UNIVERSITA' DEGLI STUDI DI NAPOLI "FEDERICO II"

DIPARTIMENTO DI AGRARIA

Sezione di Genetica e Biotecnologie Vegetali

Genomic approaches to trace the diversification history of important agronomic traits in plant

Candidato: Dott.re Antimo Di Donato

Supervisor: Prof.re Luigi Frusciante

Co-supervisor: Prof.ssa Maria Raffaella Ercolano

Genomic approaches to trace the diversification history of important agronomic traits in plant

Antimo Di Donato

University of Naples "Federico II", Department of Agricultural Sciences - Division of Plant genetics and biotechnology, Italy.

ABSTRACT

In order to investigate the diversification of important agronomic traits in plants, a conservation and evolution study of nucleotide binding genes from bacteria to plant kingdom was performed. The pathogen recognition genes were detected and classified in 102 organisms. In particular, the expansion and/or conservation of R-gene subgroups among organisms was investigated. Several large of NLR groups were found involved in important clustering events. A focus on orthologous pathogen recognition gene-rich regions in solanaceous species regions was also provided. A complete catalogue of eggplant (*Solanum melongena*) and pepper (*Capsicum annuum*) nucleotide-binding site (NBS), receptor-like protein (RLP) and receptor-like kinase (RLK) genes was generated and compared with tomato (*Solanum lycopersicum*) genomic repertoire. Orthologous relationships among clustering loci were found, and interesting reshuffling within given loci was observed for each analyzed species. The information obtained were integrated in a comparative map to highlight the evolutionary dynamics in which the PRG loci were involved. Diversification of 14 selected PRG-rich regions was also explored using a DNA target-enrichment approach. A large number of gene variants was found as well as rearrangements of single protein domain encoding sequences and changes in chromosome gene order among species. Lastly, whole-genome sequences of herbarium samples were compared to the genomes of modern tomato accessions to investigate the improvement history of the tomato crop in Italy and in Campania region in the last centuries. An aDNA extraction from herbarium tomato leaves was set up and successively used to perform aDNA sequencing sequenced. Several structural variants were detected in important genes of the ancient genomes. A comparison with a panel of wild and cultivated tomato was performed to shed light on genome pedigree history of European tomato. The findings of this thesis contribute to addressing several biological questions concerning the history of plant genome evolution and diversification.

TABLE OF CONTENTS

1. INTRODUCTION

1.1 New challenges for crop breeding

Plant breeding efforts, from the domestication of wild plant species to the present, have played a significant role in providing the food, feed, fuel, and fiber for the development of human society that currently sustains more than 6 billion individuals living in the world (Hallauer 2011).

In last 50 years the traditional crop improvement allowed to increase yield and quality traits using massive agrochemical inputs in many species (Prohens 2011). Today, the changing climate and the growing global of population requires new solutions in development of supply and agricultural production. The food demand is estimated to increase at a rate of 100–110% between 2005 and 2050 and the agricultural production cannot be implemented by increasing the cultivated area, since it would have a strong environmental impact (Tilman et al. 2011). New varieties able to efficiently use resources in changing climate should be developed.

Recent advances in genomics field made available to the scientist and breeder several tools to study the genome and its relations with phenotype, giving the opportunity to repeat the revolution triggered by plant breeding in the 20th century. Standard genetic and breeding approach permits to study only few genes, mutations or agronomic traits at one time. The availability of huge omics data source and recent sequencing technologies may improve the discovery of genetic mutations in plant disease resistance genes and other important agronomical traits. Genomic approaches can elucidate the influence of genes or genomic regions on phenotype variations and evolution, giving the access to essential information for genetic improvement. In addition, omics data sources and NGS (Next Generation Sequencing) technologies could also accelerate the cloning and the editing of genes (Kim et al. 2014a; Steuernagel et al. 2016).

1.2 Sequencing technologies

The first sequencing methods, developed and spread in the seventies, were the Maxam and Gilbert method (Maxam & Gilbert 1977) and the Sanger method. The Sanger sequencing, based on chain-terminating dideoxynucleoside analogs that caused base-specific termination of primed DNA synthesis, had been the most widely used sequencing

method approach for at least 30 years and it remains in wide use for validation of newest techniques. The first genome sequence obtained from a eukaryotic organism was the mitochondrial human genome, published in 1981 using Sanger method (Anderson et al. 1981). The great advances in automation of DNA sequencing and the development of computer programs for the analysis of sequence data made possible the sequencing of eukaryotic genomes in the mid-80s. Chain termination sequencing of bacterial artificial chromosome (BAC)-based physical maps was the main used to perform genome sequences until first decade of this century (Bevan & Uauy 2013). In the last 10 years Next Generation Sequencing platforms, in particular the 454 (http://www.454.com) and Illumina (http://www.illumina.com), had a substantial reduction in cost per base pair and times.

NGS technologies allowed to complete several important sequencing projects of crops which were begin using old sequencing technology many years before (Garcia-Mas et al. 2012; Tomato & Consortium 2012). Therefore, numerous crop sequencing projects, which integrated different NGS technologies to exploit the advantages of each method, were launched (Xu et al. 2011; Tomato & Consortium 2012; Moghe et al. 2014). In recent years, due to the higher availability of genomic data from most important crops, it was also increased the sequencing and the re-sequencing of wild and cultivated plant genomes to improve the knowledge on crop traits. Data from plant genome sequences that can be used to develop markers, to improve the genetic mapping of agronomic traits, to detect of the genetic basis of interesting phenotypes, to reconstruction evolution or domestication of plants.

Targeted sequencing

The high automation of sequencing techniques has decreased the research costs, however, analyzing an entire genome is still challenging for little research projects (Clark et al. 2011). Genomic studies often require the analysis of dozens or hundreds samples, increasing costs further. For this reason, an alternative NGS approaches called target sequencing is quickly spreading. The term Targeted Sequencing refers to a set of techniques designed to isolate and to sequence a specific fraction of a genome. These techniques are well suited to the study of plant genomes for several reasons, primarily

fewer bases to be sequenced for a sample which means lower costs. Furthermore, the plant genomes due to high repetitive sequences tend to be very large, and often few genomic regions are associated with biological functions or agronomical traits (Kiialainen et al. 2011). There are different target sequencing techniques commercially available, among these the most popular are the hybridization-based sequence capture and the PCR amplification-based methods. In the first technologies, synthetic oligonucleotides are hybridized to regions of interest; in the second method, the region of interest are amplificated using PCR. The amplification in PCR-based method is very difficult for large genomic regions because the multiple primer pairs or probes required to cover several megabases of nucleotides. An additional problem is the allele drop-out, which occurs when a variant is located in a primer binding site hindering hybridization and stopping the amplification (Neves et al. 2013). Instead, hybridization-based method has no problems with long sequences. The hybridization-based approaches has been successfully applied to identification of mutations involved in human diseases, also it has been useful to link genetic variants to agricultural phenotypic traits of interest (Gasc et al. 2016). Other potential applications of this technique include population genomics, ancient genomics, non-model organism (Gasc et al. 2016) and isolation of new genes (Witek et al. 2016).

Ancient DNA sequencing

The remarkable progress in genetics and genomics lead to the creation new and fascinating fields of study, such as the analysis of ancient DNA. Ancient DNA (aDNA) can be extracted from biological archaeological and historical material, archival collections of herbarium or medical specimens, older than 75 years (Graham 2007). The field of ancient DNA studies was probably born in 1985 with the study of DNA material from the quagga, an extinct subspecies of plain zebra that lived in South Africa until the 19th century (Higuchi et al. 1984). This work had stimulated the study of DNA of all the oldest and best-preserved samples extracted from amber or sediments.

The nucleic acids extracted from ancient samples, unlike DNA of modern samples, had a low quality, which limits the achievable information. A number of factors promote the degradation of such genetic material, such as temperature, presence of water or air, high

pressure, exposure to light, biotic and abiotic contamination. In addition, old nucleic acids may contain a large number of post-mortem mutations as the deamination of cytosine, which increase with time and of genomic structure more susceptible to miscoding lesions, potentially leading to sequence errors, or physical destruction of the DNA molecule, thus increasing the risk for preferential amplification of exogenous contaminant sequences. Furthermore, the cytoplasmic DNA concentration is usually a thousand times higher than that of nuclear in an ancient sample (Rizzi et al. 2012). Lastly, modern human DNA and microbial DNA (ancient or modern) can contaminate aDNA samples. The described issues can influence the quality and quantity of ancient sample, DNA extraction, amplification and sequencing of aDNA. The problems that plague this field of investigation require, therefore, specific technical solutions.

Many aDNA studies on different organisms have elucidated important archaeological and evolutionary questions, showing patterns of crop domestication and migration (Der Sarkissian et al. 2015). In the last few years, the advent of new sequencing technologies have considerably increased the availability of aDNA data, thus could greatly improving our knowledge on crop evolution, adaptation and domestication. An additional fascinating aspect of aDNA investigation is the discovery of lost useful mutations that could be reintroduced in modern crops. There are different sources from which obtain plant aDNA, among these herbarium collections can be an excellent font of information. The ancient collections, preserving the ancient structure of the plant, can be used to correlate genomic data with observed phenotype. Several ancient plant genomes studies could be performed in the next future in order to elucidate the patterns of plant diversification and divergence. Last year, two studies on ancient barley (Mascher et al. 2016) and maize (Ramos-Madrigal et al. 2016) genomes provided significant insights related to domestication and origin of these modern crops.

1.3 Web platforms and bioinformatic tools for crop improvement

Basic informatic systems can provide information for facilitating many aspects of crop improvement. Several organizations share with scientists and breeders information regarding crops and their relative genomes on websites.

Nowadays, data from many plant sequencing project are available completely free on different web portals. NCBI database (http://www.ncbi.nlm.nih.gov/) is the most important for the content of omics data volumes. Other databases including plant genome sequences are Plant GDB (http://www.plantgdb.org) and Phytozome (http://www.phytozome.net). Databases are often created by the same organizations that guide the sequencing projects of a certain species or botanical family. The Arabidopsis Information Resource (TAIR) maintains a database that includes the complete genome sequence along with gene structure, gene product information, gene expression, DNA and information about the Arabidopsis research community. The Sol Genomics Network (SGN) is a family-oriented database dedicated to the Solanaceae family, the portal includes genetics and omics information about important crops such as tomato, potato, pepper and tobacco (Fernandez-Pozo et al. 2015). Some databases contain information about gene family correlated with specific agronomical traits such as PRGdb (Plant Resistance Genes database), which includes data about plant resistance genes, related pathogens and diseases (Sanseverino et al. 2009). The huge amount of data produced by omic and genetic studies, requires the development informatics tools (algorithms and software), capable of analyzing large volumes of data and simplify the study of complex biological traits.

In genetics and genomics, many bioinformatics tools were develop to browse genome sequences, analyze proteins or nucleotides, assembly or mapping reads, predict and annotate genes, perform comparative and evolutionary studies. Standard NGS technology produces short sequences typically called reads. They can be assembled using two approaches: de novo or mapping. The de novo method consist in assembling overlapped reads to create longer sequences (contigs, scaffolds or pseudomolecules). De-novo assemblies are slower and more memory demanding than mapping assemblies, but they are more much precise and exhaustive. Reads mapping allows to align sequences against an existing reference genome, building a sequence that is similar but not identical to the reference. Mapping approaches are faster than de novo assemblies, it allows to detect easily new structural variation, such as deletions, insertions and rearrangements (Li & Durbin 2009). After the mapping is possible to identify single nucleotide polymorphism (SNPs) and small InDel (insertion or the deletion of bases)Classification of proteins and extraction of motifs can be performed through a variety of tools such as Pfam (Bateman

et al. 2002), InterProScan (Jones et al. 2014) or SMART (Schultz et al. 1998). Alignment of proteins and genes is important to show similarities and differences in homolog sequences. The evolutionary history of individual gene families or plant species can be followed performing a comparative analysis.

1.4 Comparative and evolution analysis

Comparative analysis uses natural variations to understand the patterns of life at all levels - from genes to communities - and the historical relationships of individuals or higher taxa and the mechanisms and patterns that drives it (Hardison 2003). Natural variants in crop plants resulted mainly from spontaneous mutations in their wild progenitors. Crop domestication and breeding have a profound influence on the genetic diversity present in modern crops. Understanding the genetic basis of phenotypic variation and the domestication processes in crops can help us efficiently utilize these diverse genetic resources for crop improvement. The use of naturally occurring alleles has greatly increased agricultural production. Through the use of germplasm resources and genetic tools such as genome sequences, genetic populations and genome-wide association studies, crop researchers are now able to extensively and rapidly mine natural variation and associate phenotypic variation with the underlying sequence variants (Bevan & Uauy 2013). Recently, the advent of second-generation sequencing has facilitated the discovery and use of natural variation in crop design and genome-wide selection. The nearly completed sequences of plant species shed light on the history of genome evolution, and provide a foundation for advancing knowledge in many agronomically important plant species.

1.5 Genomic analysis of target traits

Crop breeders explore and use the variability of the germplasm collections to improve plant characteristics. Whether these traits are associated with yield, disease and insect resistance or quality traits they are all subjected to selection pressure. Like evolution this selection process is very slow for some traits or dramatically quick for other. Many favourable traits have been introgressed in the last years using empirical methods. The next step in genetic research would be the development of a theoretical framework that

allows reliable predictions of the phenotypic consequences when making alterations to the genome make-up of a plant (Hammer et al. 2006). Genomic information has increased exponentially during the past two decades and will enhance selection process. Optimistically, it seems further genetic progress can be sustained because as greater genetic information at the molecular level is understood and integrated with phenotypic selection (Hallauer 2011). Genomic methodologies showed to be useful to elucidate the basis of genetic traits/characteristics, to understand the phenotypic of important loci throughout the in crops belonging to Poaceae and Solanaceae species (Takeda & Matsuoka 2008). In terms of developmental aspects, terminal-branching pattern and fruit-size control seem to be the predominant determinants for the yield improvement of fruits and grains (Peng et al. 1999). They display a decrease in nucleotide diversity and increased LD after strong selection, such as during domestication and subsequent crop improvement. Recent screening showed that loci that loci controlling fruit size in tomato have been important selection targets (Chakrabarti et al. 2013). Domestication genes can identified by comparing nucleotide sequence diversity between a crop species and extant populations of wild relatives as a proxy of the ancestor species.

Plant disease resistance traits

Probably the most desired crop trait is the resistance to plant pathogens. Plant disease resistance is fundamental to obtain reliable production of food, and it provides significant reductions in agricultural use of land, water, fuel and other inputs. Plants defend them self from pathogens thorough a sophisticated defense system based on the ability of plants to distinguish the phytopathogen life-styles. The circular model describes the plant–pathogen interaction in three distinct phases: (1) interaction, (2) activation, and modulation (3) effective resistance. This model schematically showed the crucial points of two components (activation and modulation) of innate plant immunity and the resultant of their combination (Andolfo & Ercolano 2015). The activation component is essentially based on the presence at the cellular levels of specific pathogen receptors called R proteins. These proteins encoded by the pathogen recognition genes (PRGs), are characterized by some common domains such as CC (coiled-coil), NB (Nucleotide binding region), TIR (Toll-interleukin region), LRR (Leucine rich region) and K (Kinase

domain). The structures that have NB-LRR domains are divided into two classes: TNL (TIR-NB-LRR) and CNL (CC-NB-LRR) which possess, respectively, either the TIR or CC domains. TNL and CNL are usually present in the cytoplasm. The CNL group includes very important genes involved in crop disease resistance. Natural and cultivated plant populations carry inherent disease resistance. Monogenic or major gene (R gene) resistance, has been widely studied at genomic level (Sekhwal et al. 2015) and employed by breeders. New approaches for exploring resistance genes dataset could be useful for shed light in molecular and evolutionary mechanisms of this gene family and for facilitating the design of diagnostic tests, comparative analysis and new breeding program.

