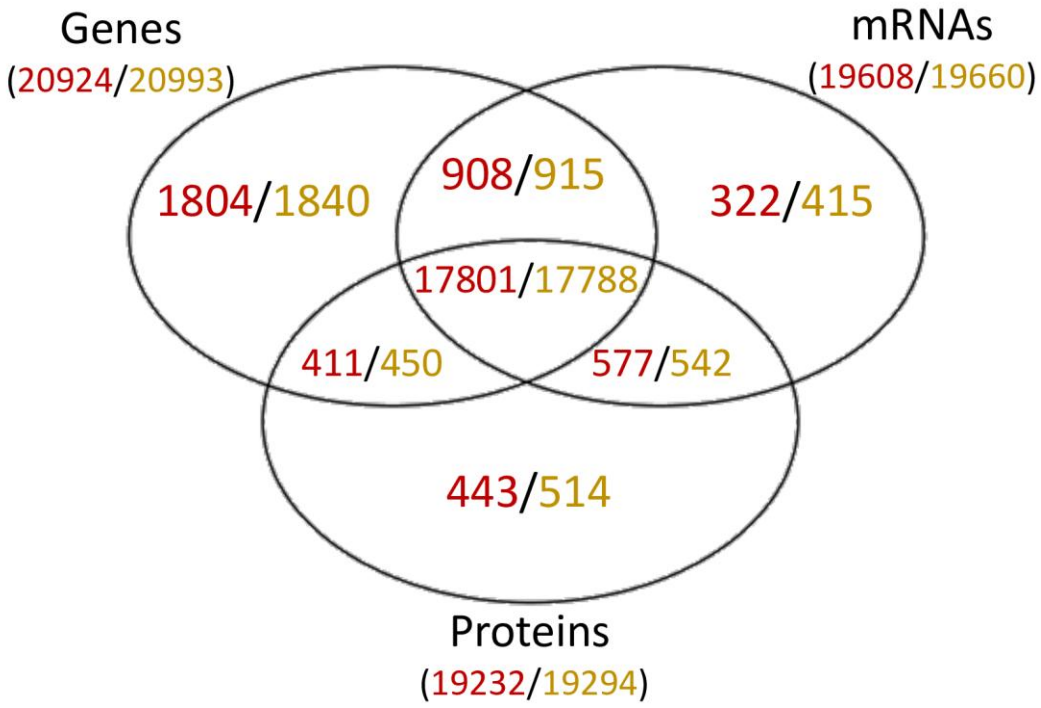


Figure 24. Venn diagram showing the number of *S. lycopersicum* genes that have a relationship with a *S. tuberosum* counterpart, and vice versa. Genes, mRNAs and proteins from Tomato are shown in red; genes, mRNAs and proteins from Potato are shown in dark yellow.

We detected similar numbers when we considered the *S. tuberosum* genes that have a relation with a *S. lycopersicum* counterpart. Again, though we observed a large number of genes that are in common among the three collections (17788), we identified 1840 genes detected exclusively by genes similarity search, 415 detected exclusively by transcripts similarity search and 514 detected exclusively by proteins similarity search (Fig. 24). A further analysis

on these species-specific datasets of genes, mRNAs and proteins, can provide valuable insights on distinctive features of Tomato and Potato.

A general overview at the results coming from the performed BBH analysis revealed the presence of 22266 *S. lycopersicum* genes that are orthologs of 22464 *S. tuberosum* genes (Fig. 25). Among them, a robust and reliable core of BBHs (17801 *S. lycopersicum* genes and 17788 *S. tuberosum* genes), defined as the consensus of a multi-level approach exploiting three different methods, was identified.

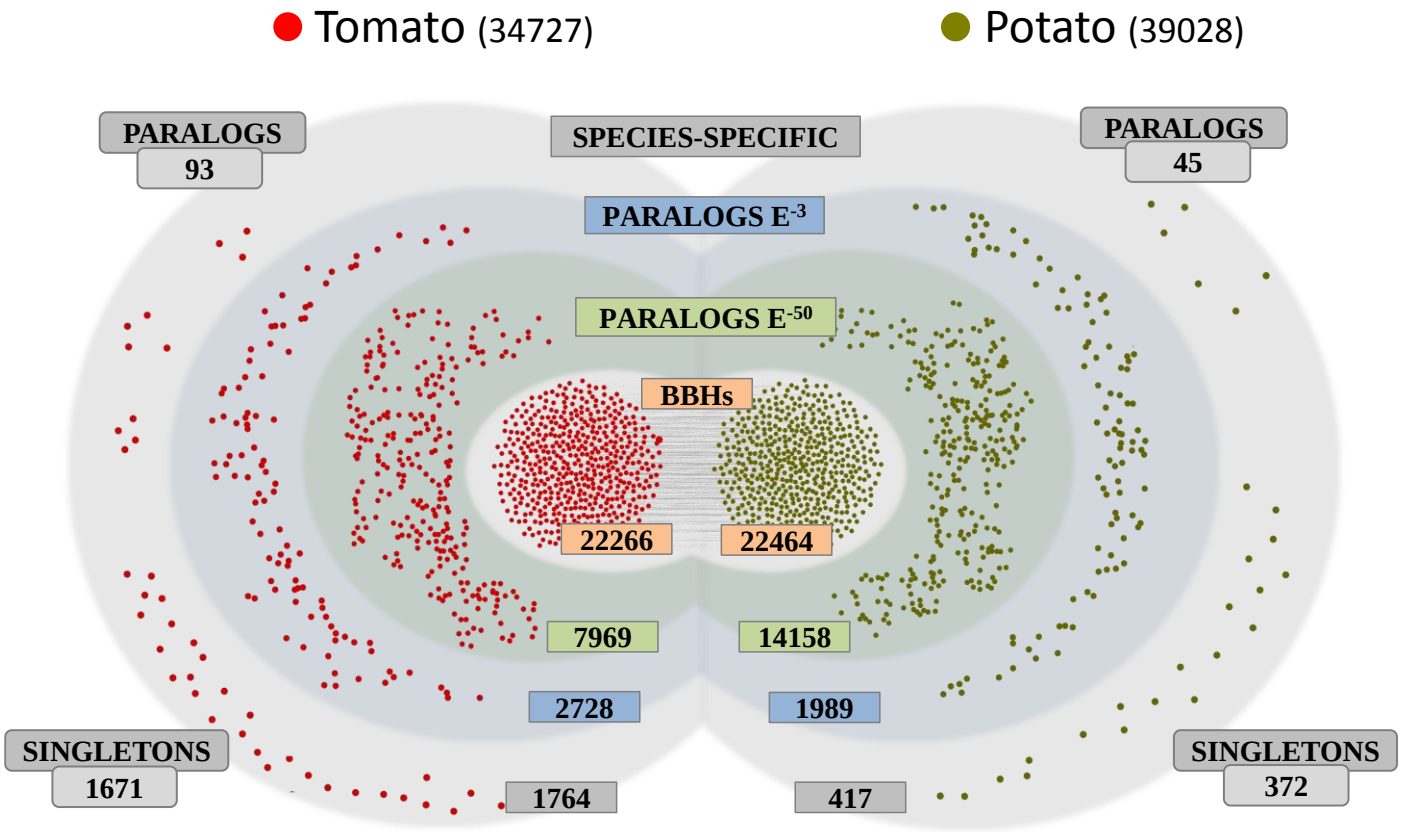


Figure 25. General overview of the cross comparison between *S. lycopersicum* and *S. tuberosum*. Tomato and Potato genes are represented in red and in dark green, respectively. BBHs are shown on the white background; paralogs detected with the stringent E-value threshold (10^{-50}) are shown on the green background; paralogs detected with the loose E-value threshold (10^{-3}) are shown on the blue background; species-specific paralogs and single-copy genes (singletons) are shown on the light gray background.

5.3.2 Intra-species relations

S. lycopersicum and *S. tuberosum* paralogs were detected by all-against-all sequence similarity searches using gene, mRNA and protein sequences, respectively. As described in the work of Rosenfeld et al. (Rosenfeld and DeSalle 2012), we set a stringent E-value threshold to E^{-50} in order to maximize the number and the accuracy of the gene families. In order to identify expansions or reductions in the number of genes of related gene families of *S. lycopersicum* and *S. tuberosum*, we detected duplicated genes starting exclusively from ortholog pairs. By this approach we expanded the ortholog collection of 7969 paralogs in Tomato and 14158 paralogs in Potato (Fig. 25), grouped together into 4924 networks. In detail, we identify 3283 two-genes networks formed by a Tomato gene and a Potato gene connected by an orthology relation, 1517 networks formed by a number of genes between 3 and 9, and 124 networks having a number of genes higher or equal to 10 (Fig. 26).

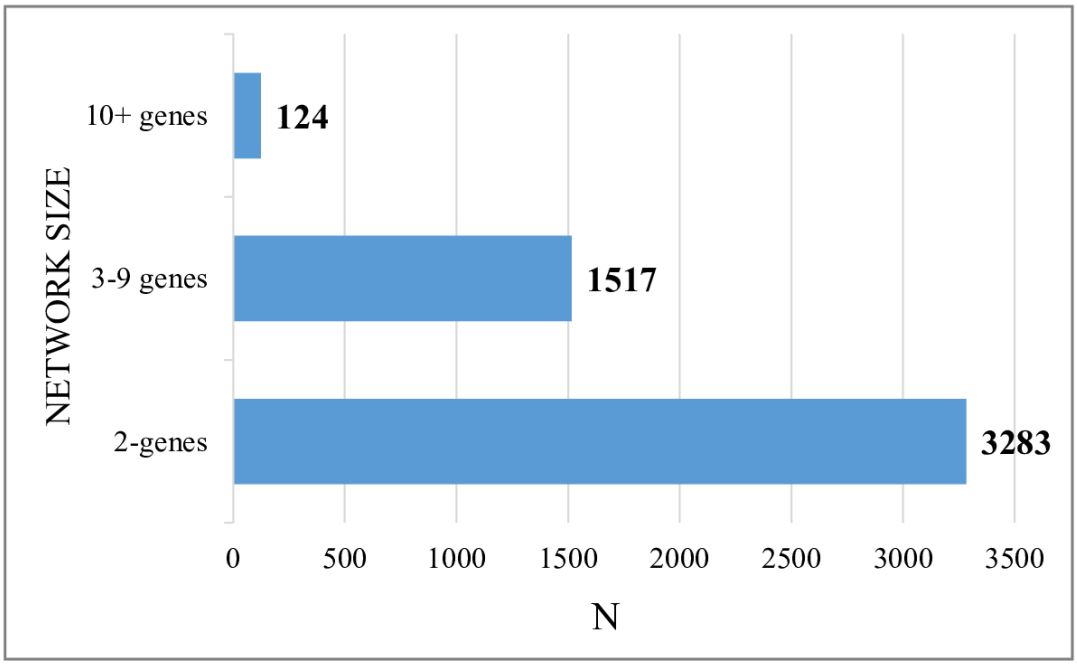


Figure 26. Bar chart showing the number of ortholog/paralog networks detected with a stringent E-value threshold (10^{-50}) and classified according to their size.

In Fig. 27, it is shown an overview of the distribution of networks containing ten or more genes based on their size and on the number of orthology relationships connecting Tomato and Potato genes.

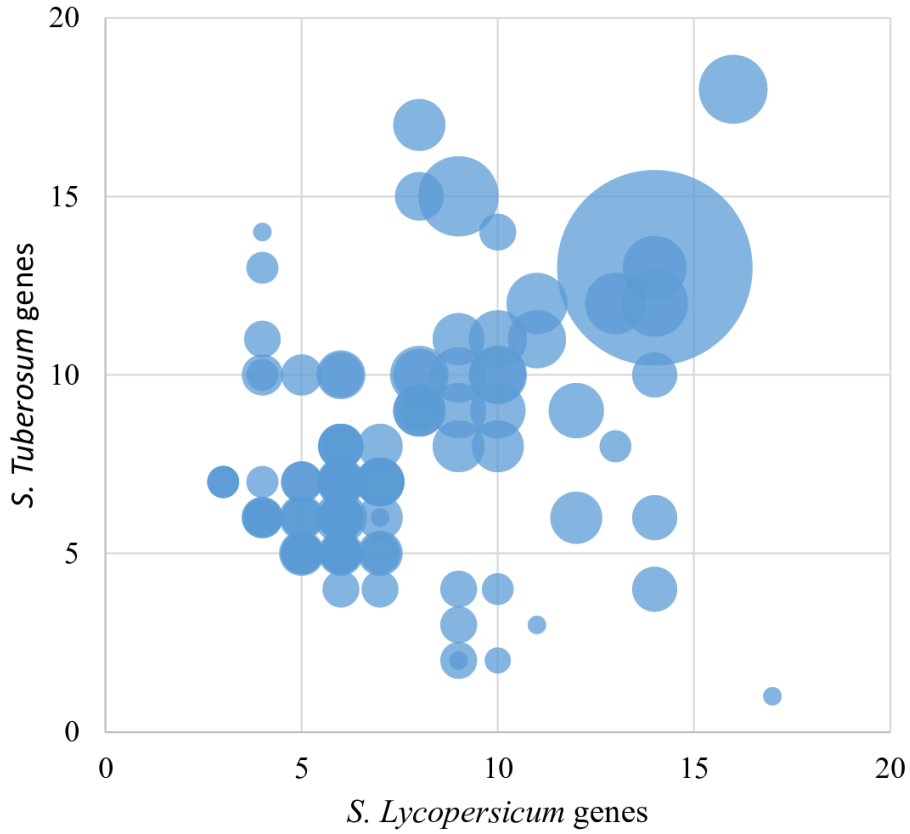


Figure 27. Scatter plot showing the distribution of the networks containing ten or more genes based on the number of *S. lycopersicum* and *S. tuberosum* genes included within them. The diameter of the circles is proportional to the number of orthology relationships inside each network.