1.6 Scientific aims

Main goal of this thesis was to study the diversification of crop agronomical traits using genomic approaches. The first section is dedicated at the study of conservation and evolution of nucleotide binding genes from bacteria to plant kingdom. The second part reports a pilot comparison of orthologous pathogen recognition gene-rich regions in solanaceous species. In the third part ancient DNA extracted from two tomato herbarium samples was sequenced and analyzed to understand the selection routes followed by tomato growers in Campania region, with a focus on variation of candidate genes involved in determination of fruit quality traits.

2. RECONSTRUCTION OF EVOLUTIONARY HISTORY OF NLR-LIKE GENE FAMILY IN METAPHYTA KINGDOM

2.1 Introduction

Most intracellular immune receptors in plants are characterized by the presence of a nucleotide-binding site and leucine-rich repeats (NLRs, also known as, NB-LRRs or NBS-LRRs), these domains are present in the majority of cloned resistance genes (R-genes) (McHale et al. 2006). NLR protein families are divided into two classes based on the presence or absence of a toll-interleukin-1 receptor (TIR) domain in TIR-NLR (TNL) or non-TIR-NLR (n-TNL). Plant NB-LRR proteins detects the presence of fungal, nematode, bacterial, or viral pathogens elicitors and trigger the immune response. Both the NB and TIR domains have a prokaryotic origin but their fusion was observed only in plant lineage. Recent studies suggest the eukaryote innate immunity originated from their endosymbionts (Dunin-Horkawicz et al. 2014). Indeed, plant and animals system had independent origin shaped after by convergent evolution (Yue et al. 2012).

A large variation in NLR complement among and within plant species both in the sequence composition of orthologs and in the number of paralogs was observed (Y. Zhang et al. 2016). The number of NLR *genes* can vary in plant genomes from <100 to >1,000 (Yue et al. 2012; Sarris et al. 2016; Shao, Wang, et al. 2016) and some gene families are more conserved in dicots and lost or modified in monocots (Tarr & Alexander 2009; Collier et al. 2011). Although the structure and function of NLR proteins have been extensively studied, the involvement of single domain to disease resistance process in plants is still not well understood. Proteins can expand their functional repertoire in a number of ways, including residue mutations, gain and loss of domain, motif arrangement (Sarris et al. 2016). The domain arrangement is important since its modification mostly promote interactions with novel substrates or new protein partners on different pathways and processes and have specific functional and spatial relationship (Lees et al. 2016). A number of alterations that can have a considerable effect were already found (Sanseverino & Ercolano 2012).

On the following pages, it will be shown a study on genes that encode nucleotide-binding and/or leucine-rich repeat domains using data from genome sequencing of bacteria, algae and plants. A comprehensive study of genes encoding NLRs and NLR-like genes across bacterial and plant species can provide insights into the presumed history of plant NLR evolution and it can lead the discovery the means of NB protein diversification. The

pathogen recognition genes were detected and classified in 102 organisms. In particular, the expansion and/or conservation of R-gene families among organism was studied. In addition, NLR-like genes that occur in cluster were identified and characterized.

2.2 Materials and methods

Identification of NLR-like genes and their cluster

The proteomes of 102 organisms were downloaded from Phytozome (http://phytozome.jgi. doe.gov/, v11) and other plant genome websites (Table S1). The proteins of each organism were scanned with Hidden Markov Model (HMM) of the Nucleotide-Binding (Pfam PF00931) and Leucine-Rich Repeat domains (Panther PTHR11017:SF191) to identify NB-encoding and NLR-related candidates using HMMER v3.0 with default parameters (Finn et al. 2011). Furthermore, a BlastP search on proteomes was performed with an E-value cut-off of 10 using typical R genes motifs released from Jupe et al. (Jupe et al. 2012). HMMER and BLAST output was annotated using IntetProScan (Jones et al. 2014) with PFAM, PANTHER, SUPERFAMILY and CDD databases. The annotated sequence were filtered on the basis of presence of NB and LRR domains and classified on the basis of domains detected from IntetProScan.

The phylogeny of all species was constructed on the basis of APG IV (Angiosperm Phylogeny Group) (Chase et al. 2016), Angiosperm Phylogeny Website (http://www.mobot.org/MOBOT/research/APweb/welcome.html) and "The Tree of Life Web Project" (http://tolweb.org/tree/).

Characterization of orthologous groups

All candidate R genes were used for a reciprocal best-hit analysis (threshold E-value <1e-5). The orthologues groups were obtained using OrthoMCL tool (Alexeyenko et al. 2006) with default parameters. The association of reference R-genes (http://prgdb.crg.eu/) and relative orthogroup were detected using Best Hit method (BlastP, E-value < 1e-5).

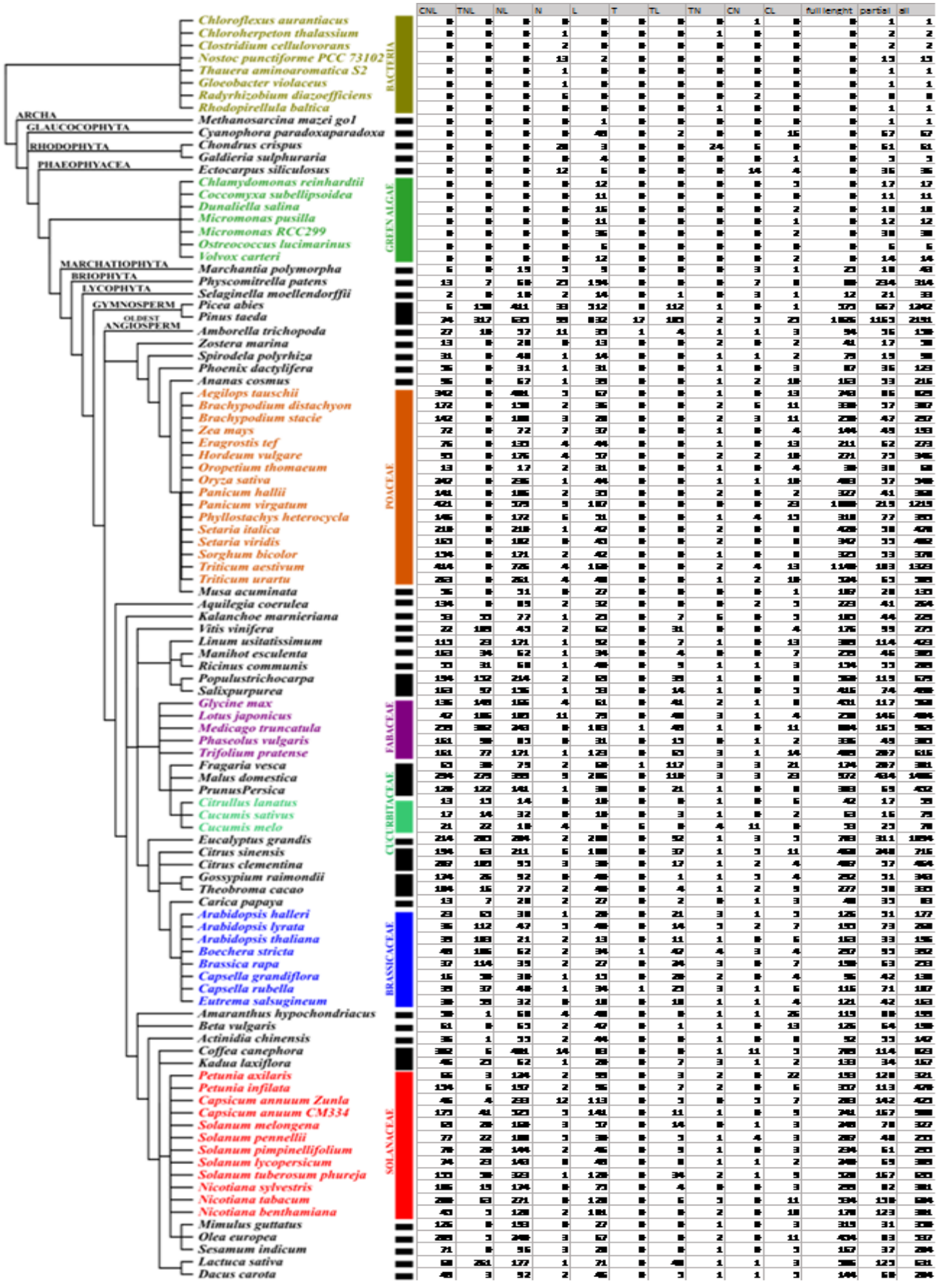
Identification of gene physical clusters

Physical clustering of candidate genes was detected using a customized scripts with gff file of corresponding genomes and Bedtools (Quinlan & Hall 2010). If two NBS genes are separated by no more than eight other genes, they are considered to be located at the same NLR-like gene cluster (Richly et al. 2002).

2.3 Results

NLR-like gene family in 102 genomes

The genes containg signatures similar to NLR domains were identified in 102 sequenced genomes representing 55 taxa, including algae, plant, bacteria and archaea divisions. About 35000 genes were identified and characterized. Peculiar NLR domains were detected in several bacteria and in one archaebacterial (Figure 1).



CNL	TNL	NL	N	L	T	TL	TN	CN	CL	fullength	partial	all
0	0	0	0	0	0	0	0	1	0	0	1	1
0	0	0	1	0	0	0	1	0	0	0	2	2
0	0	0	2	0	0	0	0	0	0	0	2	2
0	0	0	13	2	0	0	0	0	0	0	15	15
0	0	0	1	0	0	0	0	0	0	0	1	1
0	0	0	1	0	0	0	0	0	0	0	1	1
0	0	0	5	0	0	0	0	2	0	0	0	0
0	0	0	0	0	0	0	1	0	0	0	1	1
0	0	0	0	1	0	0	0	0	0	0	1	1
0	0	0	0	48	0	2	0	0	16	0	67	67
0	0	0	28	3	0	0	24	5	0	0	61	61
0	0	0	0	4	0	0	0	0	1	0	5	5
0	0	0	12	5	0	0	0	14	4	0	36	36
0	0	0	0	12	0	0	0	0	5	0	17	17
0	0	0	0	11	0	0	0	0	0	0	11	11
0	0	0	0	15	0	0	0	0	2	0	18	18
0	0	0	0	11	0	0	0	0	1	0	12	12
0	0	0	0	36	0	0	0	0	2	0	38	38
0	0	0	0	6	0	0	0	0	0	0	6	6
0	0	0	0	12	0	0	0	0	2	0	14	14
6	0	15	5	9	0	0	0	3	1	23	18	41
13	7	68	23	194	0	0	0	7	0	88	234	314
2	0	18	2	14	0	1	0	3	1	12	21	33
6	158	411	33	512	0	112	1	0	1	573	667	1240
24	317	630	99	832	17	183	2	5	23	1826	1165	2191
27	18	57	11	33	1	4	1	1	3	94	56	150
13	0	28	0	13	0	0	2	0	2	41	17	58
31	0	48	1	14	0	0	1	1	2	79	19	98
36	0	31	1	31	0	0	1	0	3	87	36	123
56	0	67	1	39	0	0	1	2	18	163	53	216
342	0	481	5	67	0	0	1	0	13	243	86	329
172	0	158	2	36	0	0	2	6	11	338	57	387
142	0	188	3	28	0	0	2	3	11	238	42	257
72	0	72	7	37	0	0	1	0	4	144	45	153
76	0	130	4	44	0	0	1	0	13	211	62	273
93	0	176	4	57	0	0	2	2	18	271	75	346
13	0	17	2	31	0	0	1	0	4	38	38	68
247	0	236	1	44	0	0	1	1	18	483	57	540
141	0	185	2	33	0	0	2	0	2	327	41	368
421	0	575	5	187	0	0	0	0	23	1088	215	1215
145	0	172	6	31	0	0	1	4	13	318	77	353
218	0	218	1	47	0	0	2	0	0	428	38	428
163	0	182	0	43	0	0	2	0	0	342	33	482
134	0	171	2	42	0	0	1	0	0	325	33	378
414	0	726	4	168	0	0	2	4	13	1148	183	1323
263	0	261	4	48	0	0	1	2	18	524	63	589
36	0	31	0	27	0	0	0	0	1	187	28	130
134	0	85	2	32	0	0	0	2	5	223	41	264
53	53	77	1	23	0	7	6	0	5	180	44	224
22	188	43	2	62	0	31	0	0	4	176	55	231
115	23	171	1	52	0	7	1	0	13	385	114	423
183	34	62	1	34	0	4	0	0	7	233	46	280
73	31	68	1	48	0	5	1	1	3	134	33	289
194	132	214	2	63	0	33	1	0	0	368	115	579
153	57	136	1	53	0	14	1	0	5	415	74	488
135	148	165	4	61	0	41	2	1	0	451	117	568
47	185	180	11	79	0	48	3	1	4	238	145	484
229	382	248	0	183	1	48	1	0	11	884	165	953
161	58	83	0	31	0	13	0	1	2	335	45	380
161	77	171	1	123	0	63	3	1	14	485	287	615
63	38	73	2	68	1	117	3	3	21	134	287	381
294	275	388	5	285	0	118	3	3	23	572	434	1486
128	122	141	1	38	0	21	1	0	0	383	65	432
13	13	14	0	18	0	0	1	0	6	42	17	59
17	14	32	0	18	0	3	1	0	2	63	16	79
21	22	18	4	0	6	0	4	11	0	53	23	78
214	283	284	2	288	0	52	1	3	5	783	311	1894
194	63	211	6	188	0	37	1	5	11	458	248	715
287	183	93	3	38	0	17	1	2	4	487	37	454
124	25	92	0	48	0	1	1	5	4	252	91	343
184	15	77	2	48	0	4	1	2	5	277	58	335
13	7	28	2	27	0	2	0	1	3	48	35	83
23	50	38	1	28	0	21	3	1	5	126	51	177
36	112	47	5	48	0	14	5	2	7	150	73	268
39	183	21	2	13	0	11	1	0	6	163	33	196
48	185	62	2	34	1	42	4	3	4	257	55	352
37	114	35	2	27	0	24	3	0	7	158	63	253
16	58	38	1	13	0	28	2	0	4	56	42	138
39	37	48	1	34	1	23	3	1	6	116	71	187
38	39	32	0	18	0	18	1	1	4	121	42	163
38	1	68	4	48	0	0	1	1	26	115	88	153
61	0	61	2	47	0	1	1	0	13	126	64	190
35	1	33	2	44	0	0	1	0	0	52	33	142
382	6	481	14	83	0	0	1	11	5	783	114	823
45	23	52	1	28	0	7	3	1	2	133	34	167
55	3	124	2	93	0	3	2	0	22	153	128	321
154	6	157	2	56	0	7	2	0	6	337	113	478
45	4	233	12	113	0	5	0	5	7	283	142	425
173	41	323	5	141	0	11	1	0	5	341	157	588
53	28	158	3	57	0	14	0	1	3	248	78	327
77	22	188	5	38	0	5	1	4	3	287	48	233
78	28	144	2	46	0	5	1	0	3	234	61	253
24	23	143	0	48	0	0	1	1	2	248	65	313
133	38	323	1	128	0	34	2	1	5	528	157	631
186	19	134	0	73	0	4	0	0	3	255	82	381
288	63	271	0	128	0	6	5	0	11	534	158	634
43	5	128	2	184	0	0	2	0	18	178	123	381
126	0	153	0	27	0	0	1	0	3	315	31	338
283	5	248	3	57	0	0	2	0	11	434	83	337
71	0	56	3	28	0	0	1	0	5	157	37	284
58	261	177	1	71	0	48	1	1	3	386	123	621
45	3	52	2	45	0	5	1	1	5	144	68	284

Figure 1. Phylogeny of the 102 organisms analysed in this study and corresponding number of detected NLR-like proteins. The phylogeny of the plants used in the analyses was constructed using different sources (see Material and Methods). NB: NB-ARC domain-containing proteins; NL: NB-ARC-LRR; TNL: TIR-NB-ARC-LRR proteins; TN: TIR-NB-ARC proteins; CC: CC-NB-ARC proteins; TL: TIR-NB-ARC-LRR proteins; Full length: CNL+TNL+NL; partial: not full length;

In the eight bacteria genomes 24 NB-encoding and 3 LRR-encoding genes were identified (Figure 1). The pluricellular red alga *Chondrus crispus* had a total of 28 NB genes and 3 LRR genes, however, in the genome of other important red alga *(Galdieria sulphuraria)*, were recorded LRR genes but not NB coding genes. Likewise, seven green algae genomes showed several LRR coding genes lacking NB domain.

NLRs were first detected in the early land plant lineage, in the liverwort (*Marchanthia polymorpha)* and moss (*Physcomitrella patens)*. Interestingly, *Marchanthia polymorpha* genome lacks TNL-like genes that in contrast were observed in *P. patens* side by side to CNL genes (13). A substantial NLR gene expansion and diversification occurred in gymnosperm group. Indeed, *P. abies* showed a number of NLR-like genes almost two times higher than *P. taeda*. By contrast a modest number of NLR genes was detected in basal flowering plant *A. trichopoda* in the seagrass *Zostera marina* and the desert plant *Spirodela polyrhiza*. The sixteen analysed Poaceae genomes lack TNL genes and showed a variable number of NLR genes (from 1323 to 68). TNLs were absent also in some eudicot genomes such as *Aquilegia coerulea*, *Beta vulgaris, Mimulus guttatus* and *Sesamum indicum*. In Eudicot, the highest number of full-length NLR genes was recorded in *Malus domestica* (972) that showed a value three time higher of other Rosaceae species. Among Fabaceae species surveyed, *Medicago truncatula* the highest number of NLR-like gene (969), more than 2-fold higher than *P. vulgaris* (385). Instead, Cucurbitaceous species showed less than 80 NLR elements. *C. papaya*, that share a common Brassicaeae ancestor (Zhang et al. 2016), possesses at least half of the NLR genes than any cruciferous genomes. In particular, we notice an increase of over 26 times of TNLs from *C. papaya (7)* to *B. stricta* (186). A large variation in the NLR-like gene number was observed among the 12 Solanaceae species analyzed with an average value 447,6 genes for species. The minimum number was observed in eggplant genome (255) and the maximum in *C. annuum* var. Zunla-1genome. *Linum usitatissimum* genome showed the lowest CNL to TNL ratio (0.20). All Brassicaceae species have a CNLs/TNLs ratio lower than one, except for *Capsella rubella*. All Solanaceae genomes are

characterized by a higher ratio of CNLs/TNLs, with the highest ratio in *P. infilata* (25.6) and the lower in *S. tuberosum* (3.11).

Orthologues group prediction and annotation using R genes

Predicted NLR-like genes clustered into 1675 orthogroups (Alexeyenko et al. 2006) that included a total of about 30000 genes. Seventy-three reference R-proteins with resistance function were assigned to orthogroups using best hit analysis. A total of 43 orthologues clusters, grouping about 14000 proteins (almost the 46,8 % of total genes in orthoMCL groups), contain R-proteins already characterized. Functional characterized R genes and relative orthologue groups were showed in Table 1.

Clonated NLR	NLR Species	Class	Best Hit ID	Best Hit Species	Orthomcl Group ID	n. of gene in group
Lr10	Triticum aestivum	CNL	TRIUR3_07291-P1	Triticum urartu	OG_1000	1259
MLA1	Hordeum vulgare	CNL	AEGTA18040	Aegilops tauschii	OG_1000	1259
MLA10	Hordeum vulgare	CNL	AEGTA18040	Aegilops tauschii	OG_1000	1259
Mla12	Hordeum vulgare subsp. Vulgare	CNL	AEGTA18040	Aegilops tauschii	OG_1000	1259
MLA13	Hordeum vulgare	CNL	AEGTA18040	Aegilops tauschii	OG_1000	1259
Mla6	Hordeum vulgare subsp. Vulgare	CNL	AEGTA18040	Aegilops tauschii	OG_1000	1259
Pi36	Oryza sativa Indica group	CNL	LOC_Os08g05440.1	Oryza sativa	OG_1000	1259
Gro1.4	Solanum tuberosum	TNL	PGSC0003DMP400030257	Solanum tuberosum phureja	OG_1001	1194
N	Nicotiana glutinosa	TNL	mRNA_87883_cds	Nicotiana tabacum	OG_1001	1194
Bs4	Solanum lycopersicum	TNL	Solyc05g007850.1.1	Solanum lycopersicum	OG_1001	1194
RY-1	Solanum tuberosum subsp. Andigena	TNL	Solyc11g011080.1.1	Solanum lycopersicum	OG_1001	1194
Pl8	Helianthus annuus	CNL	Lsa002425.1	Lactuca sativa	OG_1002	1120
Rps1-k-1	Glycine max	NL	Glyma.03G037000.1.p	Glycine max	OG_1002	1120
Rps1-k-2	Glycine max	NL	Glyma.03G034900.1.p	Glycine max	OG_1002	1120
I-2	Solanum lycopersicum	NL	Sopim11g071430.0.1	Solanum pimpinellifolium	OG_1002	1120
R3a	Solanum tuberosum	NL	PGSC0003DMP400032361	Solanum tuberosum phureja	OG_1002	1120
Rpi-blb1	Solanum bulbocastanum	CNL	PGSC0003DMP400029816	Solanum tuberosum phureja	OG_1006	686

RPM1	Arabidopsis thaliana	CNL	AT3G07040.1	Arabidopsis thaliana	OG_1007	631
FOM-2	Cucumis melo	CNL	MELO3C024725P1	Cucumis melo	OG_1008	587
RPP1	Arabidopsis thaliana	TNL	AT3G44480.1	Arabidopsis thaliana	OG_1011	528
SSI4	Arabidopsis thaliana	TNL	AT5G41750.1	Arabidopsis thaliana	OG_1011	528
RAC1	Arabidopsis thaliana	TNL	AT1G31540.2	Arabidopsis thaliana	OG_1011	528
RPP4	Arabidopsis thaliana	TNL	AT4G16860.1	Arabidopsis thaliana	OG_1011	528
RPP5	Arabidopsis thaliana	TNL	AT4G16950.1	Arabidopsis thaliana	OG_1011	528
KR1	Glycine max	TNL	Glyma.19G054900.1.p	Glycine max	OG_1012	520
RPS5	Arabidopsis thaliana	CNL	AT1G12220.1	Arabidopsis thaliana	OG_1013	520
Rps2	Arabidopsis thaliana	CNL	AT4G26090.1	Arabidopsis thaliana	OG_1013	520
Prf	Solanum pimpinellifolium	CNL	Sopim05g013280.0.1	Solanum pimpinellifolium	OG_1014	494
R1	Solanum demissum	CNL	PGSC0003DMP400044306	Solanum tuberosum phureja	OG_1014	494
Pid3	Oryza sativa Japonica group	CNL	LOC_Os06g22460.1	Oryza sativa	OG_1015	489
HRT	Arabidopsis thaliana	CNL	AT5G43470.1	Arabidopsis thaliana	OG_1016	482
RCY1	Arabidopsis thaliana	CNL	AT5G43470.1	Arabidopsis thaliana	OG_1016	482
RPP8	Arabidopsis thaliana	CNL	AT5G43470.1	Arabidopsis thaliana	OG_1016	482
PIB	Oryza sativa	CNL	Sobic.005G167400.2.p	Sorghum bicolor	OG_1017	387
Bs2	Capsicum chacoense	CNL	CA09g17480	Capsicum anuum CM334	OG_1020	334
Gpa2	Solanum tuberosum	CNL	PGSC0003DMP400013860	Solanum tuberosum phureja	OG_1020	334
Rx	Solanum tuberosum	CNL	PGSC0003DMP400013867	Solanum tuberosum phureja	OG_1020	334
Rx2	Solanum acaule	CNL	PGSC0003DMP400013860	Solanum tuberosum phureja	OG_1020	334
NRC2	Solanum lycopersicum	CNL	Solyc10g047320	Solanum lycopersicum	OG_1023	285
Hero	Solanum lycopersicum	CNL	Sopen04g003300.1	Solanum pennellii	OG_1024	279
Mi1.2	Solanum lycopersicum	CNL	Solyc06g008450.2.1	Solanum lycopersicum	OG_1024	279
Rpi-blb2	Solanum bulbocastanum	CNL	Solyc06g008790.2.1	Solanum lycopersicum	OG_1024	279
Pit	Oryza sativa	CNL	LOC_Os01g05620.1	Oryza sativa	OG_1025	266
Lr1	Triticum aestivum	CNL	AEGTA14094	Aegilops tauschii	OG_1026	236
NRG1	Nicotiana benthamiana	CNL	Niben101Scf02118g00018.1	Nicotiana benthamiana	OG_1028	226
Pm3	Triticum aestivum	CNL	AEGTA10487	Aegilops tauschii	OG_1030	219
RPP13	Arabidopsis thaliana	CNL	AT3G46530.1	Arabidopsis thaliana	OG_1031	218
ADR-1	Arabidopsis thaliana	Rpw8-NL	AT1G33560.1	Arabidopsis thaliana	OG_1033	192
Pi-ta	Oryza sativa	CNL	LOC_Os12g18360.1	Oryza sativa	OG_1035	183
L6	Linum usitatissimum	TNL	Lus10004719	Linum usitatissimum	OG_1037	156
M	Linum usitatissimum	TNL	Lus10004719	Linum usitatissimum	OG_1037	156

XA1	Oryza sativa	CNL	LOC_Os04g53120.1	Oryza sativa	OG_1038	154
Pikm2-TS	Oryza sativa Japonica group	CNL	LOC_Os11g46210.1	Oryza sativa	OG_1040	148
Rp1-D	Zea mays	CNL	GRMZM5G879178_P01	Zea mays	OG_1041	143
Rps4	Arabidopsis thaliana	TNL	AT5G45250.1	Arabidopsis thaliana	OG_1043	128
Sw-5	Solanum lycopersicum	CNL	Solyc09g098130.1.1	Solanum lycopersicum	OG_1046	117
Rdg2a	Hordeum vulgare subsp. Vulgare	CNL	TRIUR3_20950-P1	Triticum urartu	OG_1050	111
Pi2	Oryza sativa Indica group	CNL	LOC_Os06g17900.1	Oryza sativa	OG_1058	91
Pi9	Oryza sativa Indica group	CNL	LOC_Os06g17900.1	Oryza sativa	OG_1058	91
Piz-t	Oryza sativa Japonica group	CNL	LOC_Os06g17900.1	Oryza sativa	OG_1058	91
Cre1	Aegilops tauschii	CNL	Traes_2BL_E3E1888E8.1	Triticum aestivum	OG_1066	77
Lr21	Triticum aestivum	CNL	AEGTA25735	Aegilops tauschii	OG_1066	77
RRS1	Arabidopsis thaliana	TNL	AT5G45260.1	Arabidopsis thaliana	OG_1068	71
Tm-2	Solanum lycopersicum	CNL	Solyc09g018220.1.1	Solanum lycopersicum	OG_1087	51
Tm-2a	Solanum lycopersicum	CNL	Solyc09g018220.1.1	Solanum lycopersicum	OG_1087	51
Dm3	Lactuca Sativa	CNL	Lsa037896.1	Lactuca sativa	OG_1093	44
P2	Linum usitatissimum	TNL	Lus10015648	Linum usitatissimum	OG_1106	34
Pikm1-TS	Oryza sativa Japonica group	CNL	LOC_Os11g46200.1	Oryza sativa	OG_1114	30
Pikp-2	Oryza sativa Japonica group	CNL	LOC_Os11g46200.1	Oryza sativa	OG_1114	30
Pi5-2	Oryza sativa Japonica group	CNL	Pavir.9NG702800.1.p	Panicum virgatum	OG_1135	24
Pi5-1	Oryza sativa Japonica group	CNL	Seita.2G178500.1.p	Setaria italica	OG_1144	22
VAT	Cucumis melo	CNL	Cucsa.088220.1	Cucumis sativus	OG_1169	17
RLM3	Arabidopsis thaliana	TN	AT4G16990.2	Arabidopsis thaliana	OG_1824	3

Table 1. NLR orthologues groups including at least one cloned R genes identified thought a Best Hit Blast analysis.

Three R-orthogroups cluster group more than 1000 genes. Five cloned CNL genes, in grasses species, were mapped in the larger group (OG_1000) which contains 1,295 genes. The second largest orthoMCL cluster containing 1,194 TNL-like elements include four genes isolated in Solanaceae species. Most orthogroups containing cloned R-genes contain between 689 to 111 genes and only 11 orthogroups hold less than 91 genes. The smallest orthogroup (3 members) included the Arabidopsis reference genes RLM3.

References in orthogroup	Class	Poaceae	Fabaceae	Rosaceae	Cucurbitaceae	Brasicaceaee	Solanaceae
MLA1-6-10-12-13, Pi36, Lr10	CNL	1259	0	0	0	0	0
Gro1.4, N, RY-1, Bs4	TNL	0	282	139	10	3	236
Pl8, Rps1-k-1, Rps1-k-2, I-2, R3	CNL	0	344	98	0	14	211
Rpi-blb1	CNL	38	161	51	20	0	100
RPM1	CNL	21	172	94	0	5	40
FOM-2	CNL	0	0	62	27	0	81
RPP1, SSI4, RAC1,RPP4,RPS5,RPP5	TNL	0	0	0	0	485	0
KR1	TNL	0	275	65	21	0	1
RPS5, Rps2	CNL	25	9	13	0	141	0
Prf, R1	CNL	0	0	0	0	0	299
Pid3	CNL	458	0	0	0	0	0
HRT, RCY1, RPP8	CNL	0	13	53	0	138	29
PIB	CNL	387	0	0	0	0	0
Bs2, Gpa2, Rx, Rx2	CNL	0	0	0	0	0	285
NRC2	CNL	0	0	0	0	0	190
Hero, Mi1.2, Rpi-blb2	CNL	0	0	0	0	0	233
Pit	CNL	209	0	0	0	0	0
Lr1, XA1	CNL	236	0	0	0	0	0
NRG1	CNL	0	56	58	5	13	15
Pm3, Rdg2a	CNL	218	0	0	0	0	0
RPP13	CNL	1	6	2	0	12	111
ADR1	RNL	21	11	7	2	27	15
Pi-ta	CNL	183	0	0	0	0	0
L6,M	TNL	0	0	0	0	0	0
Lr1,XA1	CNL	154	0	0	0	0	0
Pikm2-TS	CNL	148	0	0	0	0	0
Rp1-D	CNL	118	0	0	0	0	0
Rps4	TNL	0	0	0	0	114	0
Sw-5	CNL	0	0	0	0	0	99
Pm3, Rdg2a	CNL	109	0	0	0	0	0
Pi2,Pi9,Piz-t	CNL	90	0	0	0	0	0
Cre1;Lr21	CNL	77	0	0	0	0	0
RRS1	TNL	0	0	0	0	64	0
Tm-2, TM2a	CNL	0	0	0	0	0	39
Dm3	CNL	0	0	0	0	0	0
P2	TNL	0	0	0	0	0	0
Pikm1-TS, Pikp-2	CNL	30	0	0	0	0	0
Pi5-2	CNL	24	0	0	0	0	0
Pi5-1	CNL	22	0	0	0	0	0

VAT	CNL	0	0	0	10	0	6
RLM3	TN	0	0	0	0	3	0

Table 2. Number of genes found in orthogroups containing a cloned R genes in six major crops family (Poaceae, Fabaceae, Rosaceae, Cucurbitaceae, Brassicaceae, Solanaceae).