In this graphic representation, if we want to look at the gene families that did not undergo significant changes in the number of members between the two plant species, we have to focus on the networks distributed along a hypothetical bisector that splits the charts. Moreover, if we want to look at the networks that

have expansions or reductions in the number of genes of the related gene families of *S. lycopersicum* and *S. tuberosum*, we have to focus on the networks that are far from a hypothetical bisector that splits the charts. Moving towards the Cartesian axes of this chart, the level of the expansion in the size of a gene family of a species compared to the other will increase. For example, in the lower right area of this graph (Fig. 27) we can note a network containing one potato gene connected by one orthology relationship to a sub-network of seventeen tomato paralogs (Fig. 28). In this specific case, we can observe how a highly duplicated gene in tomato remains on the contrary in single copy in potato. Most of these genes are classified as unknown protein, with the exception of two tomato genes annotated as subunits of two different enzymatic complexes involved in respiratory chain in mitochondria.

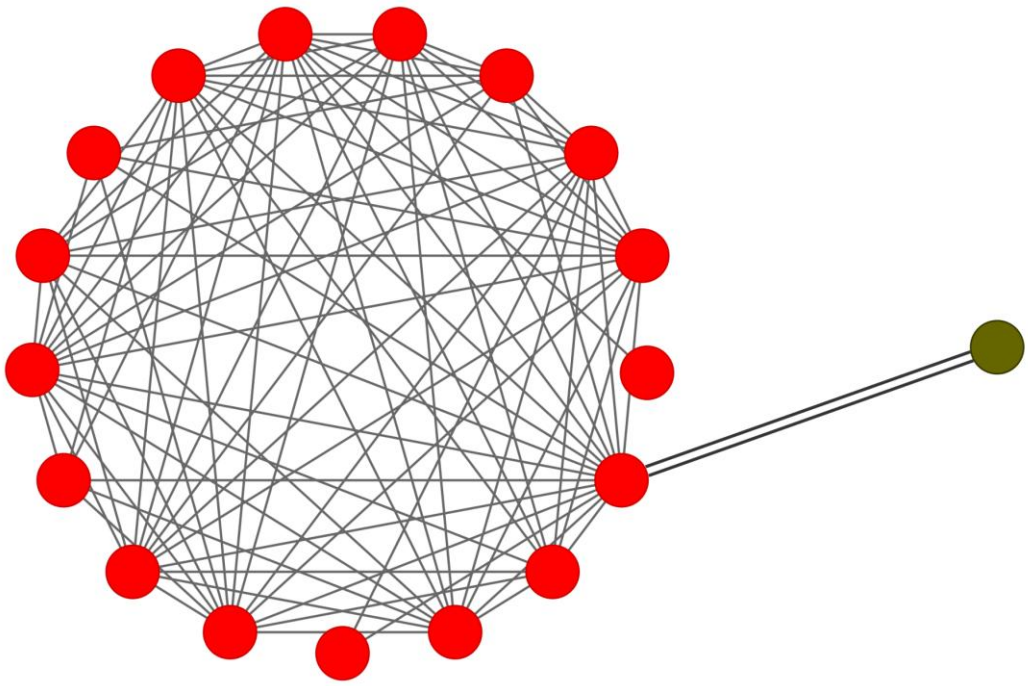


Figure 28. Cytoscape representation of a network containing seventeen tomato genes (red circles) and one potato gene (dark green circle); the double line in black represents an orthology relationship (BBH); the single lines in gray represent paralogy relationships.

Moreover, networks with larger number of orthologs are represented by circles with larger diameters (Fig. 27). In this way, based on the number of orthology relationships within each network, it is possible to infer the most conserved gene families between Tomato and Potato.

5.3.3 Species-specific genes

The species-specific genes of both species were detected filtering out from the complete *S. lycopersicum* and *S. tuberosum* gene collections, all the genes, mRNAs and proteins that have at least one ortholog counterpart, or that share at least on paralogy relation detected starting exclusively from an ortholog pair, even at a loose E-value threshold (10^{-3}). Differences between networks detected at E-value 10^{-50} and the ones detected at E-value 10^{-3} are summarized in Table 8. It is interesting to note, for both the E-value thresholds, the presence of a large network that contains most of the nodes representing *Solanum lycopersicum* and *Solanum tuberosum* genes.

	10^{-3}	10^{-50}
Total nodes	71574	66857
Total edges	3202428	954664
Orthology edges	23475	23475
Paralogy edges	3178953	931189
Tomato nodes	32963	30235
Potato nodes	38611	36622
Total networks	693	4924
Total 2-genes networks	485	3283
Total 3-9 genes networks	197	1517
Total 10+ genes networks	10	124
"big network" nodes	69619	51754
"big network" edges	3200421	938737
"big network" tomato nodes	32012	22803
"big network" potato nodes	37607	28951

Table 8. Overview of the networks statistics detected by using both E-value thresholds.

In this way, we were able to detect 93 species-specific paralogs and 1671 species-specific single-copy (singleton) genes in Tomato, and 45 species-specific paralogs and 372 species-specific singleton genes in Potato (Fig. 25). In Fig. 29, a graphical representation of all the networks of species-specific paralogs, with the exception of the two genes networks, is provided. It is interesting to note how one of these species-specific networks, which contains nine potato paralog genes (Fig. 29 on the right), shows a high degree of connection, each node being connected with each other by a paralogy relationship. In this way, we were able to identify quite clearly a new putative gene family, with a still unknown annotation.

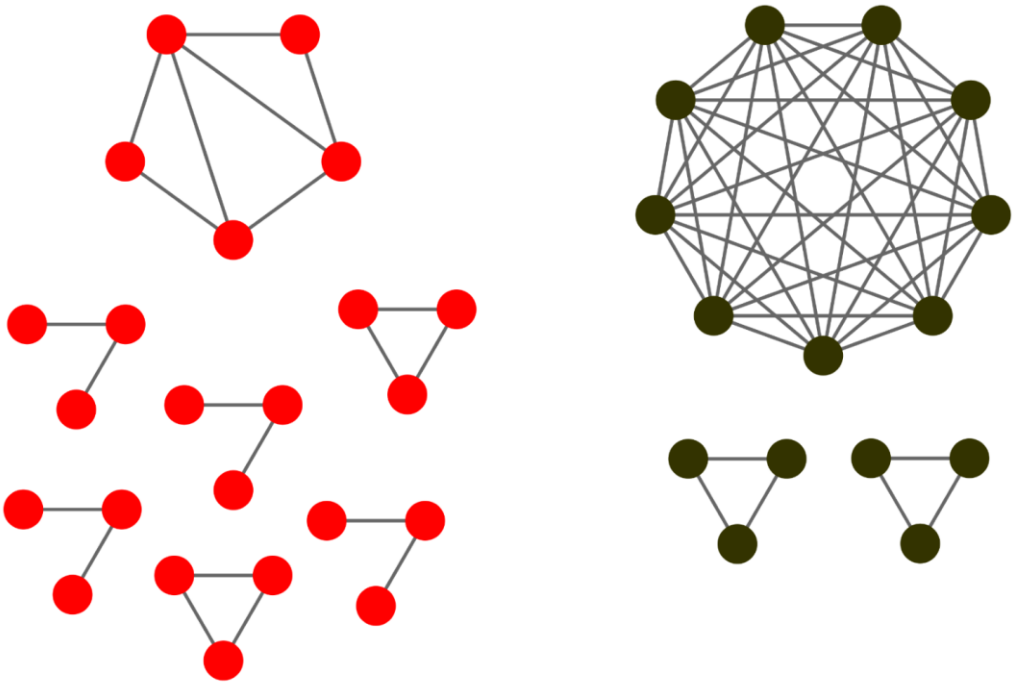


Figure 29. Cytoscape representation of species-specific paralog networks containing tomato genes (red circles) and potato genes (dark green circle); the single lines in gray represent paralogy relationships.

5.3.4 Protein domains classification

In order to provide a functional overview of Tomato and Potato, protein domains were predicted for both species by InterProScan software (Jones et al. 2014). The more frequent domains in terms of occurrence in both species are shown in Table 9.

InterPro ID	Description	N (Tom)	N (Pot)
IPR027417	P-loop containing nucleoside triphosphate hydrolase	1445	1319
IPR011009	Protein kinase-like domain	1227	1243
IPR000719	Protein kinase domain	1149	1169
IPR002290	Serine/threonine/dual specificity protein kinase, catalytic domain	848	860
IPR008271	Serine/threonine-protein kinase, active site	842	840
IPR013083	Zinc finger, RING/FYVE/PHD-type	682	641
IPR017441	Protein kinase, ATP binding site	675	712
IPR011990	Tetratricopeptide-like helical domain	575	552
IPR013320	Concanavalin A-like lectin/glucanase domain	528	572
IPR009057	Homeodomain-like	476	427
IPR002885	Pentatricopeptide repeat	474	502
IPR016040	NAD(P)-binding domain	446	432
IPR001841	Zinc finger, RING-type	435	425
IPR001245	Serine-threonine/tyrosine-protein kinase catalytic domain	419	378
IPR029058	Alpha/Beta hydrolase fold	405	376
IPR001611	Leucine-rich repeat	385	543
IPR016024	Armadillo-type fold	369	269
IPR001810	F-box domain	336	526
IPR012677	Nucleotide-binding alpha-beta plait domain	313	272
IPR015943	WD40/YVTN repeat-like-containing domain	308	239
IPR001005	SANT/Myb domain	294	265
IPR029063	S-adenosyl-L-methionine-dependent methyltransferase	294	267
IPR003593	AAA+ ATPase domain	289	243
IPR017986	WD40-repeat-containing domain	287	216
IPR011989	Armadillo-like helical	282	222
IPR000504	RNA recognition motif domain	277	246
IPR013210	Leucine-rich repeat-containing N-terminal, plant-type	271	277
IPR020846	Major facilitator superfamily domain	269	274
IPR001680	WD40 repeat	265	201
IPR012336	Thioredoxin-like fold	265	262
IPR001128	Cytochrome P450	256	482
IPR002182	NB-ARC	250	435
IPR017930	Myb domain	245	217
IPR012337	Ribonuclease H-like domain	237	192
IPR002401	Cytochrome P450, E-class, group I	236	408
IPR003591	Leucine-rich repeat, typical subtype	235	339
IPR017972	Cytochrome P450, conserved site	228	363
IPR017853	Glycoside hydrolase superfamily	212	204
IPR027443	Isopenicillin N synthase-like	199	215
IPR011992	EF-hand domain pair	198	176
IPR016177	DNA-binding domain	186	230
IPR005123	Oxoglutarate/iron-dependent dioxygenase	185	190
IPR011598	Myc-type, basic helix-loop-helix (bHLH) domain	180	163
IPR002048	EF-hand domain	177	159
IPR011991	Winged helix-turn-helix DNA-binding domain	174	166
IPR013781	Glycoside hydrolase, catalytic domain	173	166
IPR001471	AP2/ERF domain	173	219
IPR002213	UDP-glucuronosyl/UDP-glucosyltransferase	172	268
IPR012340	Nucleic acid-binding, OB-fold	171	153
IPR012334	Pectin lyase fold	164	154
IPR011050	Pectin lyase fold/virulence factor	163	154
IPR014001	Helicase superfamily 1/2, ATP-binding domain	162	106
IPR018247	EF-Hand 1, calcium-binding site	162	147
IPR003653	Ulp1 protease family, C-terminal catalytic domain	161	46
IPR007087	Zinc finger, C2H2	161	161
IPR023214	HAD-like domain	160	146
IPR001650	Helicase, C-terminal	156	107
IPR019775	WD40 repeat, conserved site	150	119
IPR007527	Zinc finger, SWIM-type	148	17
IPR003439	ABC transporter-like	144	117
IPR026992	Non-haem dioxygenase N-terminal domain	143	154
IPR029044	Nucleotide-diphospho-sugar transferases	142	125
IPR009072	Histone-fold	133	118
IPR006564	Zinc finger, PMZ-type	132	13
IPR015424	Pyridoxal phosphate-dependent transferase	131	139
IPR013026	Tetratricopeptide repeat-containing domain	129	109
IPR013830	SGNH hydrolase-type esterase domain	128	113
IPR003959	ATPase, AAA-type, core	126	101
IPR020683	Ankyrin repeat-containing domain	126	116