Intriguingly, the largest orthologues group was highly duplicated in Poaceae and not showed homologs in the other five investigated families (Table 2). In contrast, the second most populated group, containing four solanaceous TNLs (Gro1.4, N, RY-1, Bs4), was highly duplicated in Fabaceae, Solanaceae and Rosaceae, lacking any homologs in Poaceae. This group is very important because includes R genes conferring resistance to bacteria, virus and nematode. The Rpi-blb1 group is represented in all crops except for crucifers and it is highly duplicated in the legume and in the nightshade families. ADR1 group is conserved in all family of crops and its members are also present in early land plant lineage (78 genomes). In land plants, the number of these genes per genome varies from 1 to 5 excluding the pine and spruce genomes, which have respectively 32 and 35 ADR1 homologues. ADR1 mediates resistance against *Hyaloperonospora parasitica* in a salicylic acid-dependent manner, this gene had a characteristic domain, named RPW8, which is also detected in several genes grouped in the same orthogroup. Several orthologs of Fom-2 gene have been found in Rosaceae (62) and Solanaceae (81) families. Interestingly, of 587 homologs of Fom-2, 194 was detected in coffee genome. NRG1 copies are conserved in 48 analyzed eudicot genomes belonging five crop family, with a number of genes ranging from 1 to 24. NRC homologues are present in all superasterid genomes analyzed so far and are highly conserved in nightshades (Table 2). Grasses had several private orthologues groups, many of which include cloned genes conferring resistance to different fungi. Similarly, nightshades possessed several highly duplicated private groups, containing cloned genes conferring resistance to different pathogens (bacteria, nematode, fungus etc.).

NLR physical clusters

A NLR cluster identification was conducted on 46 (out of 102) assembled plant genomes. According to Luo et al. (2012) NLR-like genes separated by no more than eight non-R genes were considered part of same gene cluster. About 70% (12,902) of NLRs occurred

in 3,465 clusters on 46 assembled genomes (Table 3). The size of NLR clusters varied significantly, from a two to several dozen of genes. The largest NLR cluster was detected in the *Eucalyptus grandis* genome and contained 50 NLR genes. Interesting, 60% of these genes were homolog to KR1, a R-gene that confer resistance against soybean mosaic virus (SMV). The KR1 protein consists of a Toll/interleukin receptor (TIR) domain, a nucleotide binding site (NB) domain, an imperfect leucine-rich repeat (LRR) domain and two C-terminal transmembrane segments. Similarly, barrelclover genome contains many KR1 clusters (112).

Organism	n. of clusters	CNL	TNL	NL	N	L	T	TL	TN	CN	CL	total genes
Physcomitrella patens	65	4	2	28	19	87	0	0	0	7	2	149
Spirodela polyrhiza	16	23	0	38	0	9	0	0	1	1	0	72
Ananas cosmus	41	66	0	45	1	27	0	0	0	0	7	146
Brachypodium distachyon	74	113	0	95	1	18	0	0	0	3	4	234
Brachypodium stacie	59	85	0	68	2	15	0	0	0	1	3	174
Zea mays	24	19	0	33	1	7	0	0	0	0	1	61
Hordeum vulgare	49	29	0	56	2	15	0	0	0	1	5	108
Oryza sativa	109	172	0	157	0	29	0	0	0	0	6	364
Panicum hallii	72	79	0	116	1	21	0	0	0	0	0	217
Panicum virgatum	232	230	0	328	4	100	0	0	0	0	10	672
Setaria italica	91	129	0	149	0	23	0	0	1	0	2	304
Setaria viridis	75	100	0	111	0	17	0	0	0	0	3	231
Sorghum bicolor	81	95	0	115	1	24	0	0	0	0	2	237
Musa acuminata	21	33	0	31	0	10	0	0	0	0	0	74
Aquilegia coerulea	65	92	0	59	2	18	0	0	0	2	3	176
Vitis vinifera	77	98	19	133	1	72	0	7	0	0	9	339
Manihot esculenta	52	110	22	43	0	22	0	3	0	0	3	203
Populus trichocarpa	96	134	125	173	0	42	0	30	0	0	3	507
Salix purpurea	90	100	58	95	0	25	0	7	0	0	0	285
Glycine max	106	93	107	115	3	33	0	37	0	1	2	391
Lotus japonicus	80	15	45	68	5	49	0	34	1	0	3	220
Medicago truncatula	154	173	235	183	0	70	0	39	1	0	8	709
Phaseolus vulgaris	65	128	57	70	0	17	0	12	0	0	0	284
Trifolium pratense	111	96	58	98	0	55	0	44	1	1	9	362
Fragaria vesca	77	43	17	49	0	38	1	76	2	3	5	234

Prunus Persica	87	87	98	112	1	29	0	17	0	0	3	347
Citrullus lanatus	11	8	9	6	0	2	0	0	1	0	4	30
Eucalyptus grandis	169	190	256	231	1	157	0	74	0	2	3	914
Citrus clementina	61	181	92	81	3	24	0	17	0	2	3	403
Gossypium raimondii	59	135	24	63	0	29	0	1	0	4	0	256
Theobroma cacao	58	139	11	60	1	28	0	4	0	1	6	250
Arabidopsis thaliana	45	25	81	15	1	4	0	5	0	0	2	133
Brassica rapa	52	18	74	18	1	7	0	11	2	0	4	135
Capsella rubella	35	27	21	24	0	25	1	16	2	0	3	119
Beta vulgaris	40	37	0	44	2	28	0	0	0	0	9	120
Actinidia chinensis	27	19	0	33	2	25	0	0	0	0	4	83
Coffea canephora	147	257	5	326	12	68	0	0	0	8	3	679
Capsicum annuum Zunla	92	34	4	176	8	92	0	4	0	4	4	326
Capsicum anuum CM334	136	116	29	348	5	108	0	10	0	0	6	622
Solanum pennellii	49	38	13	63	2	19	0	4	0	2	0	141
Solanum lycopersicum	71	43	14	106	5	37	0	6	0	1	0	212
Solanum tuberosum phureja	106	88	30	184	1	68	0	16	0	0	4	391
Mimulus guttatus	58	94	0	150	0	19	0	0	0	0	1	264
Sesamum indicum	35	55	0	68	0	19	0	0	1	0	3	146
Lactuca sativa	109	52	212	117	1	52	0	35	0	1	1	471
Dacus carota	36	23	1	51	1	24	0	5	0	1	1	107

Table3. Number of clusters and the types NLR-like genes in cluster across plant genomes.

The highest number of clusters (232) was recorded in Poaceae family with highest number in *Panicum virgatum* genome, in which the 55% (672) of all annotated NLR genes were organized in clusters (Table 3). Fourteen clusters included different NLR member copies and the biggest cluster counted 16 genes. About 15% (109) of clustered NLRs were located on chromosome 2 and 3 and were homolog to MLA; R-gene that confer resistance against Powdery Mildew. A high number of clusters was also found in Fabaceae and in Asterids species. Interesting, the higher frequency of clustered genes (87% of annotated NLRs) was recorded in clementine genome. Sixteen for cent (10 out of 61) of clementine clusters contained more than 3 genes and the largest cluster contained 40 NLRs homolog to N.

Organism	MLA1-6-10-12-13, Pi36, Lr10	Gro1.4, N, RY-1, Bs4	RPM1	FOM-2	KR1	RPS5, Rps2	Pid3
Physcomitrella patens	0	0	0	0	0	0	0
Spirodela polyrhiza	0	0	0	0	0	15	2
Ananas cosmus	0	0	0	0	0	0	14
Brachypodium distachyon	51	0	0	0	0	0	7
Brachypodium stacie	30	0	1	0	0	0	9
Zea mays	8	0	0	0	0	0	8
Hordeum vulgare	15	0	0	0	0	0	4
Oryza sativa	52	0	1	0	0	0	17
Panicum hallii	49	0	0	0	0	2	10
Panicum virgatum	109	0	0	0	0	0	19
Setaria italica	49	0	0	0	0	0	25
Setaria viridis	38	0	0	0	0	0	17
Sorghum bicolor	39	0	0	0	0	0	15
Musa acuminata	0	0	2	0	0	0	0
Aquilegia coerulea	0	0	0	0	0	10	0
Vitis vinifera	0	12	2	0	0	21	0
Manihot esculenta	0	25	10	0	0	3	0
Populus trichocarpa	0	78	14	15	2	0	0
Salix purpurea	0	37	14	32	1	3	0
Glycine max	0	36	42	0	58	0	0
Lotus japonicus	0	14	10	0	15	2	0
Medicago truncatula	0	69	43	0	112	3	0
Phaseolus vulgaris	0	22	7	0	14	0	0
Trifolium pratense	0	36	35	0	13	0	0
Fragaria vesca	0	2	9	5	1	1	0

Prunus Persica	0	44	35	13	33	4	0
Citrullus lanatus	0	0	0	5	4	0	0
Eucalyptus grandis	0	8	26	0	140	18	0
Citrus clementina	0	112	7	0	0	77	0
Gossypium raimondii	0	7	1	3	0	19	0
Theobroma cacao	0	6	6	5	0	7	0
Arabidopsis thaliana	0	0	0	0	0	15	0
Brassica rapa	0	0	0	0	0	14	0
Capsella rubella	0	0	0	0	0	15	0
Beta vulgaris	0	0	4	2	0	0	0
Actinidia chinensis	0	0	3	0	0	0	0
Coffea canephora	0	9	65	164	0	0	0
Capsicum annuum Zunla	0	19	2	1	0	0	0
Capsicum anuum CM334	0	24	3	0	0	0	0
Solanum pennellii	0	6	0	3	0	0	0
Solanum lycopersicum	0	10	0	1	0	0	0
Solanum tuberosum phureja	0	20	0	18	0	0	0
Mimulus guttatus	0	0	0	8	0	0	0
Sesamum indicum	0	0	0	4	0	0	0
Lactuca sativa	0	6	0	20	0	0	0
Dacus carota	0	0	0	18	0	0	0

Table 4. Number of gene in cluster and related reference R-gene across plant genomes.

We investigated also the cluster occurrence of genes belonging to orthogroups referred to cloned R-genes (Table 4). The highest gene clustering event regard the Fom-2 orthologs genes, with 164 clustered genes in coffee genome. In addition, in *S. tuberosum* and *S. pennellii* genomes were observed four clusters of FOM-2 homologs. A conspicuous

number of RPS5 and RPS2 orthologs clustered in clementine. Furthermore, a grouping event in a specific genomic region of a basal monocot genome (*Spirodela polyrhiza*) was observed. Indeed, Pid3-like genes clustered in monocots genomes from the basal of monocotyledon lineage. Many RPM1-like genes occur in cluster in coffee (65), and barrelclover (43) and soybean genomes (42). Moreover, NRC-like genes showed clusters in all superasterid analyzed genomes except than in *A. chinensis* genome.

2.4 Discussions

The investigation of the diversification occurred in the pathogen resistance genes can provide important insights for the plant improvement. For this reason, the identification and evolution of genes involved in plant disease resistance have been hot topics in genetic field from the first cloned NBS-LRR gene (Whitham et al. 1994). In the present study, we performed a NLR genome-wide comparative analysis in 102 species (including bacterial, algal and plant genomes) and to understand the mechanisms of duplication, evolution and diversification.

Many studies have demonstrate the essential role of major domains (e.g. NBS, LRR) in the resistance function of NLR genes in plant (Dangl & Jones 2001). An important biological question to address is how and when such domains were fused to make novel functional proteins (Marone et al. 2013). In this study, in bacteria, archaea and algae genomes only independent NB or LRR domains were detected, the first NLRs was detected from *M. polymorpha* and *P. patens*, confirming the theory, which NB and LRR coding sequences are fused in the basal land plants. The mechanisms by which this domain arrangement supported functional innovations at this evolution stage is still obscure. The NLRs of *Marchantia polymoprha* not showed high similarity with cloned resistance genes, in contrast, one gene of *P. patents* is highly similar to ADR1, R gene that conferee resistance against *Hyaloperonospora parasitica* in a salicylic acid-dependent manner. Therefore, such results support the hypothesis that at this stage NLR became an immune resistance activator with a resistance function (Zhong & Cheng 2016).

The composition of NLR protein classes and number of NLR genes is very variable at both the species/genera and the family/order levels. For example, in this study the number of detected R genes is variable across pepper genomes (Zunla-1 and CM334), Solanum

spp., Arabidopsis spp., crucifers, grasses and in other botanical groups. Several plant genomes sequenced have small number of NLR genes, such as cucurbits, *Z. marina* and *O. thomaeum*. These two last species live in extreme conditions with few competitors and they had to adapt to survive with new structural and physiological challenges in the sea (*Z. marina*) (Olsen et al. 2016) or desert conditions (*O. thomaeum*). Probably, the lower amount of NLR in these grasses is due to the absence of a real prolific pathogen. In opposite, we notice an increase of copy number of NLRs gene in Brassicales order from the ancestor closest papaya to Brassicaeae species, due mainly to the expansion of TNL class. Similar expansion was observed between in plant genomes of Liliopsida class, in particular from *S. polyrrhiza* and Poaceae species. Generally, a genome duplicates resistance genes to increase the variability of resistances to different pathogens, though maintain a higher amount of R homologs has a greater fitness cost (Tian et al. 2003). However, in some cases it remains unclear why the number of R genes can vary drastically between different plant species.

Monocots and several eudicot genomes lack of TNL genes (Jacob et al. 2013). It is also interesting note that the TNLs/CNLs ratio is different across plant families, e.g. in crucifers is almost one instead in nightshades is it much lower. Interestingly, the presence and the number of some CNLs, in particular atypical and helper NLR, is correlated to TNLs presence. Indeed, *Aquilegia coerulea*, *Beta vulgaris, Mimulus guttatus*, *Sesamum indicum* and Poaceae genomes lack both TNLs and NRG-homologs. The CNL copies of NRG were not present in the plant species lacking TNLs, suggesting a correlation with this class of genes. In fact, NRG1 is required for the functioning of N (TNL) resistance gene (Peart et al. 2005). NRC homologs showed to be a helper of several R genes (Gabriëls et al. 2007; Wu et al. 2016) in superasterid lineage, an order with a lower number of TNLs in opposite to CNLs. By the contrast, the Pb1 gene, that encodes for an atypical CNL gene (Hayashi et al. 2010) showed several copies in monocots genomes. Therefore, it seems that the plant genomes counterbalanced the absence or the lower copies of TNLs altering their intrinsic function or facilitating functional innovations by combining with other proteins. The great part of detected genes have a similarity with cloned resistance genes, suggesting that a basic protein structure is selected, duplicated and diversified from plant genomes for an effective resistance function. However, several larger orthogroups lack homology to reference R genes, and need be investigated further.

The cloned R-genes (MLA, PI36, LR10 and PIB) that confer resistance to main grasses disease were highly duplicated and clustered in the Poaceae genomes. Homologs to N and Ry-1 genes, conferring resistance to disease caused by viral pathogen, had many clustered copies on clementine, a species high challenged by Citrus tristeza virus. Similarly, Fom-2-like genes are highly clustered and duplicated in coffee genome and one of most dangerous pathogen of coffee crop is *Fusarium xylarioides*, a vascular fungus. We showed that the great expansion of some gene copies was mainly due to the clustering ecent occurred in specific chromosomic regions. Such findings support the hypothesis that the formation of new R genes is mediated by genomic destabilization and consequent genomic rearrangements in the presence of these genes (Spoel & Dong 2012).

The great diversification observed in the NLR genes indicates that they are very dynamic genomic elements. These genes represent a powerful weapon to detect the presence pathogen effectors and triggering the plant immunity response. The expansion of resistance gene families across the entire land plant lineages and their frequent recombination provide a powerful arsenal for land plants, which if properly used can become an indispensable tool for humans and their agriculture production.

3. COMPARISON OF SOLANACEAE ORTHOLOGOUS PATHOGEN RECOGNITION GENE-RICH REGIONS

3.1 Introduction

The Solanaceae family comprises more than 3,000 species and includes major food crops such as tomato (*Solanum lycopersicum*), potato (*Solanum tuberosum*), eggplant (*Solanum melongena*) and pepper (*Capsicum annuum*). In recent years several important Solanaceae species have been sequenced (Fernandez-Pozo et al. 2015) and wild species genomes investigated by sequencing and resequencing (Aflitos et al. 2014; Aversano et al. 2015; Qin et al. 2014). Solanaceae species have also served as ideal systems for studying the genetics and molecular basis of plant resistance mechanisms (Ercolano et al. 2012). Given the lack of extensive studies on the eggplant and pepper genomes regarding pathogen recognition genes (PRGs: RLP - receptor-like protein, RLK - receptor-like kinase and NLR - nucleotide-binding leucine-rich repeat), information gathered from other species can be used to steer investigation in such species.

According to comparative genomic studies, Solanaceous species share extensive syntenic regions (Wang et al. 2008): many of the loci involved in disease resistance in tomato have putative orthologues, in potato, eggplant and pepper in corresponding positions (Grube et al. 2000; Vossen et al. 2014). Further characterisations have demonstrated that tomato PRG homologues in potato, tobacco and pepper are subject to dramatic reshuffling (G. Andolfo et al. 2013; Wei et al. 2016; Seo et al. 2016).

Comparison of the spatial arrangement of genes in different genomes raises important questions on how complex biological systems evolve and function. Spatial analyses of the orthologous genomic region can unravel selection processes and species history (Hoberman & Durand 2005). Over time the large- and small-scale rearrangements occurring in orthologous loci have shaped the specific genomic architecture of different species (Yeaman 2013). To reconstruct the direction and magnitude of evolutionary trajectories of a given gene family, it is critical to detect the ancient loci that can lead to the formation of gene clusters (Luo et al. 2012; Baumgarten et al. 2003).

In previous studies PRG clusters were identified using approaches based both on identifying a genomic region containing a number of PRGs or on a set of genes that delimit an interspace (Richly et al. 2002; Hoberman and Durand 2005). A more rigorous investigation of well-known Solanaceae PRG clustered loci dynamics can help to understand how their arrangement can impact disease-specific responses, since all functional resistance genes found in tomato and potato are included in a cluster (Andolfo

et al. 2014). The recently developed target-enriched strategies for next-generation sequencing could facilitate the analysis of PRG-rich regions (Grover et al. 2012). It has already been demonstrated that targeted sequencing can expand our knowledge of PRGs (Andolfo et al. 2014). Furthermore, targeted sequencing of selected PRG loci combined with availability of high-quality reference genome sequences can offer insights into the mechanisms of PRGs evolution (Gasc et al. 2016; Witek et al. 2016; Steuernagel et al. 2016). In this work, the annotation of NLRs, RLPs and RLKs coding genes in eggplant and pepper genomes was performed. The PRG clustered loci rearrangement arose was reconstructed using multiple and combined methods and a genome-wide comparative map, in the three Solanaceae species, was also realised. Finally, Solanaceae loci containing functionally characterised PRGs in such species were explored by targeted sequencing and microsynteny analysis.

3.2 Materials and methods

Pathogen recognition gene family annotation

A script developed in-house to identify tomato and potato PRG proteins by Andolfo et al. (G Andolfo et al. 2013) was implemented in this study. The HMM profiles were used to screen the pepper and eggplant proteomes (http://peppersequence.genomics.cn; http://eggplant.kazusa.or.jp) to identify pathogen recognition proteins. This proteins set was further analysed using InterProScan v5 (Jones et al. 2014) to verify the presence of characteristic domains of pathogen recognition proteins (CC: coiled coil; NB: nucleotide binding; LRR: leucine rich repeat; TIR: Toll/interleukin-1 receptor; Kin: kinase; eLRR: extracellular-leucine rich repeat; TM: transmembrane).

Cluster analysis

The calculation of local gene enrichment was conducted using two methods: an arbitrary max gap approach (MG) (Hoberman et al. 2005) for identifying the spatial arrangement of genes with similar functions separated by a gap of no more than eight non-R genes (Luo et al. 2012; Richly et al. 2002), and a sliding window (SW) approach using REEF

software (http://telethon.bio.unipd.it/bioinfo/reef/) that is able to identify chromosome regions of a given size that contain a number of adjacent PRGs based on a statistical test on genomic distribution (Coppe et al. 2006). Both analyses were conducted on ITAG Tomato Genome 2.3 (http://solgenomics.net) and Zunla-1 Pepper Genome 2.0 (http://peppersequence.genomics.cn). In particular, sliding window scanning was conducted with a setting size of 0.5 and 1 Mb, a Q-value of 0.05 and a shift length of 50 Kb. For both methods, we also varied the minimum gene number cut-off from two to four genes to highlight the clustering tendency of specific regions.

Targeted sequencing

Targeted sequencing workflow is illustrated in Figure1. The experiment was executed on four plant species: *S. lycopersicum* var. Pyrella (Sl-Pyrella), *S. peruvianum* 10543 (Sp-10543), *S. melongena* var. Cima Viola (Sm-Cima Viola) and *C. annuum* 1014 (Ca-1014). Sp-10543 is resistant to *Pyrenochaeta lycopersici*; Sp-10543 is resistant to *Meloidogyne incognita;* Sm-Cima Viola is tollerant to *Verticillium sp.;* Ca-1014 is resistan to Potato virus Y.

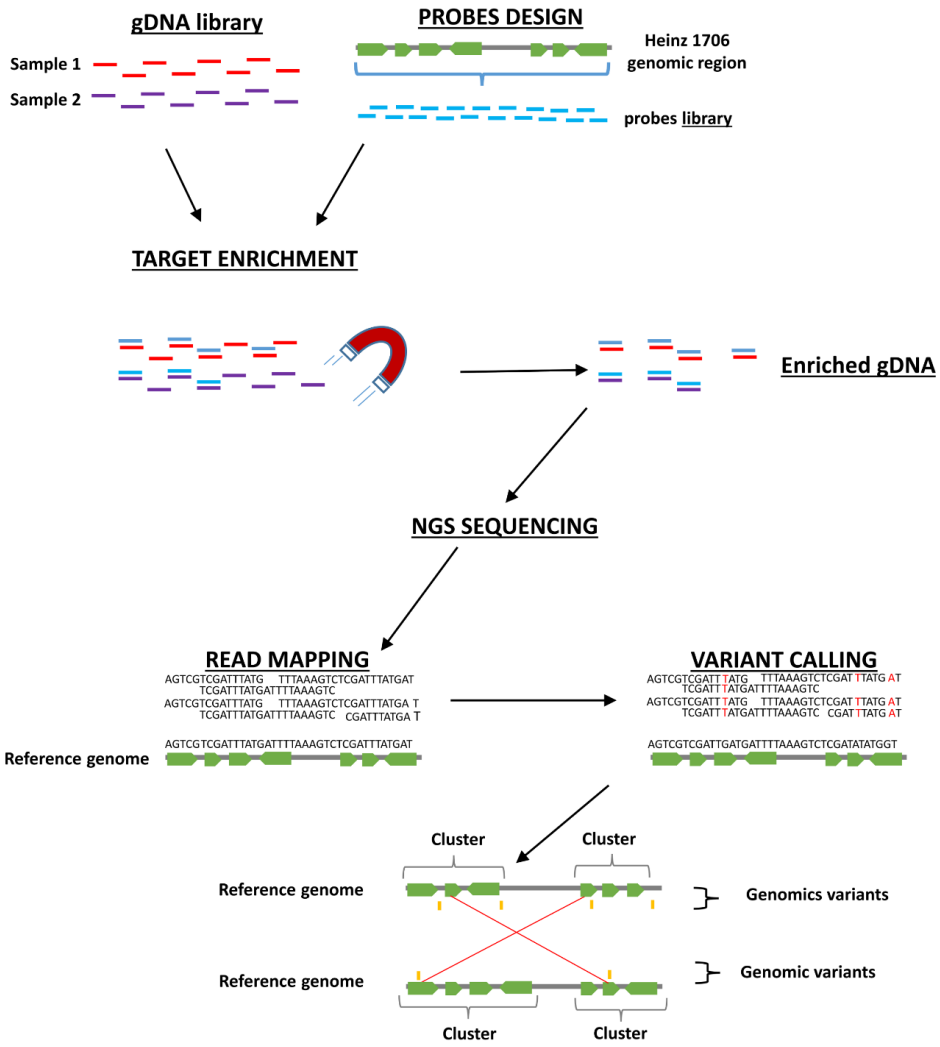


Figure 2. Workflow of the targeted sequencing and variant-calling experiment.

Fully expanded leaves were detached from three-week-old greenhouse-grown plants of each accession. Genomic DNA was extracted from young leaf tissue of the same plants, using the DNeasy Plant Mini kit (Qiagen Valencia, USA), following the manufacturer's instructions.

A custom-designed SureSelect Target Enrichment Kit (Agilent Technologies, Santa Clara, CA) was used to capture selected regions, according to the manufacturer's instructions. The probe library was designed on 14 selected regions of the ITAG v2.3 genome, identified by Andolfo et al. (2013) for a total of 5.7 Mb (see supplementary material). Library quality was determined using the Agilent High Sensitivity DNA kit on the Agilent 2100 bioanalyzer. Libraries were pooled at equimolar concentrations and

sequenced with TruSeq PE Cluster Kit v3 and TruSeq SBS Kit v3 (Illumina, San Diego, CA) on Illumina HiSeq 1000 sequencer (Illumina, San Diego, CA) generating 100-bp paired-end reads.

Mapping and variant calling

Read mapping was performed using BWA [24] on the reference genomes: ITAG Tomato Genome 2.3 (http://solgenomics.net/) for samples of Sl-Pyrella and Sp-10543; draft genome of eggplant v2.5.1 (http://eggplant.kazusa.or.jp) for the Sm-Cima Viola sample; Zunla-1 Pepper Genome 2.0 (peppersequence.genomics.cn) for the Ca-1014 sample. This procedure was executed for all samples with t = 10 and default setting. Adapter sequences were removed from sequence reads using Scythe software with default parameters (v. 0.994; https://github.com/vsbuffalo/scythe). Low quality read ends were trimmed using Sickle software (https://github.com/vsbuffalo/scythe).

Variants between four genotypes (*S. lycopersicum* var. Pyrella, *S. peruvianum* 10543, *S. melongena* var. Cima Viola and *C. annuum* 1014) and reference genomes relative were called using SAMtools software (Li & Durbin 2009) with a minimum read depth threshold of 20. Identified variants were annotated using SnpEff v3.4 (Cingolani et al. 2012) to predict their effect on the genes, using reference genomes annotations. Finally, polymorphisms were subsequently filtered for position, considering only those that were in a synteny block at the region was used for probe design indicated in eggplant and pepper sequencing work (Hirakawa et al. 2014; Qin et al. 2014).

Comparative analysis

Orthology analysis was conducted on proteomes of eggplant (SME_r2.5.1_pep_ip), tomato (ITAG2.3_proteins) and pepper (Capsicum.annuum.L_Zunla-1_v2.0_PEP) using Inparanoid Software and Multiparanoid Software with default parameters (Alexeyenko et al. 2006; Remm et al. 2001). We used a confidence score threshold = 1 to directly estimate orthology relationships between the identified PRGs.

A comparative map was constructed, merging results obtained from PRG prediction, cluster dataset and orthology results using the Circos package (Krzywinski et al. 2009).

Eggplant chromosomes were assembled using information reported in Hirakawa et al. (Hirakawa et al. 2014). Position and IDs of isolated and studied loci were reported by Andolfo et al. (G Andolfo et al. 2013) and in the Sol Genomics Network portal (https://solgenomics.net/).

Nucleotide sequences of clustered PRGs in genomic sequenced regions 4 and 10 were extracted from genomes and aligned using EINS-i algorithm of MAFFT (Katoh & Standley 2013) software. The procedure was performed separately for receptor-like proteins (RLP and RLK) and NLR genes coding; sequences with low global alignment identity (< 25%) were discarded. The phylogenetic relationships of predicted Solanaceae PRGs were inferred separately for each structural class (e.g. NLRs, RLPs) using MEGA 6 software with the maximum likelihood method general time reversible model. The bootstrap consensus tree of 100 replicates was taken to represent the evolutionary history of the sequences analysed (Tamura et al. 2013). Transposable elements were found using BLASTn search (Camacho et al. 2009) on library of transposable element reported on Pepper Genome Database (http://peppersequence.genomics.cn/page/species/download.jsp). Specific gene and protein alignments were generated in Geneious R6 platform (Kearse et al. 2012).

3.3 Results

Chromosome distribution of pepper and eggplant pathogen recognition genes

A total of 1097 and 775 pathogen recognition genes were identified in *C. annuum* Zunla-1 (Zun) and *S. melongena* Nakate-Shinkuro (Nak) genomes, respectively (Table 5).

	Protein domains*	S. melogena Nakate-Shinkuro	C. annum Zunla-1	S. lycopersicum Heinz 1706
Full-length	CC-NB-LRR	72	73	101
	TIR-NB-LRR	18	5	19
	NB-LRR	65	87	59
	RLP	185	291	168

	RLK	233	338	263
Total full-length		573	794	610
Partial	CC-NB	37	38	17
	TIR-LRR	3	-	2
	TIR-NB	3	1	6
	NB	66	101	52
	TIR	15	8	11
	LRR	78	155	62
Total partial		202	303	150
Total		775	1097	760

*Table 5. Classification of S. melongena and C. annuum pathogen recognition genes that encode domains similar to plant R proteins. * RLP: receptor-like protein; RLK: receptor-like kinase; CC: coiled coil; NB: nucleotide binding; LRR: leucine rich repeat; TIR: Toll/interleukin-1 receptor.*

The chromosome distribution of PRG classes retrieved in the two analysed species in comparison with the tomato PRGs profile obtained by Andolfo et al. (G Andolfo et al. 2013) is shown in Figure 3.

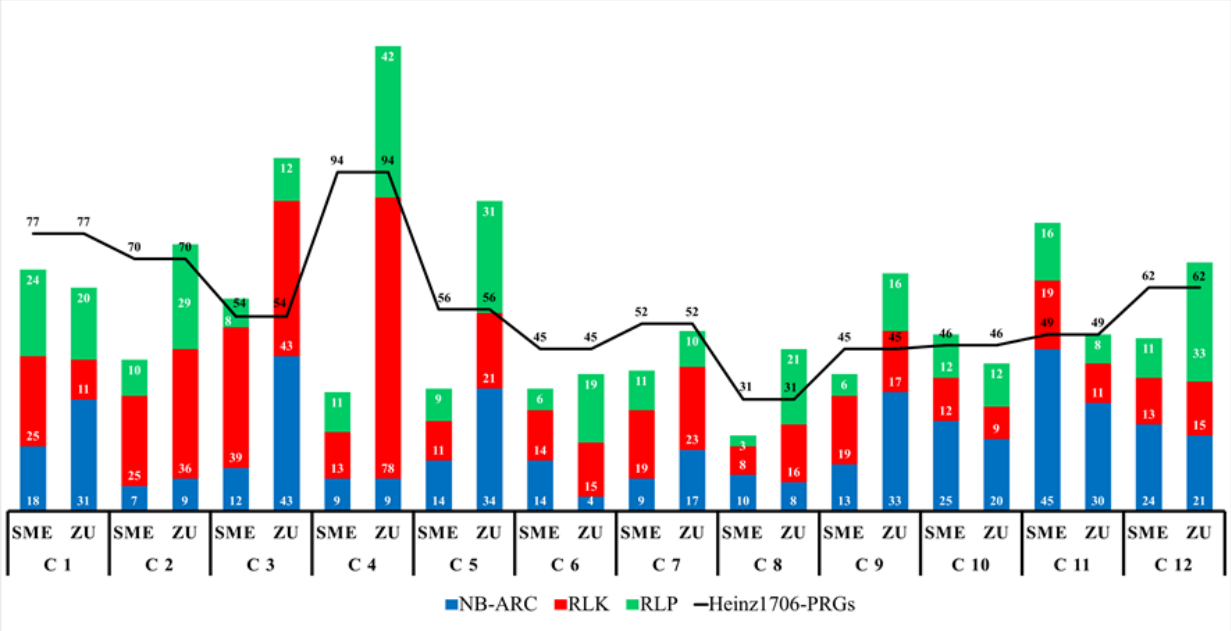


Figure3. Chromosome distribution of main pathogen recognition protein classes (RLP: receptor-like protein, RLK: receptor-like kinase and NLR: nucleotide-binding leucine rich repeat) of S. melongena Nakate-Shinkuro (NAK) and C. annuum Zunla-1 (ZUN). For comparative propose we reported the total number (black line) of PRGs identified in S. lycopersicum Heinz 1706 (Heinz 1706-PRGs) for each chromosome by Andolfo et al. 2013.