IPR000008	C2 domain	125	102
IPR010255	Haem peroxidase	125	129
IPR015421	Pyridoxal phosphate-dependent transferase, major region, subdomain 1	125	130
IPR014729	Rossmann-like alpha/beta/alpha sandwich fold	123	99
IPR008972	Cupredoxin	122	116
IPR005225	Small GTP-binding protein domain	121	109
IPR019734	Tetratricopeptide repeat	120	98
IPR029071	Ubiquitin-related domain	119	110
IPR002016	Haem peroxidase, plant/fungal/bacterial	119	123
IPR001878	Zinc finger, CCHC-type	119	177
IPR014710	RmlC-like jelly roll fold	118	127
IPR011011	Zinc finger, FYVE/PHD-type	116	88
IPR002110	Ankyrin repeat	115	104
IPR013785	Aldolase-type TIM barrel	113	100
IPR001623	DnaJ domain	113	105
IPR015300	DNA-binding pseudobarrel domain	113	110
IPR019557	Aminotransferase-like, plant mobile domain	112	59
IPR023213	Chloramphenicol acetyltransferase-like domain	112	156
IPR016140	Bifunctional inhibitor/plant lipid transfer protein/seed storage helical domain	111	119
IPR001932	PPM-type phosphatase domain	110	95
IPR001356	Homeobox domain	109	101
IPR006447	Myb domain, plants	108	98
IPR023393	START-like domain	108	132
IPR000823	Plant peroxidase	107	108
IPR003676	Small auxin-up RNA	106	144
IPR002100	Transcription factor, MADS-box	105	157
IPR003340	B3 DNA binding domain	103	96
IPR021109	Aspartic peptidase domain	103	110
IPR003441	NAC domain	102	116
IPR015410	Domain of unknown function DUF1985	101	28
IPR002198	Short-chain dehydrogenase/reductase SDR	101	104
IPR002347	Glucose/ribitol dehydrogenase	101	105
IPR003480	Transferase	101	150
IPR017451	F-box associated interaction domain	101	255
IPR020472	G-protein beta WD-40 repeat	100	78
IPR019793	Peroxidases heam-ligand binding site	100	97
IPR011333	POZ domain	90	108
IPR006501	Pectinesterase inhibitor domain	87	103
IPR025558	Domain of unknown function DUF4283	60	135

Table 9. Summary of protein domains common to both species detected by scanning the IntertPro database. All the domains with at least 100 occurrences in Tomato or Potato are listed.

It’s interesting to note how some domains have much more occurrences in a species compared to the other. It’s the case of two zinc-finger domain, the SWIM-type and the PMZ-type, which have 148 occurrences in Tomato and 17 occurrences in Potato, and 132 occurrences in Tomato and 13 occurrences in Potato, respectively (Tab. 9).

If we focus on the proteins domains exclusively detected in one species rather than the other, we notice different type of transposases or transposons that are present exclusively in Tomato and not in Potato, or vice versa (Tab. 10 and 11).

InterPro ID	Description	N
IPR018289	MULE transposase domain	119
IPR004332	Transposase, MuDR, plant	69
IPR004252	Probable transposase, Ptta/En/Spm, plant	24
IPR005162	Retrotransposon gag domain	15
IPR030386	GB1/RHD3-type guanine nucleotide-binding (G) domain	12
IPR006912	Harbinger transposase-derived protein	10
IPR015894	Guanylate-binding protein, N-terminal	10
IPR005798	Cytochrome b/b6, C-terminal	8
IPR011759	Cytochrome C oxidase subunit II, transmembrane domain	8
IPR019645	Uncharacterised protein family Ycf15	8
IPR000515	ABC transporter type 1, transmembrane domain MetI-like	7
IPR005559	CG-1 DNA-binding domain	7
IPR016151	DNA mismatch repair protein MutS, N-terminal	6
IPR001457	NADH:ubiquinone/plastoquinone oxidoreductase, chain 6	5
IPR007836	Ribosomal protein L41	5
IPR017452	GPCR, rhodopsin-like, 7TM	5
IPR024733	Alpha-N-acetylglucosaminidase, tim-barrel domain	5
IPR000409	BEACH domain	4
IPR001463	Sodium:alanine symporter	4
IPR003359	Photosystem I Ycf4, assembly	4
IPR006133	DNA-directed DNA polymerase, family B, exonuclease domain	4
IPR006134	DNA-directed DNA polymerase, family B, multifunctional domain	4
IPR009543	Vacuolar protein sorting-associated protein 13, SHR-binding domain	4
IPR014012	Helicase/SANT-associated domain	4
IPR016961	Diacylglycerol kinase, plant	4
IPR022546	Uncharacterised protein family Ycf68	4
IPR023211	DNA polymerase, palm domain	4
IPR028275	Clustered mitochondria protein, N-terminal	4

Table 10. Summary of protein domains exclusively detected in Tomato. All the domains with at least 4 occurrences are shown.

InterPro ID	Description	N
IPR004242	Transposon, En/Spm-like	12
IPR029466	No apical meristem-associated, C-terminal domain	5
IPR002397	Cytochrome P450, B-class	3
IPR011773	DNA-directed RNA polymerase, alpha subunit	3
IPR025314	Domain of unknown function DUF4219	3
IPR012171	Fatty acid desaturase	3

Table 11. Summary of protein domains exclusively detected in Potato. All the domains with at least 3 occurrences are shown.

5.3.5 Metabolic pathways and enzyme classification

In order to highlight common or peculiar metabolic features of the compared species, we performed sequence similarity searches between Tomato, Potato and the entire Swiss-Prot protein collection, identifying the enzyme-coding genes of both Solanaceae. The represented Venn diagram of the detected enzymatic classes (Fig. 30) shows that 31 and 17 of them were detected exclusively in Tomato and Potato, respectively. The most represented enzymatic class exclusively detected in Tomato belongs to the transferases (Tab. 12), while in Potato belongs to the oxidoreductases (Tab. 13).

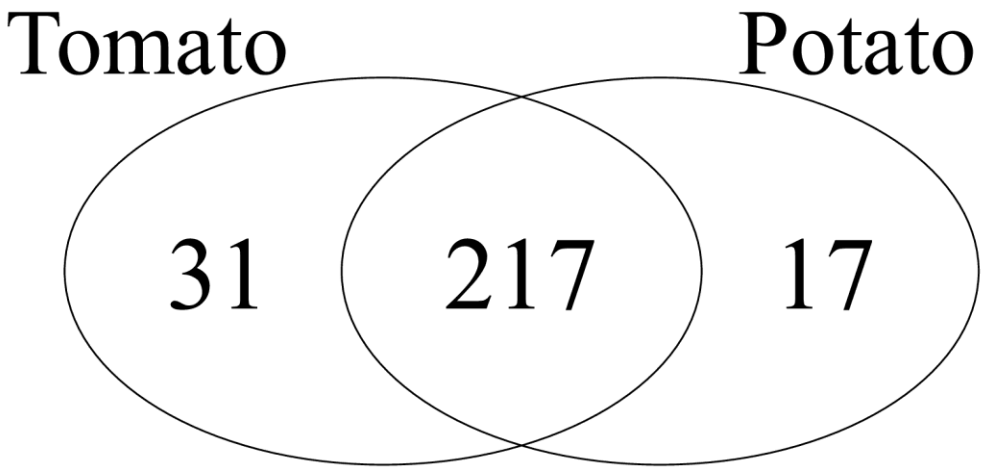


Figure 30. Venn diagram showing peculiar or common enzymatic classes detected in Tomato and Potato.

EC NUMBER	CLASS
1.2.1.41	-
2.3.1.-	-
2.3.1.12	Transferases; Acyltransferases; Transferring groups other than aminoacyl groups
2.4.1.18	Transferases; Glycosyltransferases; Hexosyltransferases
2.4.2.9	Transferases; Glycosyltransferases; Pentosyltransferases
2.5.1.21	Transferases; Transferring alkyl or aryl groups, other than methyl groups
2.5.1.28	Transferases; Transferring alkyl or aryl groups, other than methyl groups
2.5.1.46	Transferases; Transferring alkyl or aryl groups, other than methyl groups
2.5.1.92	Transferases; Transferring alkyl or aryl groups, other than methyl groups
2.6.1.85	Transferases; Transferring nitrogenous groups; Transaminases
2.7.10.1	-
2.7.2.11	Transferases; Transferring phosphorus-containing groups; Phosphotransferases
2.7.7.41	Transferases; Transferring phosphorus-containing groups; Nucleotidyltransferases
2.8.1.9	Transferases; Transferring sulfur-containing groups; Sulfurtransferases
3.1.27.1	Hydrolases; Acting on ester bonds; Endoribonucleases producing 3'-phosphomonoesters
3.2.1.15	Hydrolases; Glycosylases; Glycosidases
3.2.1.78	Hydrolases; Glycosylases; Glycosidases
3.4.21.92	Hydrolases; Acting on peptide bonds (peptidases); Serine endopeptidases
3.5.1.-	-
3.5.4.2	Hydrolases; Acting on carbon-nitrogen bonds, other than peptide bonds; In cyclic amidines
3.6.1.5	Hydrolases; Acting on acid anhydrides; In phosphorus-containing anhydrides
4.1.1.22	Lyases; Carbon-carbon lyases; Carboxy-lyases
4.2.3.105	-
4.2.3.117	-
4.2.3.15	-
4.2.3.16	-
5.3.1.9	Isomerases; Intramolecular oxidoreductases; Interconverting aldoses and ketoses
5.4.2.2	Isomerases; Intramolecular transferases; Phosphotransferases (phosphomutases)
5.99.1.3	Isomerases; Other isomerases
6.1.1.17	Ligases; Forming carbon-oxygen bonds; Ligases forming aminoacyl-tRNA
6.4.1.2	Ligases; Forming carbon-carbon bonds; Ligases that form carbon-carbon bonds

Table 12. Summary of enzymatic classes exclusively detected in Tomato.

EC NUMBER	CLASS
1.1.1.284	Oxidoreductases; Acting on the CH-OH group of donors
1.1.1.44	Oxidoreductases; Acting on the CH-OH group of donors
1.10.3.9	Oxidoreductases; Acting on diphenols and related substances as donors
1.14.13.11	Oxidoreductases; Acting on paired donors
1.14.13.129	Oxidoreductases; Acting on paired donors
1.14.13.88	Oxidoreductases; Acting on paired donors
1.23.5.1	Oxidoreductases; Reducing C-O-C group as acceptor
1.97.1.12	Oxidoreductases; Other oxidoreductases
2.1.1.127	Transferases; Transferring one-carbon groups; Methyltransferases
2.1.1.68	Transferases; Transferring one-carbon groups; Methyltransferases
2.3.1.61	Transferases; Acyltransferases; Transferring groups other than aminoacyl groups
2.7.11.25	Transferases; Transferring phosphorus-containing groups; Protein-serine/threonine kinases
2.7.6.1	Transferases; Transferring phosphorus-containing groups; Diphosphotransferases
3.1.1.-	-
4.1.2.13	Lyases; Carbon-carbon lyases; Aldehyde-lyases
4.3.3.7	Lyases; Carbon-nitrogen lyases; Amine-lyases
5.3.1.23	Isomerases; Intramolecular oxidoreductases; Interconverting aldoses and ketoses

Table 13. Summary of enzymatic classes exclusively detected in Potato.

Moreover we exploited the KEGG database (Kanehisa and Goto 2000) to investigate about the metabolic pathways associated to the previously detected enzymatic classes. In this way, we were able to detect pathways that contain enzymatic classes exclusively detected in one species (Tab 14), assuming the existence of some metabolic pathways preferentially activated in Tomato rather than in Potato, and vice versa.

As an example, the monoterpenoid biosynthesis, associated to some enzymatic classes detected exclusively in Tomato (Tab. 14), is a metabolic pathway that leads to the production of some plant secondary metabolites, which belong to a large family of compounds of valuable applications in medicine and cosmetics (Oswald et al. 2007).

KEGG_ID	PATHWAY DESCRIPTION	SPECIES
ec00254	Aflatoxin biosynthesis	Tomato
ec00332	Carbapenem biosynthesis	Tomato
ec00642	Ethylbenzene degradation	Tomato
ec00760	Nicotinate and nicotinamide metabolism	Tomato
ec00902	Monoterpenoid biosynthesis	Tomato
ec00300	Lysine biosynthesis	Potato
ec00534	Glycosaminoglycan biosynthesis - heparan sulfate / heparin	Potato

Table 14. Summary of metabolic pathways exclusively detected in Tomato or Potato.

5.3.6 A 3-species comparison between Tomato, Potato and Grapevine

In order to simultaneously provide a roughly overview of the conserved or specific functional features and metabolic traits of the species analyzed in the fourth and fifth chapter, we integrated the information of the protein domain analyses and the enzymatic and pathway analyses obtained from the Tomato-Grapevine comparison and from the Tomato-Potato comparison, producing a 3-species comparative analysis.