The total number of genes varied along the chromosomes of each species. Chromosome 4 was enriched in PRGs both in tomato and pepper (94 vs 129) whilst the highest number of genes in eggplant was observed on chromosome 11. A large variation in genome arrangement was observed for the assessed classes in pepper and eggplant. The RLK class showed the largest variation in pepper, ranging from the 78 members on chromosome 4 to 9 members on chromosome 10. Instead, eggplant showed a similar RLP (3 to 24) and RLK (8 to 25) trend distribution. NLR path distribution showed marked differences between the two species: in pepper a conspicuous number of NLR-related genes were identified on chromosomes 3 (43) and 5 (34) whilst in eggplant they are enriched on chromosome 11 (45). Eggplant and pepper genomes also displayed a large number of PRGs (157 and 173 respectively) located on chromosome 0, which could lead to a bias in genome distribution.

Identification of resistance orthologous groups

To translate the resistance gene information from the tomato model species to other Solanaceae crops, we performed orthology prediction analysis. The three species shared a core set of 320 selected as *bona fide* resistance orthologous groups (ROGs), including 1076 PRG proteins (Figure 4), of which 182 were RLP, 297 NLR and 498 RLK.

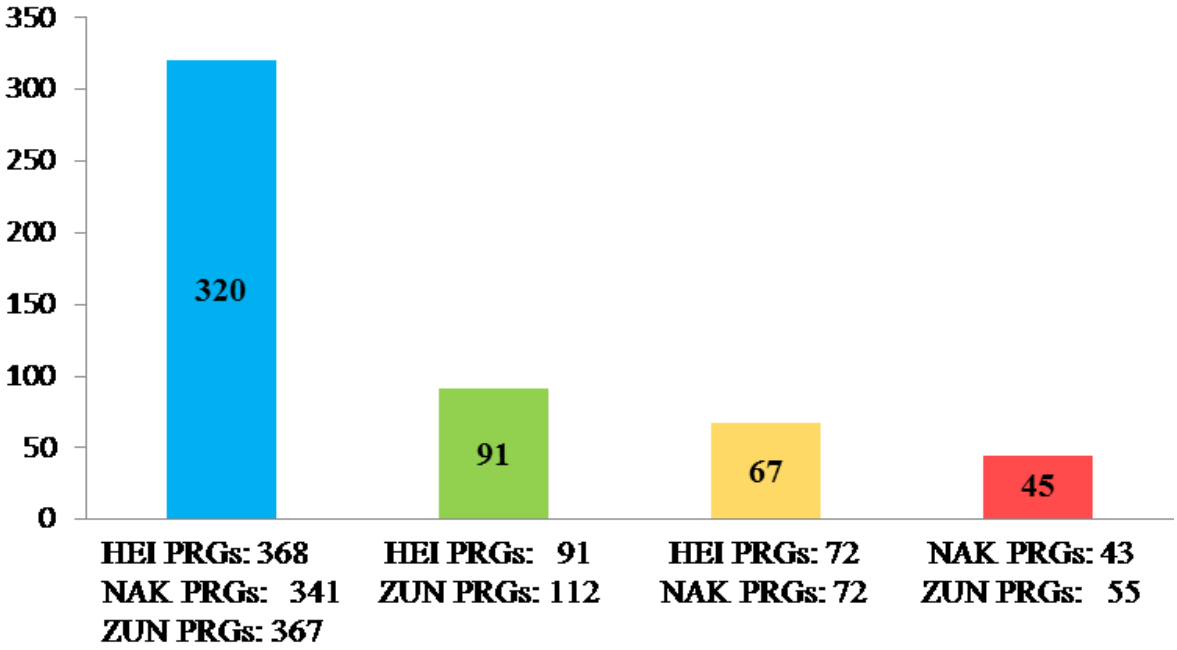


Figure 4. Bar chart of Solanaceae resistance orthologous groups (ROGs). Three species (pepper, tomato and eggplant) were used to generate the diagram. The shared ROGs between pepper-tomato-eggplant, tomato-pepper, tomato-eggplant and pepper-eggplant in cyan, green, yellow and red bars were showed, respectively. The number of pathogen recognition genes (PRGs) of common groups for each plant species (HEI: tomato; NAK: eggplant; ZUN: pepper) is reported below the horizontal line.

In common ROGs a larger number (89, 95) of PRG paralogues were shared between *C. annuum* and *S. lycopersicum*. Furthermore, 91 ROGs were only shared between tomato and pepper genomes and 67 between tomato and eggplant. Orthology analysis also allowed ortho-groups containing homologues of cloned genes to be found. Six very important Solanaceae *R*-gene loci (Cf9, Hero, Prf, Tm2 and Mi1.2) showed a highly confident orthologue in each three species analysed. Two orthologues to tomato LeEIX1 and LeEIX2 homologues (Solyc07g008620 and Solyc07g008630), in pepper and eggplant genomes were identified (Capana00g004962 and Sme2.5_01783.1_g00006). The tomato RpiBlb1 homologous gene (Solyc08g076000) presented two and one orthologues in pepper eggplant genomes, respectively (Capana01g000864 and Capana01g000870; Sme2.5_11213.1_g00002.1). Ve2 gene presented only two orthologues in tomato and pepper genomes (Solyc09g005080; Capana09g001153). Interestingly, a large diversification of I2 homologous genes was found in tomato and eggplant genomes. The I-2 gene homologues (Solyc11g071430- Solyc11g071420),

located on chr11, showed several paralogues, including a tomato LRR gene (Solyc11g065800) similar to Capana11g000417 and Sme2.5_15564.1_g00001.
Other tomato genes implicated in the resistance process, but not included in the principal four pathogen recognition protein classes, showed an orthologous relationship in pepper and eggplant genome, namely ASC-1 (locus name Solyc03g114600) which belongs to the same group as Capana03g000932 and Sme2.5_04467.1_g00005, located on chromosome 3, and the Mlo1 gene (locus name Solyc04g049090) which showed orthologues both in pepper and eggplant (Capana05g002411, Capana06g001935, Capana11g000102, Sme2.5_00266.1_g00009, Sme2.5_12945.1_g00001).

Detection of genomic regions rich in pathogen recognition genes

In order to investigate the genomic arrangement of pathogen recognition loci and to evaluate the grouping tendency of such genes in tomato and pepper genomes we performed both a sliding-window (SW) scan and max-gap (MG) analysis. The eggplant genome was not included in this analysis since its genome is assembled partially.
Taking into account the structural differences of the compared genomes, SW scanning was performed on a window of 0.5 or 1 Mb, varying the minimum gene number cut-off from two to four genes (Table 2). Tomato displayed from 39 to 66 clusters, with a number of genes varying from 5.6 to 6.9. The number of PRG clusters, as well as the number of genes included in a cluster, showed a higher variation in pepper, ranging from 106 to 59 and to 7 to 10, respectively. Using the MG method, we were able to identify from 40 to 146 clusters in tomato, with the number of genes per cluster varying from 3.2 to 5.7, and from 201 to 72 in pepper, with the number of genes per cluster varying from 3.9 to 6.6 (Table 6).

	HEINZ 1706			ZUNLA-1		
Sliding window analysis (0.5 M)	2 GENES	3 GENES	4 GENES	2 GENES	3 GENES	4 GENES
average n. genes for cluster	5,6	6,1	6,9	4,3	5,3	9,3
n. of clusters	66	61	50	147	101	59
n. of genes in cluster	368	370	344	632	540	412
NB-ARC genes in cluster	159	147	153	211	178	126
RLP genes in cluster	83	80	67	162	134	112

RLK genes in cluster	79	94	87	154	140	110
average lenght of clusters (bp)	281917	314068	355599	255315	313651	393368
Sliding window analysis (1 Mb)						
average n. genes for cluster	8,1	8,1	8,1	4,6	6,5	8,1
n. of clusters	39	45	47	106	91	62
n. of genes in cluster	317	365	382	597	591	504
NB-ARC genes in cluster	140	160	151	193	189	161
RLP genes in cluster	56	80	84	148	151	128
RLK genes in cluster	81	93	103	154	139	138
average lenght of clusters (bp)	721002	829940	780434	510068	647393	818992
Max Gap analysis						
average n. genes for cluster	3,2	4,5	5,7	3,9	5,3	6,6
n. of clusters	146	72	40	201	114	72
n. of genes in cluster	471	323	227	778	604	478
NB-ARC genes in cluster	192	142	100	250	205	164
RLP genes in cluster	110	74	49	197	156	120
RLK genes in cluster	112	72	54	206	150	121
average lenght of clusters (bp)	67980	106576	120984	407856	510818	655767

Table 6. Results of identification of PRG clusters using Sliding window (window size: 0.5 and 1 Mb) and Max Gap analysis.

The SW and MG analyses differ in the number of clusters identified and the number of genes included in a single cluster. In general, the average size (Kb) of tomato clusters using MG was lower than in SW. The higher data point match was obtained by using three genes per cluster as a cut-off both for MG and SW (with a sliding window setting of 0.5 Mb for tomato and 1 Mb for pepper). In particular, 23 and 57 clusters in tomato and pepper, respectively, showed exactly the same matching genes. Total MG and SW cluster-datasets were filtered for the presence of at least two orthologues among the analysed species. Following this criterion we detected four tomato clusters shared with pepper identified only by the MG approach and three clusters identified only by the SW approach. Therefore, to avoid data loss, we merged MG and WM data to obtain a list of conserved Solanaceae PRG clusters.

Comparative analysis of Solanaceae PRG groups

A comparative genomic map was obtained by merging the results obtained by PRG prediction (2632 PRGs), with cluster analysis results filtered for orthology (176 clusters) (Figure 5).

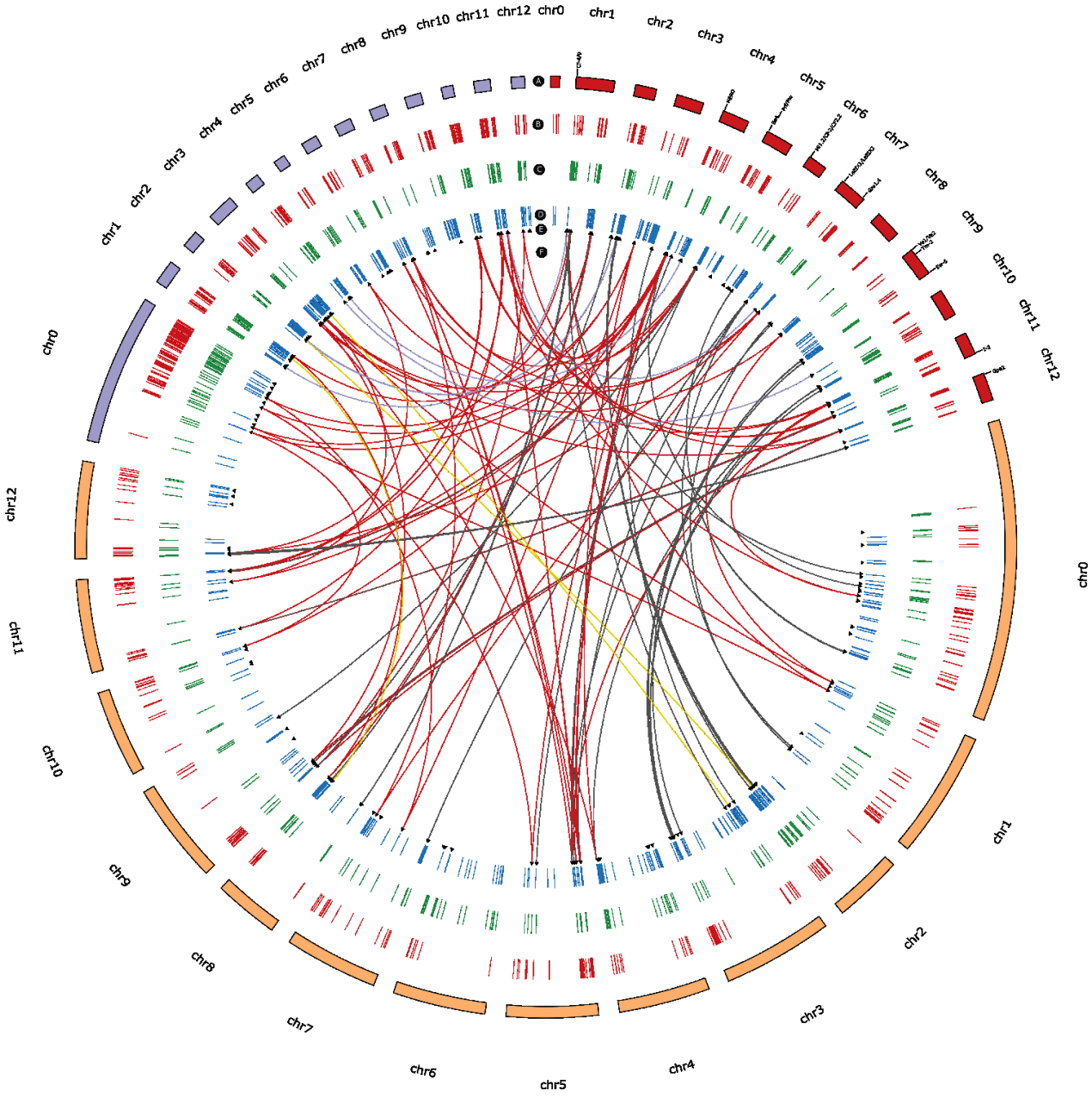


Figure5. Bar chart of Solanaceae resistance orthologous groups (ROGs). Three species (pepper, tomato and eggplant) were used to generate the diagram. The shared ROGs between pepper-tomato-eggplant, tomato-pepper, tomato-eggplant and pepper-eggplant in cyan, green, yellow and red bars were showed, respectively. The number of pathogen recognition genes (PRGs) of common groups for each plant species (HEI: tomato; NAK: eggplant; ZUN: pepper) is reported below the horizontal line.

PRG arrangement data of eggplant draft genome were computed by looking at least at two adjacent genes on a single scaffold. In *S. melongena* 86 PRG groups with an average size of ~22 Kbp were recorded. Seventeen tomato, 21 pepper and 22 eggplant clusters shared at least one PRG orthologue among all three species. Furthermore, 23 tomato clusters shared PRGs exclusively with 34 pepper clusters, and exclusively with five

eggplant clusters. In particular, the Cf4/Cf9 locus (tomato chromosome 1) showed orthologous genes in eggplant and pepper syntenic regions. Four PRG clusters detected on tomato chromosome 4 shared orthologous-PRG located on chromosome 5 and 11 in pepper and eggplant, respectively. In particular, Hero cluster share orthologous genes with a pepper cluster on chromosome 5. Two tomato clusters, located on chromosome 5 and containing BS4 gene and homologue R1 genes (Solyc05g007350, Solyc05g007610, Solyc05g007630 and Solyc05g007640), shared orthologous clustered genes on pepper chromosome 5. A tomato cluster on chromosome 5 (Solyc05g006570, Solyc05g006620, Solyc05g006630 and Solyc05g006670) is highly conserved in the pepper and eggplant genome. Clusters Pto and Prf shared orthologous genes with clusters located on pepper chromosomes 9 and 11. Clusters LeEix1 and LeEIX2 located on chromosome 7 showed orthology with a cluster on the corresponding chromosome in pepper. A tomato cluster containing Rblb1 homologues on chromosome 8 shares orthologous genes with clusters located on pepper chromosome 1 and eggplant chromosome 3. PRG organised in clusters on chromosome 9 have orthologues on pepper chromosome 3. Clusters Tm2 and Sw5 showed orthologous clusters on the corresponding chromosome in pepper. The I2 clusters located on tomato chromosome 11 showed orthologous clusters on chromosome 11 of eggplant and pepper. The tomato cluster flanked by Gpa2 markers showed orthologous clustered genes on pepper chromosome 9.

Sequence diversity in selected PRG orthologous loci

Fourteen orthologous loci of tomato, eggplant and pepper cultivated species (*S. lycopersicum* var. Pyrella: Sl-Pyrella; *S. melongena* var. Cima Viola: Sm-Cima Viola; *C. annuum* 1014: Ca-1014) and of a wild tomato species (*S. peruvianum* 10543: Sp-10543), were re-sequenced using a targeted sequencing approach (Figure 2).

A total of 33542210, 3094959, 32242713 and 33888476 sequencing reads were generated for Sl-Pyrella, Sp-10543, Sm-Cima Viola and Ca-1014, respectively (Table 7).

Target region ID	Genomic region length (bp)			Number of mapped reads on reference genomes			
	S. lycopersicum var. Heinz 1706	*S. melongena cv.* Nakate-Shinkuro	*C. annuum* Zunla-1	*S. lycopersicum* var. Pyrella	*S. peruvianum* 10535	*S. melongena* Cima Viola	*C. annuum* 1014

1	397939	553443	1602289	1290056	658805	177389	878031
2	380992	804748	8418917	1102759	630488	115428	1044283
3	409182	757886	1640039	1421008	779073	78260	210649
4*	434286	1339189	13771232	1333691	715046	293687	1456244
5	482675	847302	1636991	1517143	808522	125596	373254
6	369026	529512	3114100	1072281	564252	77492	162143
7	325151	324163	3065185	930927	525767	72020	1378748
8	531986	122141	2083026	1490319	770220	17332	827943
9	332217	696036	-	885939	512408	58542	-
10*	340231	431921	8506575	1113391	614764	139320	1055478
11	344709	1674430	-	1133286	706345	75888	-
12	402578	1207545	2135689	1305886	728515	180626	1271819
13	651409	1077628	1715216	1966412	1010844	192566	1824726
14	311256	362147	1218009	924032	475723	46980	521771
Total	5713637	10728091	48907268	17487130	9500772	1651126	11005089

*Table7. Syntenic genomic regions sequenced in tomato, eggplant and pepper genomes. The regions length and the number of mapped reds are related to three reference genomes (S. lycopersicum_ Heinz 1706; S. melongena_Nakate-Shinkuro; C. annuum Zunla-1). *Target regions used for microsynteny analysis.*

Using this approach, we analysed the evolutionary dynamics of 14 selected regions, containing PRGs putatively implicated in plant disease resistance.

The size of genomic regions captured and the number of reads mapping to the reference genomes are reported in Table 3. In each region, a large number of homologous genes to cloned resistance genes was found. The reads mapped on the respective genomic regions ranged from 17487130 to 9500772 in cultivated and wild tomatoes and from 11005089 to 1651126 in Ca-1014 and Sm-Cima Viola. All variants obtained with respect to the reference gene annotations were filtered for genome position, taking into account only those that were in the regions used for the probe design in pepper and eggplant syntenic regions. The number of variants identified in the analysed regions ranged from 101579 to 1484. The *S. peruvianum (*Sp-10543) sample showed the highest number of variants (Table 8), since the reads are mapped to the heterologous *S. lycopersicum* "Heinz 1706" genome. Ca-1014 also showed a larger quantity of variants (6,930), perhaps due to the greater length of syntenic regions (Table 8).

Name region	Chr	R-locus	PRGs annotation (Andolfo et al. 2013)	Coordinates		Lenght (bp)	Variants (SNPs; InDel)			
							S. lycopersicum var. Pyrella	*S. peruvianum 10543*	*S. melongena Cima Viola*	*C. annuum 1014*
1	1	Cf4_Cf9	6	336,518	734,457	397,939	560	8.192	97	183
2	4	Hero	8	1,645,385	2,026,377	380,992	103	6.244	163	524
3	4	-	15	2,479,262	2,888,444	409,182	76	8.147	72	113
4	5	L6	4	1,057,083	1,491,369	434,286	19	8.051	116	609
5	5	BS4_R1	4	1,820,981	2,303,656	482,675	80	9.835	180	105
6	5	Pto_Prf	4	6,182,878	6,551,904	369,026	23	5.141	90	70
7	6	Cf2_Cf5	2	1,989,526	2,314,677	325,151	62	5.475	96	2.058
8	6	Mi1.2_RpiBlb2	7	2,268,556	2,800,542	531,986	63	7.492	32	552
9	7	LeEix1_LeEix2	5	3,399,546	3,731,763	332,217	878	5.614	52	-
10	8	RpiBlb1	5	56,755,597	57,095,828	340,231	19	6.379	146	495
11	9	Ve1_Ve2	3	28,645	373,354	344,709	82	5.098	104	-
12	9	Tm2_Tm2a	5	66,612,472	67,015,050	402,578	92	7.972	112	245
13	11	I2_Rx2	7	51,345,297	51,996,706	651,409	325	11.953	186	1.548
14	12	Gpa2	6	2,565,757	2,877,013	311,256	69	5.986	38	428
Total							2.451	101.579	1.484	6.93

Table8. Detected variants (SNPs and InDel) of selected genomic regions for target sequencing experiment.

The I2 region showed a high level of diversification in all species. In tomato, regions 1, 2 and 9 also showed a conspicuous variation. In Sp-10543 five regions (1, 3, 4, 5 and 13) showed a number of variants up to 8000. In Sl-Pyrella and in eggplant, detected variants were nearly all homozygotes (88.8 % and 97.7 %) (Figure 6), while most of the wild tomato variants were heterozygotes (58.1 %).

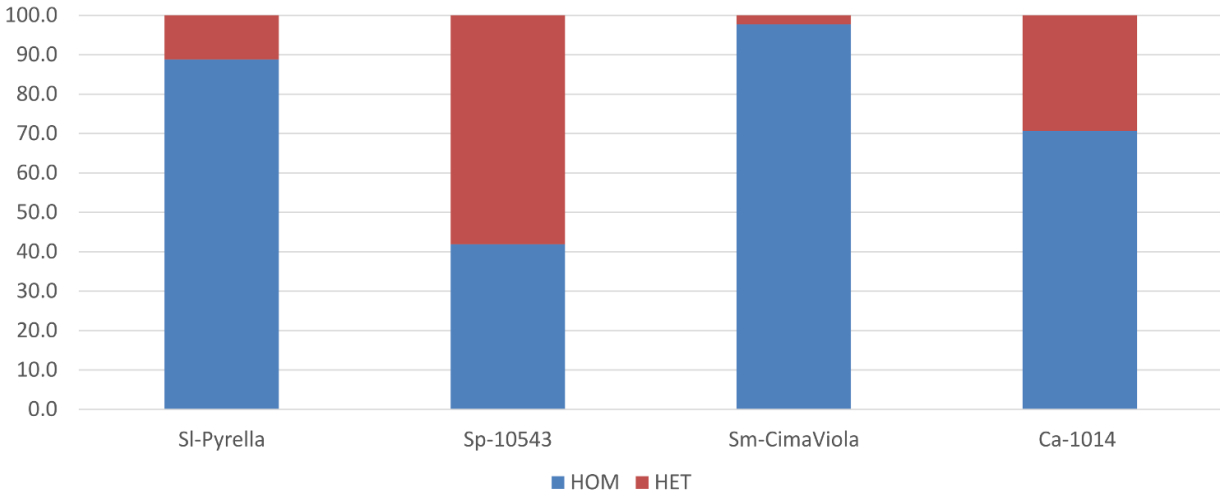


Figure 6. Proportion of homozygous and heterozygous SNPs assessed in four species. Zygosity classes are colour-coded as indicated. Accession IDs and the percentage of SNPs are indicated on the x and y axes, respectively.

Sl-Pyrella showed the highest percentage of INDELs with 33.83% of insertions and 9.55% of deletions (Figure 7).

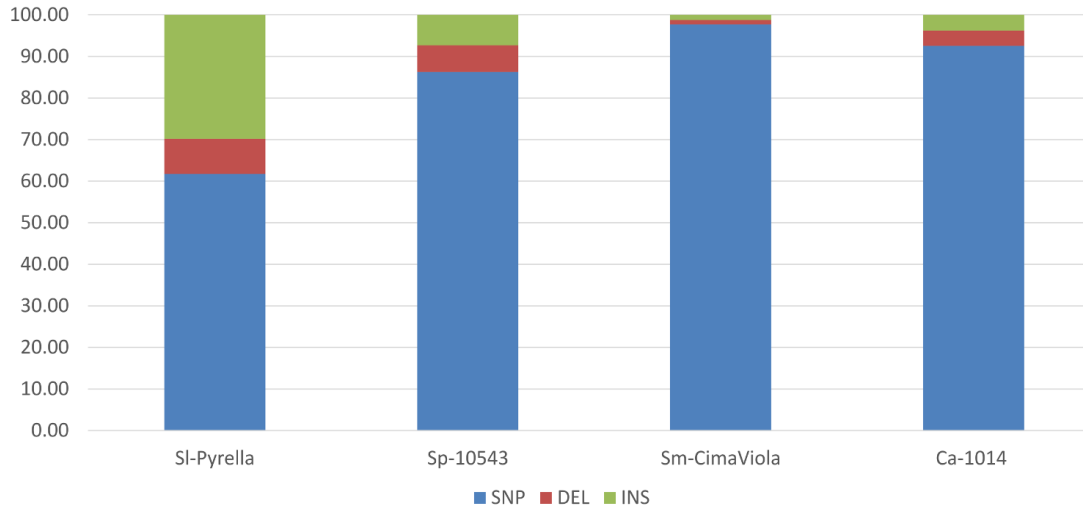


Figure 7. Proportion of DNA variant types assessed in the four species. Variant types are colour-coded as indicated. Accession IDs and the percentage of SNPs are indicated on the x and y axes, respectively.

When compared to the corresponding annotated genomes, we observed a significantly higher polymorphism frequency in intergenic regions than in genic regions for all species

except for Sm-Cima Viola (9.23%), which was the only sample with reads mapped on a draft genome.

	S. lycopersicum var. Pyrella .	S. peruvianum 10535	S. melongena Cima Viola	C. annuum 1014
Upstream	21	3121	25	101
Codon_change	0	18	0	1
Frame_shift	1	14	0	1
Intron	8	648	0	17
Non_synonymous_coding	5	1567	11	52
Splice_site_acceptor	0	2	0	0
Splice_site_donor	0	1	0	0
Splice_site_region	0	17	0	1
Stop_gained/lost	0	20	0	5
Synonymous_coding	2	980	19	13
Downstream	35	3080	25	52
Total	72	9468	80	243

Table9. Gene variant categories detected by targeted sequencing in pathogen recognition genes. Downstream and upstream variants located in coding sequences and their putative promoter regions (2Kb upstream the translation start site).

In all, 72, 9468, 80 and 243 variants of different types were detected on PRG genomic loci in Sl-Pyrella, Sp-10543, Sm-Cima Viola and Ca-1014, respectively (Table 9).

Microsyntenic PRG region reshuffling

The tomato PRG target region 4, located on chromosome 5, containing one RLP, one RLK and two TNLs, is highly conserved in pepper and eggplant (Figure 8).

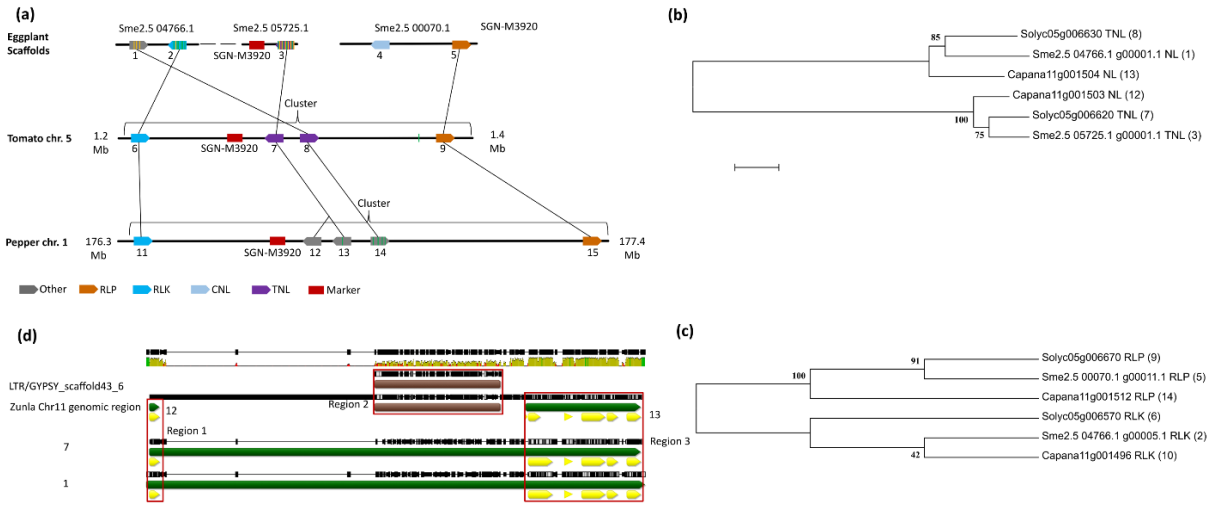


Figure 8. Reconstruction of the gene duplication history of target genomic region 4. (a) Schematic representation of postulated gene duplication events occurring in the genomic region. Detected gene variants are reported as green (low impact on coding gene) or yellow (medium impact on coding gene) ticks. Phylogenetic analysis performed using the maximum likelihood method, based on the general time reversible model, for homologous sequences of cytoplasmic (NLR; panel b) and transmembrane (RLP and RLK; panel c) receptor proteins. Bootstrap values are indicated above branches. (d) Multiple alignment of pepper, tomato and eggplant genomic sequences. We reported gene locus (green), the exons (yellow) and the transposable element (brown). Pepper region includes Capana11g001502 (12), Capana11g001503 (13) genes and a transposable element in the intergenic region while tomato (7) and eggplant (1) homologous genes lack any transposon insertion. Red rectangles display the most highly conserved sequences, indicated as Region 1 (Pairwise Identity 82.4%; Identical Sites 73,5%), Region 2 (Pairwise Identity 100%; Identical Sites 100 %) and Region 3 (Pairwise Identity 79.3 %; Identical Sites 70,9 %).