Figure 31 shows a distribution of the most common protein domains shared between the three species, mainly including different kind of protein kinases. Further analyses will clarify if these domains have a similar distribution in other than plants eukaryotes, or if they are preferentially more abundant in plants.

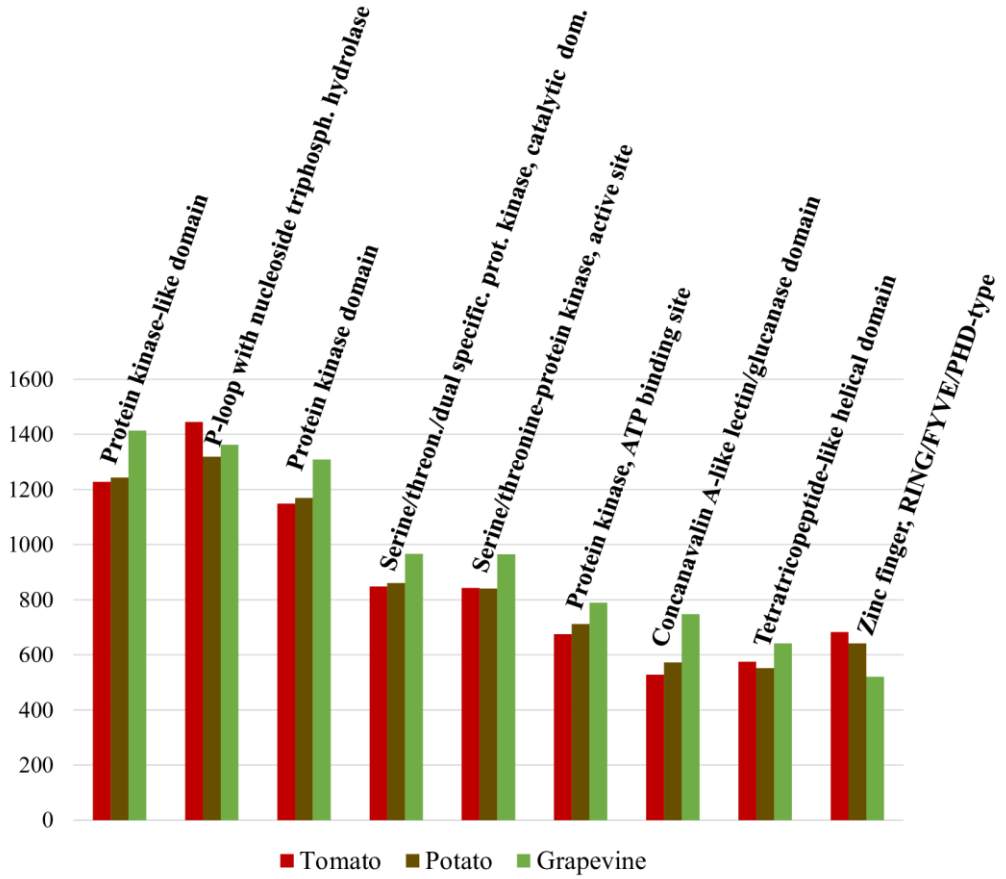


Figure 31. Distribution of the most common protein domains shared between Tomato, Potato and Grapevine. All the domains with at least 500 occurrences in each species are shown.

Focusing on domains exclusively working in one of the three compared species, we are probably looking at peculiar molecular functions of Tomato, Potato and Grapevine. As an example, the Harbinger transposase-derived protein domain has been exclusively detected in Tomato (Tab. 15). The majority of the members of this family are from plants and have an hydrolase activity, acting on ester bonds. An in-depth targeted analysis on this protein could reveal the functional role that this domain plays exclusively in Tomato.

InterPro ID	Description	N
IPR006912	Harbinger transposase-derived protein	10
IPR005798	Cytochrome b/b6, C-terminal	8
IPR019645	Uncharacterised protein family Ycf15	8
IPR000515	ABC transporter type 1, transmembrane domain MetI-like	7
IPR001457	NADH:ubiquinone/plastoquinone oxidoreductase, chain 6	5
IPR007836	Ribosomal protein L41	5
IPR017452	GPCR, rhodopsin-like, 7TM	5
IPR001463	Sodium:alanine symporter	4
IPR022546	Uncharacterised protein family Ycf68	4
IPR004242	Transposon, En/Spm-like	12
IPR029466	No apical meristem-associated, C-terminal domain	5
IPR011713	Leucine-rich repeat 3	15
IPR005830	Aerolysin	4
IPR008998	Agglutinin domain	4
IPR023307	Aerolysin-like toxin, beta complex domain	4

Table 15. Summary of protein domains exclusively detected in Tomato (in red), Potato (in brown) and Grapevine (in green). All the domains with at least 4 occurrences are shown.

If we further look at the pathways containing enzymatic classes exclusively detected in one of the three compared species, we may infer about metabolic features preferentially activated in one species rather than the others. A general overview of the integrated metabolic information from Tomato, Potato and Grapevine (Fig. 32), shows that, beside a numerous group of conserved pathways (87) within the three species and 44 pathways shared between the two Solanaceae species, three pathways contain enzymes detected exclusively in Tomato, and two pathways contain enzymes detected exclusively in Potato.

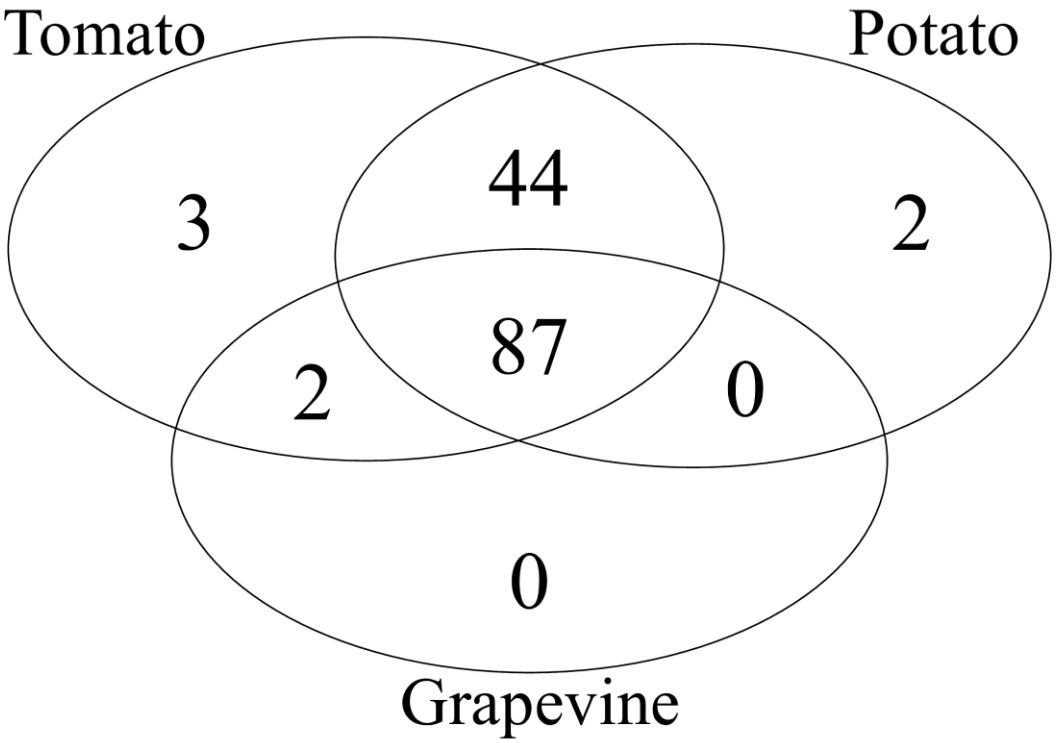


Figure 32. Venn diagram showing peculiar or common metabolic pathways detected in Tomato, Potato and Grapevine.

Interestingly two metabolic pathways, i.e. the Aflatoxin and the Monoterpenoid biosynthesis, were detected exclusively in Tomato and Grapevine, highlighting that two distantly-related species may show common metabolic features, which two members of the same family, namely the Solanaceae, do not share.

Table 16 lists the pathways containing enzymes detected exclusively in one of the three compared species. The Carbapenem biosynthesis (Fig. 33), as an example, is a metabolic pathway containing enzymatic classes detected exclusively in Tomato. Carbapenems are antibiotics used for the treatment of infections caused by multidrug-resistant (MDR) bacteria. They are members of the beta lactam class of antibiotics, as well as the penicillins and cephalosporins, which kill bacteria by inhibiting the cell wall synthesis. This class of secondary metabolites, exhibiting a broader spectrum of activity compared to

cephalosporins and penicillins, attracts particular attention by the scientific community. Moreover their effectiveness is less affected by the mechanisms of antibiotic resistance than other beta lactams (Meletis 2016).

KEGG ID	Description
ec00332	Carbapenem biosynthesis
ec00642	Ethylbenzene degradation
ec00760	Nicotinate and nicotinamide metabolism
ec00300	Lysine biosynthesis
ec00534	Glycosaminoglycan biosynthesis - heparan sulfate / heparin

Table 16. Metabolic pathway exclusively detected in each of the three species. Tomato pathways are shown in red, Potato pathways are shown in brown. No specific Grapevine pathways were detected.

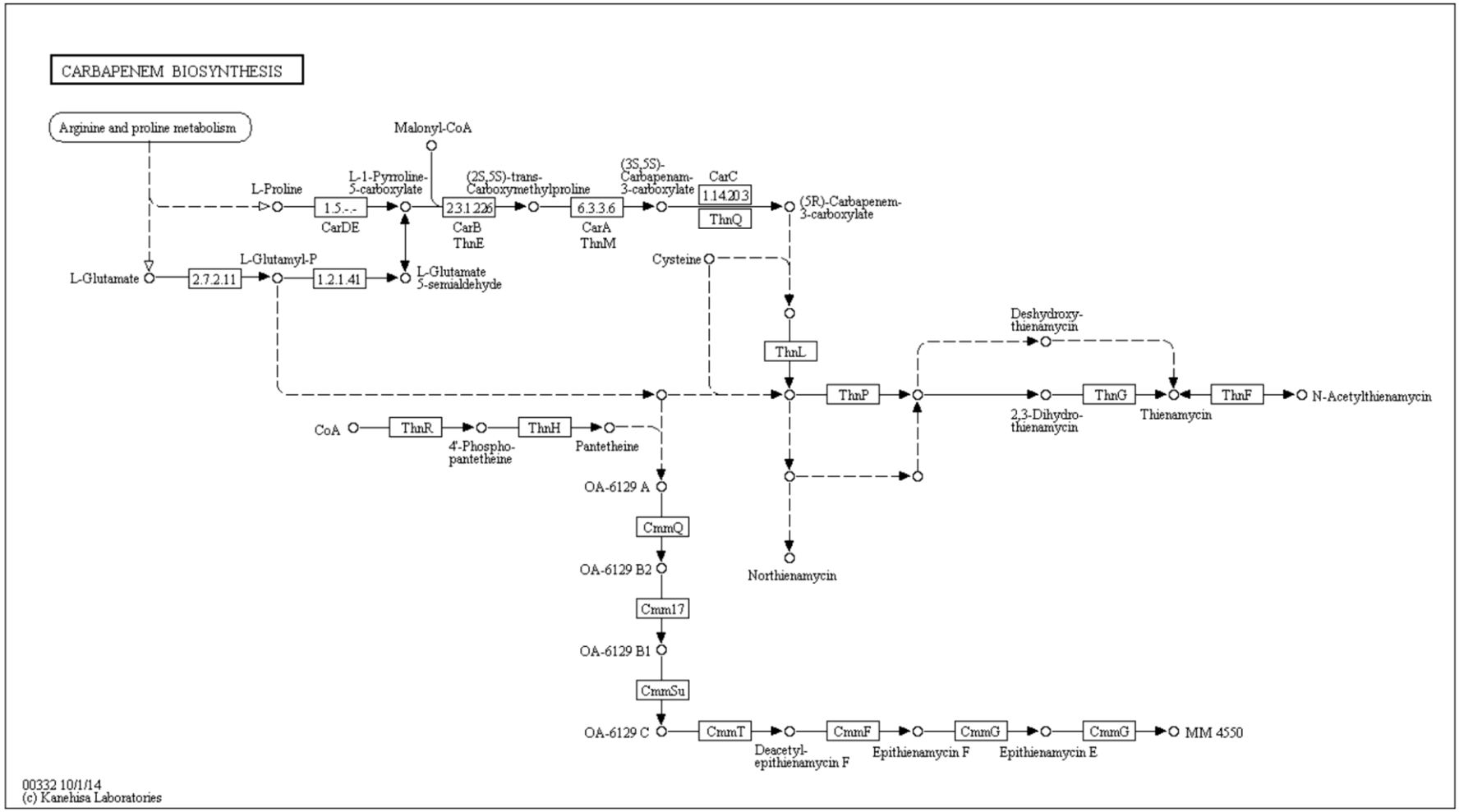


Figure 33. Diagram of the Carbapenem biosynthesis pathway. Image extracted from KEGG database (available at http://www.genome.jp/kegg/).

5.4 Conclusions

Tomato and Potato, two closely related species belonging to the family of Solanaceae, presented a high level of similarity between their genomic features, as shown by the presence of a strong consensus of orthologs genes. Accordingly to a multi-level procedure that considers three different methods, more than 80% of the gene, mRNA and protein content of both species resulted to have an ortholog counterpart.

Moreover, we built networks of paralog genes for Tomato and Potato, connected by orthology relationships, to investigate about the organization and the evolution of gene families in both Solanaceae species. By this approach, we predicted gene families of one species that underwent an expansion or a reduction in the number of their elements when compared to the other species. Furthermore, a deeper analysis of such networks allows the identification of cases in which genes with a well-known function are related by a homology relationship to genes with an unknown function, enabling the transfer of the information relating to the gene annotation.

In order to detect the specific genes of Tomato and Potato, all the genes that share even a low sequence similarity level with the networks of orthologs/paralogs previously detected were filtered out from the entire gene collections of both species.

We detected the metabolic pathways and the enzymatic classes associated to the Tomato and Potato genomes, with the aim of roughly compare the two Solanaceae species at metabolic level. The presence of some peculiar enzymes and preferential metabolic pathways was inferred in both species. Moreover, a general overview of a 3-species cross-comparison, considering a protein domain and metabolic pathway characterization of Tomato, Potato and Grapevine, was provided.

This chapter highlighted how closely related species of the same family, although showing a strong similarity at genomic level with a high number of conserved genes that preside similar functions and regulative mechanisms, own distinctive genes that give to each organism its peculiar features.

Chapter 6. Summary and conclusions

In this thesis, we focused the attention on the biological relevance of orthology and paralogy relationships, as major forces that drive evolutionary speciation events and gene function innovation, respectively. Comparative genomics strategies massively use orthologs and paralogs detection techniques for purposes of gene function prediction, transferring the biological information from model species to less known organisms. In this context, it is crucial that the available resources from reference organisms, such as *Arabidopsis thaliana* among plants, are properly validated and organized for a suitable exploitation by the science community. In the second chapter, we presented a web-accessible resource, developed with my contribution in the laboratory where I worked during the last three years, which aims to efficiently explore the single-copy genes and the paralogs, organized into networks, of the model *A. thaliana*, for comparative genomics approaches or gene families investigations. The network organization provides an immediate access to the paralogy information, supporting the unraveling of the complexity of the Arabidopsis genome. Moreover, the case study of acetyltransferases family described in the second chapter showed how the use of different E-value threshold for the definition of paralogy relationships favors the investigation of gene family organization and splitting.

Orthologs prediction has always been based on protein sequence similarity searches, even though these type of analysis can lead to errors due to the lack of a comprehensive definition of such sequences in recently sequenced organisms with a still preliminary annotation. In the third chapter, we presented a methodology for predicting orthologs between two species by comparing mRNA sequences instead of proteins sequences. The developed algorithm, based on the widespread Bidirectional Best Hit approach, includes a routine able

to improve the quality of the alignment, in order to avoid that a mutation event, such as the insertion or deletion of one or two bases, may result in a loss of the real information content of a sequence similarity search. We showed that in many cases the orthologs detected using transcriptomic data have higher scoring, probably taking advantages of reconstructed alignments that include also regions with different coding frame. In the era of fast genome and transcript sequencing, draft gene annotations are often released without consistently human curation. Although these efforts are usually supported by incredible enrichment of transcriptome datasets, the proteome complement is still limited and alternative approaches for ortholog detection may lead to more reliable results.

The developed tools and the acquired knowledge were applied to the comparative analyses presented in chapters four and five. The cross comparison between two distantly related fleshy fruit species, such as tomato and grapevine, presented in the fourth chapter, showed a significant amount of common features between these members of two different clades, i.e. the Asterids and the Rosids. In the fifth chapter, instead, we investigated about the differences and similarities between potato and tomato, two closely related species of Solanaceae, an economically relevant family among plants.

In the first step of both comparisons, we detected a strong core of orthology relationships defined as the consensus of three different methods based on gene, mRNAs and protein sequences, respectively. Moreover, networks of ortholog/paralog genes of the compared species were created, with the aim of investigating about the organization and evolution of gene families within different species. By this approach, it was possible to detect gene families of one species undergoing an expansion or a reduction in the other species, and vice versa. In addition, the analysis of such networks allowed to detect cases in which genes relating to a given gene family are closely connected by relationships of orthology/paralogy to genes of unknown or incomplete function

annotation, enabling the transfer of the information concerning the considered gene family. In the next step of the comparison, we predicted the distinctive genes of the compared species. The species-specific genes were detected filtering out from the complete gene collections of both organisms, all the genes, mRNAs and proteins that have at least one ortholog counterpart, or that share at least one paralogy relation detected starting exclusively from an ortholog pair, even at a loose E-value threshold (10^{-3}). A further exhaustive study and characterization of these distinctive genes will highlight unique features of the compared species. Finally, in the last steps of the comparative analyses, a roughly comparison of the protein domains and the metabolic features highlighted the presence of distinctive function and peculiar enzymatic classes associated to preferential pathways in each of the investigated species.

Annex A

EC NUMBER	CLASS
1.-.-.-	-
1.1.1.-	-
1.1.1.195	Oxidoreductases; Acting on the CH-OH group of donors; With NAD+ or NADP+ as acceptor
1.1.1.236	Oxidoreductases; Acting on the CH-OH group of donors; With NAD+ or NADP+ as acceptor
1.1.1.39	Oxidoreductases; Acting on the CH-OH group of donors; With NAD+ or NADP+ as acceptor
1.1.1.49	Oxidoreductases; Acting on the CH-OH group of donors; With NAD+ or NADP+ as acceptor
1.1.1.85	Oxidoreductases; Acting on the CH-OH group of donors; With NAD+ or NADP+ as acceptor
1.10.2.2	Oxidoreductases; Acting on diphenols and related substances as donors; With a cytochrome as acceptor
1.10.3.-	-
1.10.3.1	Oxidoreductases; Acting on diphenols and related substances as donors; With oxygen as acceptor
1.10.9.1	Oxidoreductases; Acting on diphenols and related substances as donors; With a copper protein as acceptor
1.11.1.-	-
1.11.1.12	Oxidoreductases; Acting on a peroxide as acceptor; Peroxidases
1.11.1.15	Oxidoreductases; Acting on a peroxide as acceptor; Peroxidases
1.13.11.12	Oxidoreductases; Acting on single donors with incorporation of molecular oxygen (oxygenases)
1.13.11.58	Oxidoreductases; Acting on single donors with incorporation of molecular oxygen (oxygenases)
1.14.-.-	-
1.14.11.23	Oxidoreductases; Acting on paired donors, with incorporation or reduction of molecular oxygen
1.14.13.121	Oxidoreductases; Acting on paired donors, with incorporation or reduction of molecular oxygen
1.14.13.90	Oxidoreductases; Acting on paired donors, with incorporation or reduction of molecular oxygen
1.14.17.4	Oxidoreductases; Acting on paired donors, with incorporation or reduction of molecular oxygen
1.14.19.2	Oxidoreductases; Acting on paired donors, with incorporation or reduction of molecular oxygen
1.14.21.6	Oxidoreductases; Acting on paired donors, with incorporation or reduction of molecular oxygen
1.17.4.1	Oxidoreductases; Acting on CH or CH2 groups; With a disulfide as acceptor
1.2.1.2	Oxidoreductases; Acting on the aldehyde or oxo group of donors; With NAD+ or NADP+ as acceptor
1.2.1.41	-
1.2.4.1	Oxidoreductases; Acting on the aldehyde or oxo group of donors; With a disulfide as acceptor
1.3.1.-	-
1.3.1.105	Oxidoreductases; Acting on the CH-CH group of donors; With NAD+ or NADP+ as acceptor
1.3.1.22	Oxidoreductases; Acting on the CH-CH group of donors; With NAD+ or NADP+ as acceptor
1.3.1.42	Oxidoreductases; Acting on the CH-CH group of donors; With NAD+ or NADP+ as acceptor
1.3.1.83	Oxidoreductases; Acting on the CH-CH group of donors; With NAD+ or NADP+ as acceptor
1.3.3.4	Oxidoreductases; Acting on the CH-CH group of donors; With oxygen as acceptor
1.3.5.5	Oxidoreductases; Acting on the CH-CH group of donors; With a quinone or related compound as acceptor
1.3.5.6	Oxidoreductases; Acting on the CH-CH group of donors; With a quinone or related compound as acceptor
1.3.8.4	Oxidoreductases; Acting on the CH-CH group of donors; With a flavin as acceptor
1.3.99.-	-
1.3.99.12	Oxidoreductases; Acting on the CH-CH group of donors; With unknown physiological acceptors
1.4.4.2	Oxidoreductases; Acting on the CH-NH2 group of donors; With a disulfide as acceptor
1.6.3.-	-
1.6.5.4	Oxidoreductases; Acting on NADH or NADPH; With a quinone or similar compound as acceptor
1.6.5.9	Oxidoreductases; Acting on NADH or NADPH; With a quinone or similar compound as acceptor
1.7.1.1	Oxidoreductases; Acting on other nitrogenous compounds as donors; With NAD+ or NADP+ as acceptor
1.8.-.-	-
1.8.1.7	Oxidoreductases; Acting on a sulfur group of donors; With NAD+ or NADP+ as acceptor
1.8.4.11	Oxidoreductases; Acting on a sulfur group of donors; With a disulfide as acceptor
1.8.7.1	Oxidoreductases; Acting on a sulfur group of donors; With an iron-sulfur protein as acceptor
1.9.3.1	Oxidoreductases; Acting on a heme group of donors; With oxygen as acceptor
2.2.1.6	Transferases; Transferring aldehyde or ketonic groups; Transketolases and transaldolases
2.2.1.7	Transferases; Transferring aldehyde or ketonic groups; Transketolases and transaldolases
2.3.1.-	-
2.3.1.12	Transferases; Acyltransferases; Transferring groups other than aminoacyl groups
2.3.1.133	Transferases; Acyltransferases; Transferring groups other than aminoacyl groups
2.3.3.13	Transferases; Acyltransferases; Acyl groups converted into alkyl groups on transfer
2.3.3.16	Transferases; Acyltransferases; Acyl groups converted into alkyl groups on transfer
2.4.1.1	Transferases; Glycosyltransferases; Hexosyltransferases
2.4.1.123	Transferases; Glycosyltransferases; Hexosyltransferases
2.4.1.13	Transferases; Glycosyltransferases; Hexosyltransferases
2.4.1.14	Transferases; Glycosyltransferases; Hexosyltransferases
2.4.1.18	Transferases; Glycosyltransferases; Hexosyltransferases
2.4.1.207	Transferases; Glycosyltransferases; Hexosyltransferases
2.4.1.21	Transferases; Glycosyltransferases; Hexosyltransferases
2.4.1.242	Transferases; Glycosyltransferases; Hexosyltransferases
2.4.1.25	Transferases; Glycosyltransferases; Hexosyltransferases
2.4.2.10	Transferases; Glycosyltransferases; Pentosyltransferases

2.4.2.9	Transferases; Glycosyltransferases; Pentosyltransferases
2.5.1.18	Transferases; Transferring alkyl or aryl groups, other than methyl groups
2.5.1.19	Transferases; Transferring alkyl or aryl groups, other than methyl groups
2.5.1.21	Transferases; Transferring alkyl or aryl groups, other than methyl groups
2.5.1.28	Transferases; Transferring alkyl or aryl groups, other than methyl groups
2.5.1.32	Transferases; Transferring alkyl or aryl groups, other than methyl groups
2.5.1.43	Transferases; Transferring alkyl or aryl groups, other than methyl groups
2.5.1.46	Transferases; Transferring alkyl or aryl groups, other than methyl groups
2.5.1.54	Transferases; Transferring alkyl or aryl groups, other than methyl groups
2.5.1.58	Transferases; Transferring alkyl or aryl groups, other than methyl groups
2.5.1.59	Transferases; Transferring alkyl or aryl groups, other than methyl groups
2.5.1.92	Transferases; Transferring alkyl or aryl groups, other than methyl groups
2.6.1.1	Transferases; Transferring nitrogenous groups; Transaminases
2.6.1.78	Transferases; Transferring nitrogenous groups; Transaminases
2.6.1.79	Transferases; Transferring nitrogenous groups; Transaminases
2.6.1.85	Transferases; Transferring nitrogenous groups; Transaminases
2.6.1.9	Transferases; Transferring nitrogenous groups; Transaminases
2.6.1.96	Transferases; Transferring nitrogenous groups; Transaminases
2.7.1.-	-
2.7.1.1	Transferases; Transferring phosphorus-containing groups; Phosphotransferases with an alcohol group
2.7.1.148	Transferases; Transferring phosphorus-containing groups; Phosphotransferases with an alcohol group
2.7.1.4	Transferases; Transferring phosphorus-containing groups; Phosphotransferases with an alcohol group
2.7.1.71	Transferases; Transferring phosphorus-containing groups; Phosphotransferases with an alcohol group
2.7.1.90	Transferases; Transferring phosphorus-containing groups; Phosphotransferases with an alcohol group
2.7.13.3	Transferases; Transferring phosphorus-containing groups; Protein-histidine kinases
2.7.2.11	Transferases; Transferring phosphorus-containing groups; Phosphotransferases with a carboxy group as acceptor
2.7.4.6	Transferases; Transferring phosphorus-containing groups; Phosphotransferases with a phosphate group
2.7.7.41	Transferases; Transferring phosphorus-containing groups; Nucleotidyltransferases
2.7.7.9	Transferases; Transferring phosphorus-containing groups; Nucleotidyltransferases
2.7.9.4	Transferases; Transferring phosphorus-containing groups; Phosphotransferases with paired acceptors
2.8.1.9	Transferases; Transferring sulfur-containing groups; Sulfurtransferases
3.1.1.11	Hydrolases; Acting on ester bonds; Carboxylic-ester hydrolases
3.1.27.1	Hydrolases; Acting on ester bonds; Endoribonucleases producing 3'-phosphomonoesters
3.1.3.2	Hydrolases; Acting on ester bonds; Phosphoric-monoester hydrolases
3.1.3.24	Hydrolases; Acting on ester bonds; Phosphoric-monoester hydrolases
3.1.3.25	Hydrolases; Acting on ester bonds; Phosphoric-monoester hydrolases
3.1.4.4	Hydrolases; Acting on ester bonds; Phosphoric-diester hydrolases
3.2.1.23	Hydrolases; Glycosylases; Glycosidases, i.e. enzymes that hydrolyse O- and S-glycosyl compounds
3.2.1.26	Hydrolases; Glycosylases; Glycosidases, i.e. enzymes that hydrolyse O- and S-glycosyl compounds
3.2.1.78	Hydrolases; Glycosylases; Glycosidases, i.e. enzymes that hydrolyse O- and S-glycosyl compounds
3.4.11.1	Hydrolases; Acting on peptide bonds (peptidases); Aminopeptidases
3.4.11.5	Hydrolases; Acting on peptide bonds (peptidases); Aminopeptidases
3.4.21.92	Hydrolases; Acting on peptide bonds (peptidases); Serine endopeptidases
3.4.22.-	-
3.4.24.64	Hydrolases; Acting on peptide bonds (peptidases); Metalloendopeptidases
3.5.1.-	-
3.5.1.53	Hydrolases; Acting on carbon-nitrogen bonds, other than peptide bonds; In linear amides
3.5.1.88	Hydrolases; Acting on carbon-nitrogen bonds, other than peptide bonds; In linear amides
3.5.4.16	Hydrolases; Acting on carbon-nitrogen bonds, other than peptide bonds; In cyclic amidines
3.5.4.2	Hydrolases; Acting on carbon-nitrogen bonds, other than peptide bonds; In cyclic amidines
3.5.5.1	Hydrolases; Acting on carbon-nitrogen bonds, other than peptide bonds; In nitriles
3.5.5.4	Hydrolases; Acting on carbon-nitrogen bonds, other than peptide bonds; In nitriles
3.6.1.23	Hydrolases; Acting on acid anhydrides; In phosphorus-containing anhydrides
3.6.1.5	Hydrolases; Acting on acid anhydrides; In phosphorus-containing anhydrides
3.6.3.8	Hydrolases; Acting on acid anhydrides; catalyse transmembrane movement of subst.
4.1.1.1	Lyases; Carbon-carbon lyases; Carboxy-lyases
4.1.1.15	Lyases; Carbon-carbon lyases; Carboxy-lyases
4.1.1.17	Lyases; Carbon-carbon lyases; Carboxy-lyases
4.1.1.19	Lyases; Carbon-carbon lyases; Carboxy-lyases
4.1.1.22	Lyases; Carbon-carbon lyases; Carboxy-lyases
4.1.1.23	Lyases; Carbon-carbon lyases; Carboxy-lyases
4.1.1.37	Lyases; Carbon-carbon lyases; Carboxy-lyases
4.1.1.39	Lyases; Carbon-carbon lyases; Carboxy-lyases
4.1.1.50	Lyases; Carbon-carbon lyases; Carboxy-lyases
4.2.1.1	Lyases; Carbon-oxygen lyases; Hydro-lyases
4.2.1.121	Lyases; Carbon-oxygen lyases; Hydro-lyases
4.2.1.65	Lyases; Carbon-oxygen lyases; Hydro-lyases
4.2.2.2	Lyases; Carbon-oxygen lyases; Acting on polysaccharides
4.2.3.1	Lyases; Carbon-oxygen lyases; Acting on phosphates
4.2.3.105	-
4.2.3.117	-
4.2.3.15	-
4.2.3.16	-

4.2.3.21	Lyases; Carbon-oxygen lyases; Acting on phosphates
4.2.3.4	Lyases; Carbon-oxygen lyases; Acting on phosphates
4.2.3.5	Lyases; Carbon-oxygen lyases; Acting on phosphates
4.2.3.88	Lyases; Carbon-oxygen lyases; Acting on phosphates
4.3.1.19	Lyases; Carbon-nitrogen lyases; Ammonia-lyases
4.4.1.14	Lyases; Carbon-sulfur lyases; Carbon-sulfur lyases (only sub-subclass identified to date)
4.4.1.5	Lyases; Carbon-sulfur lyases; Carbon-sulfur lyases (only sub-subclass identified to date)
4.4.1.9	Lyases; Carbon-sulfur lyases; Carbon-sulfur lyases (only sub-subclass identified to date)
5.1.3.1	Isomerases; Racemases and epimerases; Acting on carbohydrates and derivatives
5.2.1.13	Isomerases; cis-trans-Isomerases; cis-trans Isomerases (only sub-subclass identified to date)
5.3.1.1	Isomerases; Intramolecular oxidoreductases; Interconverting aldoses and ketoses, and related compounds
5.3.1.9	Isomerases; Intramolecular oxidoreductases; Interconverting aldoses and ketoses, and related compounds
5.3.99.9	Isomerases; Intramolecular oxidoreductases; Other intramolecular oxidoreductases
5.4.2.12	Isomerases; Intramolecular transferases; Phosphotransferases (phosphomutases)
5.4.2.2	Isomerases; Intramolecular transferases; Phosphotransferases (phosphomutases)
5.4.3.8	Isomerases; Intramolecular transferases; Transferring amino groups
5.4.99.39	Isomerases; Intramolecular transferases; Transferring other groups
5.4.99.40	Isomerases; Intramolecular transferases; Transferring other groups
5.4.99.55	Isomerases; Intramolecular transferases; Transferring other groups
5.5.1.18	Isomerases; Intramolecular lyases; Intramolecular lyases (only sub-subclass identified to date)
5.5.1.19	Isomerases; Intramolecular lyases; Intramolecular lyases (only sub-subclass identified to date)
5.99.1.3	Isomerases; Other isomerases; Sole sub-subclass for isomerases that do not belong in the other subclasses
6.1.1.17	Ligases; Forming carbon-oxygen bonds; Ligases forming aminoacyl-tRNA and related compounds
6.1.1.6	Ligases; Forming carbon-oxygen bonds; Ligases forming aminoacyl-tRNA and related compounds
6.2.1.12	Ligases; Forming carbon-sulfur bonds; Acid-thiol ligases
6.2.1.5	Ligases; Forming carbon-sulfur bonds; Acid-thiol ligases
6.3.2.2	Ligases; Forming carbon-nitrogen bonds; Acid-D-amino-acid ligases (peptide synthases)
6.3.2.3	Ligases; Forming carbon-nitrogen bonds; Acid-D-amino-acid ligases (peptide synthases)
6.3.4.4	Ligases; Forming carbon-nitrogen bonds; Other carbon-nitrogen ligases
6.6.1.1	Ligases; Forming nitrogen-D-metal bonds; Forming coordination complexes

Table 17. Summary of enzymatic classes exclusively detected in Tomato, in a Tomato-Grapevine comparison.

Annex B

EC NUMBER	CLASS
1.1.1.366	Oxidoreductases; Acting on the CH-OH group of donors; With NAD+ or NADP+ as acceptor
1.13.11.54	Oxidoreductases; Acting on single donors with incorporation of molecular oxygen
1.2.4.4	Oxidoreductases; Acting on the aldehyde or oxo group of donors; With a disulfide as acceptor
2.1.1.228	Transferases; Transferring one-carbon groups; Methyltransferases
2.1.1.240	Transferases; Transferring one-carbon groups; Methyltransferases
2.1.1.267	Transferases; Transferring one-carbon groups; Methyltransferases
2.3.1.196	Transferases; Acyltransferases; Transferring groups other than aminoacyl groups
2.3.1.232	Transferases; Acyltransferases; Transferring groups other than aminoacyl groups
2.3.1.95	Transferases; Acyltransferases; Transferring groups other than aminoacyl groups
2.4.1.115	Transferases; Glycosyltransferases; Hexosyltransferases
2.5.1.51	Transferases; Transferring alkyl or aryl groups, other than methyl groups
2.5.1.52	Transferases; Transferring alkyl or aryl groups, other than methyl groups
2.7.1.137	Transferases; Transferring phosphorus-containing groups; Phosphotr. with an alcohol group
2.7.4.3	Transferases; Transferring phosphorus-containing groups; Phosphotr. with a phosphate group
2.7.7.13	Transferases; Transferring phosphorus-containing groups; Nucleotidyltransferases
3.1.3.77	Hydrolases; Acting on ester bonds; Phosphoric-monoester hydrolases
3.2.1.17	Hydrolases; Glycosylases; Glycosidases
3.4.19.12	Hydrolases; Acting on peptide bonds (peptidases); Omega peptidases
3.6.1.19	Hydrolases; Acting on acid anhydrides; In phosphorus-containing anhydrides
3.6.4.-	-
3.6.5.-	-
4.2.1.104	Lyases; Carbon-oxygen lyases; Hydro-lyases
4.2.1.109	Lyases; Carbon-oxygen lyases; Hydro-lyases
4.2.1.50	Lyases; Carbon-oxygen lyases; Hydro-lyases
4.2.1.93	Lyases; Carbon-oxygen lyases; Hydro-lyases
4.2.3.111	Lyases; Carbon-oxygen lyases; Acting on phosphates
4.2.3.22	Lyases; Carbon-oxygen lyases; Acting on phosphates
4.2.3.75	Lyases; Carbon-oxygen lyases; Acting on phosphates
5.5.1.6	Isomerases; Intramolecular lyases; Intramolecular lyases (only sub-subclass identified to date)
6.3.4.14	Ligases; Forming carbon-nitrogen bonds; Other carbon-nitrogen ligases
6.3.5.-	-
6.3.5.7	Ligases; Forming carbon-nitrogen bonds; Carbon-nitr. ligases with glutam. as amido-N-donor

Table 18. Summary of enzymatic classes exclusively detected in Grapevine, in a Tomato-Grapevine comparison.

Acknowledgements

I would like to thank my supervisor Maria Luisa Chiusano for her support, encouragement and advice. I feel privileged to work with her in the past and I hope I will be able to collaborate with her again in the future.

Special thanks to Hamed Bostan, Pasquale Di Salle, Mara Sangiovanni and Alessandra Vigilante, the people without whom the work of described in the second chapter would not be possible. Thanks a lot to Valentino Ruggieri and Chiara Colantuono for their constant interest in my work and their availability to answer to my question. Special thanks to Alfonso Esposito, for providing me valuable help in the InterPro analyses showed in this thesis.

Thanks a lot to Daniele Del Monaco, Carlo Impradice and Francesco Monticolo, the youngest members of our lab, always willing to help if needed.

Thanks in particular to all the people I met in the "Genopom" building during these three years. I would like to thank all of them for the nice discussions, for the friendly atmosphere and for good coffees. Special thanks to Prof. Luigi Frusciante, our mentor and trusted guide.

I would like to thank all my professors and all my colleagues met in the past years at the University of Naples "Federico II".

Finally, I want to thanks the people who supported me in my personal life. Thanks a lot to all my friends, to my parents, to my sister, to my parents in-law and to all my family.

This thesis is dedicated to my wife Stefania and to my daughter Arianna, my two girls, the driving force of my life.

List of Figures

List of Tables

References

Albert, R. and Barabási, A.-L. 2002. Statistical mechanics of complex networks. *Reviews of Modern Physics.* **74**: 47-97.

Alexeyenko, A., Lindberg, J., Pérez-Bercoff, Å., et al. 2006. Overview and comparison of ortholog databases. *Drug Discovery Today: Technologies.* **3**: 137-143.

Alm, E. and Arkin, A. P. 2003. Biological networks. *Curr Opin Struct Biol.* **13**: 193-202.

Alon, U. 2003. Biological networks: the tinkerer as an engineer. *Science.* **301**: 1866-1867.

Altenhoff, A. M. and Dessimoz, C. 2009. Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Comput Biol.* **5**: e1000262.

Altenhoff, A. M. and Dessimoz, C. 2012. Inferring orthology and paralogy. *Methods Mol Biol.* **855**: 259-279.

Altenhoff, A. M., Schneider, A., Gonnet, G. H., et al. 2011. OMA 2011: orthology inference among 1000 complete genomes. *Nucleic Acids Res.* **39**: D289-294.

Altenhoff, A. M., Studer, R. A., Robinson-Rechavi, M., et al. 2012. Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. *PLoS Comput Biol.* **8**: e1002514.

Altschul, S. F., Gish, W., Miller, W., et al. 1990. Basic local alignment search tool. *J Mol Biol.* **215**: 403-410.

Ambrosino, L., Bostan, H., di Salle, P., et al. 2016. pATsi: Paralogs and Singleton Genes from Arabidopsis thaliana. *Evol Bioinform Online.* **12**: 1-7.

Barabasi, A. L. and Oltvai, Z. N. 2004. Network biology: understanding the cell's functional organization. *Nat Rev Genet.* **5**: 101-113.

Beilstein, M. A., Al-Shehbaz, I. A., Mathews, S., et al. 2008. Brassicaceae phylogeny inferred from phytochrome A and ndhF sequence data: tribes and trichomes revisited. *American Journal of Botany.* **95**: 1307-1327.

Bevan, M. and Walsh, S. 2005. The Arabidopsis genome: a foundation for plant research. *Genome Res.* **15**: 1632-1642.

Blanc, G., Barakat, A., Guyot, R., et al. 2000. Extensive duplication and reshuffling in the Arabidopsis genome. *Plant Cell.* **12**: 1093-1101.

Blanc, G., Hokamp, K. and Wolfe, K. H. 2003. A recent polyploidy superimposed on older large-scale duplications in the Arabidopsis genome. *Genome Res.* **13**: 137-144.

Blanc, G. and Wolfe, K. H. 2004. Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. *Plant Cell.* **16**: 1679-1691.

Bonierbale, M. W., Plaisted, R. L. and Tanksley, S. D. 1988. RFLP Maps Based on a Common Set of Clones Reveal Modes of Chromosomal Evolution in Potato and Tomato. *Genetics.* **120**: 1095-1103.

Bradbury, L. M., Niehaus, T. D. and Hanson, A. D. 2013. Comparative genomics approaches to understanding and manipulating plant metabolism. *Curr Opin Biotechnol.* **24**: 278-284.

Bray, D. 2003. Molecular networks: the top-down view. *Science.* **301**: 1864-1865.

Burlingame, B., Mouillé, B. and Charrondière, R. 2009. Nutrients, bioactive non-nutrients and anti-nutrients in potatoes. *Journal of Food Composition and Analysis.* **22**: 494-502.

Camacho, C., Coulouris, G., Avagyan, V., et al. 2009. BLAST+: architecture and applications. *BMC Bioinformatics.* **10**: 421.

Chen, F., Mackey, A. J., Stoeckert, C. J., Jr., et al. 2006. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.* **34**: D363-368.

Chen, F., Mackey, A. J., Vermunt, J. K., et al. 2007. Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS One.* **2**: e383.

Chen, X. and Zhang, J. 2012. The ortholog conjecture is untestable by the current gene ontology but is supported by RNA sequencing data. *PLoS Comput Biol.* **8**: e1002784.

Conner, J. A., Conner, P., Nasrallah, M. E., et al. 1998. Comparative mapping of the Brassica S locus region and its homeolog in Arabidopsis. Implications for the evolution of mating systems in the Brassicaceae. *Plant Cell.* **10**: 801-812.

Costanzo, M. C., Hogan, J. D., Cusick, M. E., et al. 2000. The yeast proteome database (YPD) and Caenorhabditis elegans proteome database (WormPD): comprehensive resources for the organization and comparison of model organism protein information. *Nucleic Acids Res.* **28**: 73-76.

Coutinho, T. J., Franco, G. R. and Lobo, F. P. 2015. Homology-independent metrics for comparative genomics. *Comput Struct Biotechnol J.* **13**: 352-357.

CRIBI Available online: http://genomes.cribi.unipd.it.

Cui, L., Wall, P. K., Leebens-Mack, J. H., et al. 2006. Widespread genome duplications throughout the history of flowering plants. *Genome Res.* **16**: 738-749.

de Crecy-Lagard, V. and Hanson, A. D. 2007. Finding novel metabolic genes through plant-prokaryote phylogenomics. *Trends Microbiol.* **15**: 563-570.

Dessimoz, C., Cannarozzi, G., Gil, M., et al. 2005. OMA, A Comprehensive, Automated Project for the Identification of Orthologs from Complete Genome Data: Introduction and First Achievements. *Comparative Genomics.* **3678**: 61-72.

Dessimoz, C., Gabaldon, T., Roos, D. S., et al. 2012. Toward community standards in the quest for orthologs. *Bioinformatics.* **28**: 900-904.

Diestel, R. (2010). Graph Theory, Springer-Verlag Berlin Heidelberg.

Ding, G., Sun, Y., Li, H., et al. 2008. EPGD: a comprehensive web resource for integrating and displaying eukaryotic paralog/paralogon information. *Nucleic Acids Res.* **36**: D255-262.

Dolinski, K. and Botstein, D. 2007. Orthology and functional conservation in eukaryotes. *Annu Rev Genet.* **41**: 465-507.

Dorman, C. J. 2013. Genome architecture and global gene regulation in bacteria: making progress towards a unified model? *Nat Rev Microbiol.* **11**: 349-355.

Dorogovtsev, S. N. and Mendes, J. F. F. 2002. Evolution of networks. *Advances in Physics.* **51**: 1079-1187.

Duarte, J. M., Wall, P. K., Edger, P. P., et al. 2010. Identification of shared single copy nuclear genes in Arabidopsis, Populus, Vitis and Oryza and their phylogenetic utility across various taxonomic levels. *BMC Evol Biol.* **10**: 61.

Eubank, S., Guclu, H., Kumar, V. S., et al. 2004. Modelling disease outbreaks in realistic urban social networks. *Nature.* **429**: 180-184.

Fernandez-Pozo, N., Menda, N., Edwards, J. D., et al. 2015. The Sol Genomics Network (SGN)--from genotype to phenotype to breeding. *Nucleic Acids Res.* **43**: D1036-1041.

Fitch, W. M. 1970. Distinguishing homologous from analogous proteins. *Syst Zool.* **19**: 99-113.

Flagel, L. E. and Wendel, J. F. 2009. Gene duplication and evolutionary novelty in plants. *New Phytol.* **183**: 557-564.

Flicek, P., Ahmed, I., Amode, M. R., et al. 2013. Ensembl 2013. *Nucleic Acids Res.* **41**: D48-55.

Foolad, M. R. (2007). Current Status Of Breeding Tomatoes For Salt And Drought Tolerance. Advances in Molecular Breeding Toward Drought and Salt Tolerant Crops. M. A. Jenks, P. M. Hasegawa and S. M. Jain. Dordrecht, Springer Netherlands**:** 669-700.

Frishman, D. 2007. Protein annotation at genomic scale: the current status. *Chem Rev.* **107**: 3448-3466.

Fulton, D. L., Li, Y. Y., Laird, M. R., et al. 2006. Improving the specificity of high-throughput ortholog prediction. *BMC Bioinformatics.* **7**: 270.

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

Gabaldon, T. and Huynen, M. A. 2004. Prediction of protein function and pathways in the genome era. *Cell Mol Life Sci.* **61**: 930-944.

Gapper, N. E., McQuinn, R. P. and Giovannoni, J. J. 2013. Molecular and genetic regulation of fruit ripening. *Plant Mol Biol.* **82**: 575-591.

Gaut, B. S. 2001. Patterns of chromosomal duplication in maize and their implications for comparative maps of the grasses. *Genome Res.* **11**: 55-66.

Giot, L., Bader, J. S., Brouwer, C., et al. 2003. A protein interaction map of Drosophila melanogaster. *Science.* **302**: 1727-1736.

Grimplet, J., Van Hemert, J., Carbonell-Bejerano, P., et al. 2012. Comparative analysis of grapevine whole-genome gene predictions, functional annotation, categorization and integration of the predicted gene sequences. *BMC Res Notes.* **5**: 213.

Guyot, R., Lefebvre-Pautigny, F., Tranchant-Dubreuil, C., et al. 2012. Ancestral synteny shared between distantly-related plant species from the asterid (Coffea canephora and Solanum Sp.) and rosid (Vitis vinifera) clades. *BMC Genomics.* **13**: 103.

Hanson, A. D., Pribat, A., Waller, J. C., et al. 2010. 'Unknown' proteins and 'orphan' enzymes: the missing half of the engineering parts list--and how to find it. *Biochem J.* **425**: 1-11.

He, C., Munster, T. and Saedler, H. 2004. On the origin of floral morphological novelties. *FEBS Lett.* **567**: 147-151.

He, X. and Zhang, J. 2005. Gene complexity and gene duplicability. *Curr Biol.* **15**: 1016-1021.

Hirsch, C. D., Hamilton, J. P., Childs, K. L., et al. 2014. Spud DB: A Resource for Mining Sequences, Genotypes, and Phenotypes to Accelerate Potato Breeding. *The Plant Genome.* **7**.

Huerta-Cepas, J., Capella-Gutierrez, S., Pryszcz, L. P., et al. 2014. PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome. *Nucleic Acids Res.* **42**: D897-902.

Hughes, A. L. 2005. Gene duplication and the origin of novel proteins. *Proc Natl Acad Sci U S A.* **102**: 8791-8792.

Hulsen, T., Huynen, M. A., de Vlieg, J., et al. 2006. Benchmarking ortholog identification methods using functional genomics data. *Genome Biol.* **7**: R31.

Huynen, M. A. and Bork, P. 1998. Measuring genome evolution. *Proc Natl Acad Sci U S A.* **95**: 5849-5856.

Ihmels, J., Friedlander, G., Bergmann, S., et al. 2002. Revealing modular organization in the yeast transcriptional network. *Nat Genet.* **31**: 370-377.

Iovene, M., Wielgus, S. M., Simon, P. W., et al. 2008. Chromatin Structure and Physical Mapping of Chromosome 6 of Potato and Comparative Analyses With Tomato. *Genetics.* **180**: 1307-1317.

Ito, T., Chiba, T., Ozawa, R., et al. 2001. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A.* **98**: 4569-4574.

Jaillon, O., Aury, J. M., Noel, B., et al. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature.* **449**: 463-467.

Jansen, R. K., Kaittanis, C., Saski, C., et al. 2006. Phylogenetic analyses of Vitis (Vitaceae) based on complete chloroplast genome sequences: effects of taxon sampling and phylogenetic methods on resolving relationships among rosids. *BMC Evol Biol.* **6**: 32.

Jeong, H., Oltvai, Z. N. and Barabási, A. L. 2003. Prediction of Protein Essentiality Based on Genomic Data. *Complexus.* **1**: 19-28.

Jiao, Y., Leebens-Mack, J., Ayyampalayam, S., et al. 2012. A genome triplication associated with early diversification of the core eudicots. *Genome Biol.* **13**: R3.

Jiao, Y., Wickett, N. J., Ayyampalayam, S., et al. 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature.* **473**: 97-100.

Johnston, J. S., Pepper, A. E., Hall, A. E., et al. 2005. Evolution of genome size in Brassicaceae. *Ann Bot.* **95**: 229-235.

Joint_Genome_Institute 2008. JGI. Available at ftp://ftp.jgi-psf.org/pub/JGI_data/Sorghum_bicolor/v1.0/Sbi/annotation/Sbi1.4/.

Jones, P., Binns, D., Chang, H. Y., et al. 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics.* **30**: 1236-1240.

Kaessmann, H. 2010. Origins, evolution, and phenotypic impact of new genes. *Genome Res.* **20**: 1313-1326.

Kanehisa, M. and Goto, S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**: 27-30.

Karp, P. D., Ouzounis, C. A., Moore-Kochlacs, C., et al. 2005. Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res.* **33**: 6083-6089.

Kinsella, R. J., Kahari, A., Haider, S., et al. 2011. Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database (Oxford).* **2011**: bar030.

Knapp, S. 2002. Tobacco to tomatoes: a phylogenetic perspective on fruit diversity in the Solanaceae. *J Exp Bot.* **53**: 2001-2022.

Knapp, S., Bohs, L., Nee, M., et al. 2004. Solanaceae--a model for linking genomics with biodiversity. *Comp Funct Genomics.* **5**: 285-291.

Koonin, E. V. 2005. Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet.* **39**: 309-338.

Koornneef, M. and Meinke, D. 2010. The development of Arabidopsis as a model plant. *Plant J.* **61**: 909-921.

Kristensen, D. M., Wolf, Y. I., Mushegian, A. R., et al. 2011. Computational methods for Gene Orthology inference. *Brief Bioinform.* **12**: 379-391.

Ku, H. M., Vision, T., Liu, J., et al. 2000. Comparing sequenced segments of the tomato and Arabidopsis genomes: large-scale duplication followed by selective gene loss creates a network of synteny. *Proc Natl Acad Sci U S A.* **97**: 9121-9126.

Kuzniar, A., van Ham, R. C., Pongor, S., et al. 2008. The quest for orthologs: finding the corresponding gene across genomes. *Trends Genet.* **24**: 539-551.

Lamesch, P., Berardini, T. Z., Li, D., et al. 2012. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.* **40**: D1202-1210.

Lewis, W. H. 1979. Polyploidy in angiosperms: dicotyledons. *Basic Life Sci.* **13**: 241-268.

Li, S., Armstrong, C. M., Bertin, N., et al. 2004. A map of the interactome network of the metazoan C. elegans. *Science.* **303**: 540-543.

Lodhi, M. A. and Reisch, B. I. 1995. Nuclear DNA content of Vitis species, cultivars, and other genera of the Vitaceae. *Theor Appl Genet.* **90**: 11-16.

Long, M., Betran, E., Thornton, K., et al. 2003. The origin of new genes: glimpses from the young and old. *Nat Rev Genet.* **4**: 865-875.

Lopes, C. T., Franz, M., Kazi, F., et al. 2010. Cytoscape Web: an interactive web-based network browser. *Bioinformatics.* **26**: 2347-2348.

Luz, H. and Vingron, M. 2006. Family specific rates of protein evolution. *Bioinformatics.* **22**: 1166-1171.

Lynch, M. and Conery, J. S. 2000. The evolutionary fate and consequences of duplicate genes. *Science.* **290**: 1151-1155.

Lysak, M. A., Koch, M. A., Pecinka, A., et al. 2005. Chromosome triplication found across the tribe Brassiceae. *Genome Res.* **15**: 516-525.

Maere, S., De Bodt, S., Raes, J., et al. 2005. Modeling gene and genome duplications in eukaryotes. *Proc Natl Acad Sci U S A.* **102**: 5454-5459.

Magadum, S., Banerjee, U., Murugan, P., et al. 2013. Gene duplication as a major force in evolution. *J Genet.* **92**: 155-161.

Mason, O. and Verwoerd, M. 2007. Graph theory and networks in Biology. *IET Syst Biol.* **1**: 89-119.

Meinke, D. W., Cherry, J. M., Dean, C., et al. 1998. Arabidopsis thaliana: a model plant for genome analysis. *Science.* **282**: 662, 679-682.

Meletis, G. 2016. Carbapenem resistance: overview of the problem and future perspectives. *Ther Adv Infect Dis.* **3**: 15-21.

Mitchell, A., Chang, H. Y., Daugherty, L., et al. 2015. The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res.* **43**: D213-221.

Moniz de Sa, M. and Drouin, G. 1996. Phylogeny and substitution rates of angiosperm actin genes. *Mol Biol Evol.* **13**: 1198-1212.

Moreno-Hagelsieb, G. and Latimer, K. 2008. Choosing BLAST options for better detection of orthologs as reciprocal best hits. *Bioinformatics.* **24**: 319-324.

Moriya, Y., Itoh, M., Okuda, S., et al. 2007. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.* **35**: W182-185.

NCBI_Resource_Coordinators 2013. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*

Newman, M. E. J. 2003. The Structure and Function of Complex Networks. *SIAM Review.* **45**: 167-256.

Nguyen, H. C., Hoefgen, R. and Hesse, H. 2012. Improving the nutritive value of rice seeds: elevation of cysteine and methionine contents in rice plants by ectopic expression of a bacterial serine acetyltransferase. *J Exp Bot.* **63**: 5991-6001.

O'Brien, K. P., Remm, M. and Sonnhammer, E. L. 2005. Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res.* **33**: D476-480.

Oswald, M., Fischer, M., Dirninger, N., et al. 2007. Monoterpenoid biosynthesis in Saccharomyces cerevisiae. *FEMS Yeast Res.* **7**: 413-421.

Overbeek, R., Fonstein, M., D'Souza, M., et al. 1999. The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A.* **96**: 2896-2901.

Paterson, A. H., Bowers, J. E., Bruggmann, R., et al. 2009. The Sorghum bicolor genome and the diversification of grasses. *Nature.* **457**: 551-556.

Paterson, A. H., Bowers, J. E., Burow, M. D., et al. 2000. Comparative genomics of plant chromosomes. *Plant Cell.* **12**: 1523-1540.

Pavlopoulos, G. A., Secrier, M., Moschopoulos, C. N., et al. 2011. Using graph theory to analyze biological networks. *BioData Min.* **4**: 10.

Pereira, C., Denise, A. and Lespinet, O. 2014. A meta-approach for improving the prediction and the functional annotation of ortholog groups. *BMC Genomics.* **15 Suppl 6**: S16.

Peters, S. A., Bargsten, J. W., Szinay, D., et al. 2012. Structural homology in the Solanaceae: analysis of genomic regions in support of synteny studies in tomato, potato and pepper. *The Plant Journal.* **71**: 602-614.

Powell, S., Forslund, K., Szklarczyk, D., et al. 2014. eggNOG v4.0: nested orthology inference across 3686 organisms. *Nucleic Acids Res.* **42**: D231-239.

Proost, S., Van Bel, M., Sterck, L., et al. 2009. PLAZA: a comparative genomics resource to study gene and genome evolution in plants. *Plant Cell.* **21**: 3718-3731.

Rain, J. C., Selig, L., De Reuse, H., et al. 2001. The protein-protein interaction map of Helicobacter pylori. *Nature.* **409**: 211-215.

Ravasz, E., Somera, A. L., Mongru, D. A., et al. 2002. Hierarchical organization of modularity in metabolic networks. *Science.* **297**: 1551-1555.

Remm, M., Storm, C. E. and Sonnhammer, E. L. 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol.* **314**: 1041-1052.

Rensing, S. A. 2014. Gene duplication as a driver of plant morphogenetic evolution. *Curr Opin Plant Biol.* **17c**: 43-48.

Rogozin, I. B., Managadze, D., Shabalina, S. A., et al. 2014. Gene family level comparative analysis of gene expression in mammals validates the ortholog conjecture. *Genome Biol Evol.* **6**: 754-762.

Rosenfeld, J. A. and DeSalle, R. 2012. E value cutoff and eukaryotic genome content phylogenetics. *Mol Phylogenet Evol.* **63**: 342-350.

Rost, B. 1999. Twilight zone of protein sequence alignments. *Protein Eng.* **12**: 85-94.

Rouard, M., Guignon, V., Aluome, C., et al. 2011. GreenPhylDB v2.0: comparative and functional genomics in plants. *Nucleic Acids Res.* **39**: D1095-1102.

Rubin, G. M., Yandell, M. D., Wortman, J. R., et al. 2000. Comparative genomics of the eukaryotes. *Science*. **287**: 2204-2215.

Salgado, H., Santos-Zavaleta, A., Gama-Castro, S., et al. 2006. The comprehensive updated regulatory network of Escherichia coli K-12. *BMC Bioinformatics.* **7**: 1-5.

Samanta, M. P. and Liang, S. 2003. Predicting protein functions from redundancies in large-scale protein interaction networks. *Proceedings of the National Academy of Sciences.* **100**: 12579-12583.

Sangiovanni, M., Vigilante, A. and Chiusano, M. L. 2013. Exploiting a Reference Genome in Terms of Duplications: The Network of Paralogs and Single Copy Genes in Arabidopsis thaliana. *Biology.* **2**: 1465-1487.

Schreiber, F., Patricio, M., Muffato, M., et al. 2014. TreeFam v9: a new website, more species and orthology-on-the-fly. *Nucleic Acids Res.* **42**: D922-925.

Seoighe, C. and Gehring, C. 2004. Genome duplication led to highly selective expansion of the Arabidopsis thaliana proteome. *Trends Genet.* **20**: 461-464.

Sesso, H. D., Liu, S., Gaziano, J. M., et al. 2003. Dietary lycopene, tomato-based food products and cardiovascular disease in women. *J Nutr.* **133**: 2336-2341.

Shannon, P., Markiel, A., Ozier, O., et al. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**: 2498-2504.

Sharma, A., Li, X. and Lim, Y. P. 2014. Comparative genomics of Brassicaceae crops. *Breed Sci.* **64**: 3-13.

Shen-Orr, S. S., Milo, R., Mangan, S., et al. 2002. Network motifs in the transcriptional regulation network of Escherichia coli. *Nat Genet.* **31**: 64-68.

Simillion, C., Vandepoele, K., Van Montagu, M. C., et al. 2002. The hidden duplication past of Arabidopsis thaliana. *Proc Natl Acad Sci U S A.* **99**: 13627-13632.

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

Snel, B., Bork, P. and Huynen, M. 2000. Genome evolution. Gene fusion versus gene fission. *Trends Genet.* **16**: 9-11.

Snel, B., Bork, P. and Huynen, M. A. 2002. Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Res.* **12**: 17-25.

Somerville, C. and Dangl 2000. Genomics. Plant biology in 2010. *Science.* **290**: 2077-2078.

Somerville, C. and Koornneef, M. 2002. A fortunate choice: the history of Arabidopsis as a model plant. *Nat Rev Genet.* **3**: 883-889.

Sonnhammer, E. L. and Koonin, E. V. 2002. Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet.* **18**: 619-620.

Tabe, L., Wirtz, M., Molvig, L., et al. 2010. Overexpression of serine acetlytransferase produced large increases in O-acetylserine and free cysteine in developing seeds of a grain legume. *J Exp Bot.* **61**: 721-733.

Tanksley, S. D., Ganal, M. W., Prince, J. P., et al. 1992. High density molecular linkage maps of the tomato and potato genomes. *Genetics.* **132**: 1141-1160.

Tatusov, R. L., Koonin, E. V. and Lipman, D. J. 1997. A genomic perspective on protein families. *Science.* **278**: 631-637.

Tavares, S., Wirtz, M., Beier, M. P., et al. 2015. Characterization of the serine acetyltransferase gene family of Vitis vinifera uncovers differences in regulation of OAS synthesis in woody plants. *Front Plant Sci.* **6**: 74.

Teichmann, S. A. and Babu, M. M. 2004. Gene regulatory network growth by duplication. *Nat Genet.* **36**: 492-496.

The_Arabidopsis_Genome_Initiative 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature.* **408**: 796-815.

The_Arabidopsis_Information_Resource 2011. TAIR. Available at http://www.arabidopsis.org/.

The_Tomato_Genome_Consortium 2012. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature.* **485**: 635-641.

Trachana, K., Forslund, K., Larsson, T., et al. 2014. A phylogeny-based benchmarking test for orthology inference reveals the limitations of function-based validation. *PLoS One.* **9**: e111122.

Trachana, K., Larsson, T. A., Powell, S., et al. 2011. Orthology prediction methods: a quality assessment using curated protein families. *Bioessays.* **33**: 769-780.

Uetz, P., Giot, L., Cagney, G., et al. 2000. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature.* **403**: 623-627.

Uniprot_consortium 2015. UniProt: a hub for protein information. *Nucleic Acids Res.* **43**: D204-212.

Van Bel, M., Proost, S., Wischnitzki, E., et al. 2012. Dissecting plant genomes with the PLAZA comparative genomics platform. *Plant Physiol.* **158**: 590-600.

Van de Peer, Y. 2011. A mystery unveiled. *Genome Biol.* **12**: 113.

Vision, T. J., Brown, D. G. and Tanksley, S. D. 2000. The origins of genomic duplications in Arabidopsis. *Science.* **290**: 2114-2117.

Vogelstein, B., Lane, D. and Levine, A. J. 2000. Surfing the p53 network. *Nature.* **408**: 307-310.

Wang, L., Li, j., Zhao, J., et al. 2015. Evolutionary developmental genetics of fruit morphological variation within the Solanaceae. *Frontiers in Plant Science.* **6**.

Watanabe, K. 2015. Potato genetics, genomics, and applications. *Breed Sci.* **65**: 53-68.

Waterhouse, R. M., Tegenfeldt, F., Li, J., et al. 2013. OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. *Nucleic Acids Res.* **41**: D358-365.

Wikstrom, N., Savolainen, V. and Chase, M. W. 2001. Evolution of the angiosperms: calibrating the family tree. *Proc Biol Sci.* **268**: 2211-2220.

Wolfe, K. H. 2001. Yesterday's polyploids and the mystery of diploidization. *Nat Rev Genet.* **2**: 333-341.

Xia, X. (2013). Comparative Genomics, Springer.

Xu, X., Pan, S., Cheng, S., et al. 2011. Genome sequence and analysis of the tuber crop potato. *Nature.* **475**: 189-195.

Yi, H., Dey, S., Kumaran, S., et al. 2013. Structure of soybean serine acetyltransferase and formation of the cysteine regulatory complex as a molecular chaperone. *J Biol Chem.* **288**: 36463-36472.

Yilmaz, A., Mejia-Guerra, M. K., Kurz, K., et al. 2011. AGRIS: the Arabidopsis Gene Regulatory Information Server, an update. *Nucleic Acids Res.* **39**: D1118-1122.

Zhang, H., Jin, J., Tang, L., et al. 2011. PlantTFDB 2.0: update and improvement of the comprehensive plant transcription factor database. *Nucleic Acids Res.* **39**: D1114-1117.