Orthologous genes in pepper show the same tomato order and orientation, except for an inversion occurring between Capana11g001503 and Capana11g001504 (Figure 8). By contrast, in eggplant a first inversion was observed between genes 1 and 3 (7 and 8 in tomato) and a second inversion between genes 1 and 2 (6 and 8 in tomato). Pepper genes 12 and 13 were orthologous to tomato loci 7 and 8, but unlike the latter, do not encode the TIR domain. The Solyc05g006620 gene in pepper was divided into Capana11g001502 and Capana11g001503, probably due to a transposable element insertion, as showed in figure 4, panel c. In eggplant changes in the protein structure of Sme2.5_04766.1_g00001 were recorded. Moreover, three missense SNP variants on gene Sme2.5_04766.1_g00001 and one gene Sme2.5_04766.1_g00005 were identified by targeted sequencing in the Sm-Cima Viola accession. The coding region of Sme 2.5_05725.1 g00001 displays six non-synonymous mutations. Non-synonymous

mutations were also identified in coding regions of Capana11g001503 and Capana11g001504 in the *C. annuum* 1014 genome. Instead, PRGs in this region were highly conserved in tomato (Pirella vs Heinz 1706).

The tomato target region 10, including a PRG cluster, located on chromosome 8 also showed a good level of collinearity with the pepper and eggplant genome. Figure 9 presents the region flanked by the marker est_ae501f12.

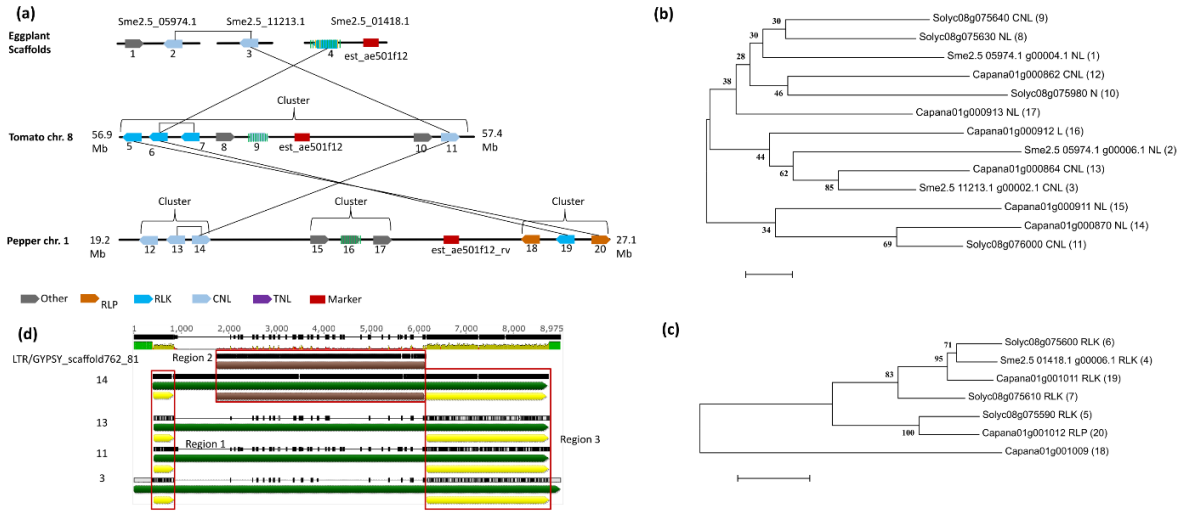


Figure 9. Reconstruction of the gene duplication history of target genomic region 4. region 10. (a) Representation of postulated duplication and of identified variants in green (low impact on coding gene) or in yellow (medium impact on coding gene). Phylogenetic analysis performed on homologous sequences of cytoplasmic (NLR; panel b) and transmembrane (RLP and RLK; panel c) receptor proteins. Bootstrap values are indicated above branches. (d) Multiple alignment of pepper (13,14), tomato (11) and eggplant (3) sequences and for each genes, we reports: locus (green), exons (yellow) and transposable element (brown). An insertion of transposable element on sequence of pepper gene 14 was evidenced. Red rectangles display highly conserved sequences, indicated as Region 1 (Pairwise Identity 78.3%;, Identical Sites 64.4%), Region 2 (Pairwise Identity 89.6%; Identical Sites 89.6%) and Region 3 (Pairwise Identity 75.8%; Identical Sites 85.6%).

This cluster in tomato consists of three RLKs, three CNLs and one NL, genes that are split into two clusters in pepper, separated by a third cluster without orthologous genes. An inversion was observed between tomato genes 5, 6, 11 and pepper genes 14, 19 and 20. The tomato gene 7 was originated by tandem duplication of gene 6. Discordant results were indicated in phylogenetic analysis and nucleotide sequence alignment due to a large TE insertion on the intron of Capana01g00870 (Figure 8). Gene 20 coding an RLP protein

showed partial orthology with an RLK gene in tomato. A similar occurrence was found for the NLR gene 14, lacking the CC domain and proving orthologous to full gene 11 (CC-NB-LRR). Several synonymous mutations were identified on gene 4 and gene 16 in Cima Viola (eggplant) as well as on tomato genes 8 and 9.

3.4 Discussions

The plant kingdom exhibits a large variation in PRG repertories among species but also among single individuals (Shao, Xue, et al. 2016). In our PRGs scanning of eggplant (Nakate-Shinkuro) genome showed more or less the same number of tomato resistance genes. By contrast, the PRGs of pepper Zunla-1 was contracted respect that pepper CM334 genome (Kim et al. 2014b), in terms of NLRs and RLKs. Genotype loci rearrangements could occur in response to specific phytopathogens. Indeed, CM334 pepper is resistant to *Phytophthora spp.* and potyviruses, whereas Zunla-1 is resistant to Fusarium wilt, *Phytophthora spp*. and Anthracnose (Qin et al. 2014). The variation in PRG number observed in Solanaceae was also found in other taxa, such as Rosaceae (Jia et al. 2015), Graminaceae (Li et al. 2010), Brassicaceae (Peele et al. 2014) and Fabaceae (Shao et al. 2014).

In our analysis the highest NLR concentration in pepper was evidenced on chromosome 5, that characterized by presence of QTL for resistance to *Phytophthora capsici* (Rehrig et al. 2014), whilst the eggplant genome showed the highest concentration of NLR genes on chromosome 11. On this chromosome in tomato, the I2 gene is located, as are other important resistant gene loci. The approach involving target-sequencing and co-localization with candidate *R*-genes may help to identify putative genes for major diseases also in this species (G Andolfo et al. 2013).

Our results confirm that most Solanaceae PRGs tend to be physically clustered. However, to explore the clustering tendency of such genes in several species it was necessary to perform a first analysis with a dynamic setting in order to identify the most suitable method. Indeed, the partitioning of genes into clusters could be hampered by genome architecture, including gene density and gene order (Hoberman & Durand 2005).

In some studies, following an ordinal data strategy, a locus with two or more PRGs separated by a number of non-PR genes (Richly et al. 2002; Luo et al. 2012) was

identified as a cluster. In other studies, following a spatial vision a gene cluster was defined as a physical region that contains more than three or more genes within 200 Kb or less (Holub 2001). We noted that by looking at least for three adjacent PRGs of 0.5 Mb in tomato and of 1 Mb in pepper most well-known Solanaceae clusters were identified. The task became more challenging when homologous regions were scrambled by rearrangement events that modified the global genome architecture (Joshi & Nayak 2013). Therefore, we combined the results obtained by using different methods to refine the annotation for size and number of genes included in a cluster because the differences in the size and organisation of the two genomes analysed made the comparison difficult.

The size of *R*-gene clusters can vary significantly, from a few genes to several hundred genes, e.g. Ve (Kawchuk et al. 2001) and Dm3 loci (Meyers et al. 1998). Most isolated resistance genes occurring in clusters evidenced genome rearrangement that could be triggered by plant-pathogen interaction (Spoel & Dong 2012). Recently, it was observed that increased pathogen pressure induces epigenetic changes and promotes PRGs rearrangements (Molinier et al. 2006; Boyko et al. 2007; Alvarez et al. 2010). Indeed, the PRGs clusters play a leading role in new functional resistance gene generation, since they represented a localized island of genetic variability in the plant genome.

Orthology prediction analysis performed by comparing tomato, eggplant and pepper proteomes evidenced more than 1000 pathogen recognition orthologues and related inparalogues conserved across the three cultivated Solanaceae spp. Some cloned tomato *R*-genes showed a putative orthologue in pepper and eggplant, whilst other tomato PRGs were shared just with one species. Merging orthology-related data and cluster data, we confirmed that PRGs comprise one of the most plastic gene families in plants, associated with gene loss, gene conservation and gene clustering (Zhang et al. 2014). Tomato chromosome 4 showed the highest number of conserved PRG clusters even if some clusters were fragmented into different chromosomes in other species (Destefanis et al. 2015). The putative Gpa2 locus in tomato, located on chromosome 12, shares an orthologous relationship with two clusters located on chromosome 9 in pepper. PRG clusters on tomato chromosome 9, including genes Sw5 and Tm2, showed orthologous clustering on pepper chromosome 3 (Djian-Caporalino et al. 2007; Grube et al. 2000), where some genes involved in resistance to viruses are located (Caranta et al. 1997).

Targeted capture sequencing allowed the detection of a big set of variations in genes located in important resistance loci. Polymorphisms in the genomic region containing *R*-genes were pronounced in several species (McHale et al. 2012; González et al. 2013). This technique displayed its usefulness for sequencing large genomic regions, offering a simple method to analyse gene polymorphism in a relatively efficient and economic manner (Grover et al. 2012). The sequenced syntenic regions in eggplant and pepper were larger than those in tomato, possibly due to the difference in genomic size reported for the species analysed (Tomato & Consortium 2012; Hirakawa et al. 2014; Qin et al. 2014). In particular, pepper showed a massive genomic insertion of transposable elements (Kim et al. 2014b). Moreover, a lower number of reads was mapped in eggplant since its genome was not assembled in pseudomolecules. In general, the efficiency of target sequencing was similar to that observed in cross-species microarray experiments (Nazar et al. 2010; Lu et al. 2009; Bar-Or et al. 2007). The highest number of variants was obtained in *S. peruvianum*, confirming the level of polymorphism reported in the literature for this species (Aflitos et al. 2014).

The cataloguing of genes and the concerted use of genomic information (clustering tendency; orthology relationship and variant detection) showed to be a valuable strategy for identifying important genes or alleles and for exchanging information related to coding protein function across plant species. Comparative analysis of two selected PRG loci showed a high level of genome rearrangement (gene losses, duplicated genes, genome shuffling and transposable element insertions). PRG polymorphisms (SNPs, IN/DEL, domain loss or insertions) that can play an important role in gene recombination (Baumgarten et al. 2003; Meyers et al. 2003; McHale et al. 2012; Sanseverino & Ercolano 2012). Even if the mechanisms underlying enhanced recombination at these loci have not been clearly established, pathogen recognition protein structure changes can have a great impact in specific disease response (Zhang et al. 2004; Nandety et al. 2013; Wang et al. 1998). PRG architecture seems to be modified by the interplay of large-scale gene organisation that determines global conservation of locus order genome-wide and extensive local genome rearrangements mediated by tandem duplication, transposons and other shuffling elements that lead to distinct local arrangements (Zhang et al. 2014; Aversano et al. 2015). Regions including genes involved in defence responses have been shown to be hot-spots of genomic variability across genomes (Spoel & Dong 2012).

Extant local arrangements of Solanaceae pathogen recognition genes within a genome are indicative of biological and environmental factors influencing genotype adaptation, and have significant influence on phenotypic resistance diversity (Aversano et al. 2015).

4. INVESTIGATION OF EUROPEAN TOMATO IMPROVEMENT HISTORY THROUGH aDNA SEQUENCING

4.1 Introduction

Genetic analyses of ancient DNA have been used to dissect the genetic basis of traits underlying domestication in a wide range of organisms (Mascher et al. 2016). Current knowledge of plant domestication is largely derived from morphological analysis of archaeological and herbarium remains and/or population genetic analysis of present-day samples. Trace the selection history of a species can provide insights into the selection of important traits, facilitating both the management germplasm repository and the use of genetic resources (Blanca et al. 2015).

The evolutionary history of tomato (*Solanum lycopersicum*) has been clarified comparing genomes of cultivated varieties and wild species (Aflitos et al. 2014; Lin et al. 2014). Tomato domestication probably occurred in the Andean region of Ecuador and Peru and was completed in Mesoamerica (Blanca et al. 2012). Subsequently, a rapid evolution of populations under human selection led to conspicuous phenotypic transformations, as well as adaptations to varied environments (Bai & Lindhout 2007). Extensive breeding activities have modified tomato over the last centuries. Breeding was mainly focused on improving yield production, fruit quality and disease resistance traits. These efforts resulted in the introduction of many introgressions from tomato relatives and more distant wild species (Sim et al. 2011). Selection sweeps promoted the diversification and genetic differentiation in fresh and processing tomato market classes (Lin et al. 2014). The traits that most likely have been selected during the domestication of tomato were fruit morphological traits.

However, many questions about the events occurred during the domestication process remain unanswered. Notably, some changes in fruit shape became in 'modern' cultivars may originated after the tomato was brought to Europe about 500 years ago, albeit is not well understood when and where these alleles arose and how they spread through the germplasm. Multiple evolutionary processes in small cherry fruit, round large fruit, and elongated fruit have been postulated. For example, elongated accessions are evolutionary intermediates between large round and small size accessions (Lin et al. 2014). In recent years, several genes affecting these traits have been identified (Liu et al. 2002; Frary 2000; Xiao et al. 2008). Xiao et al asserted that elongated variants derived by Sun gene duplication (Xiao et al. 2008). However, other authors hypothesized that elongated

tomato fruits originated as hybrids between large round and small size tomato, and based on their distribution, they originate in Europe (Rodriguez et al. 2011). Furthermore, although several hypotheses have been proposed, the exact geographical origin of the elongated groups has not been established (Rodriguez et al. 2011). Small-scale aDNA studies can help to reveal patterns of crops adaptation and migration, however, they can't investigated the impact of these events on whole crop genomes. For this reason, whole genome scale studies on ancient genomes have been conducted in recent years, paving the way for many future studies in this fascinating field of research. Here it is reported the genome sequences of two tomato herbarium samples, which are part of the *Herbarium Porticense* collection (http://www.herbariumporticense.unina.it/it/). Whole-genome sequences of herbarium samples were compared to modern tomato accessions to reveal the relationship with wild and cultivated landraces and to investigate the improvement history of the tomato crop in Italy and in Campania region in the last centuries.

4.2 Materials and methods

Collection of Samples

The samples were taken from the *Herbarium Porticense* collection in MUSA Museum (http://www.centromusa.it/it/), University of Naples Federico II. The older samples were called SET17. According to the label, reporting information related to the identity of the species, the identity of the collector, the oldest herbarium material is 250 years old since it was collect in the eighteenth century in the historical herbaria of Neapolitan botanist Domenico Cirillo (Ricciardi & Castellano 2014a), at the time it was catalogued as "*Solanum (Lycopersicon)*". The second called LEO90 is part of the personal collection of botanist Orazio Comes (Ricciardi & Castellano 2014b), dated in 1890 and catalogued as "*Lycopersicum esculentum var. oblungum*".

aDNA extraction and PCR amplification

Total genomic DNA was isolated from herbarium leaves dated between 1700 and 1890. Approximately 0.005 g of tissue was ground in sterile 1.5 ml tubes using sterilized

micropestles or Tissue Lyser following the Ames et al protocol (Ames & Spooner 2008)or Yoshida et protocol (Yoshida et al. 2013) with slight modification.Moreover. DNeasy Plant Mini Kit (Qiagen, Hilden, Germany) was tested on the same samples. A pair of primers trn V/ndh C fw (F: 5 ′ AAG TTT ACT CAC GGCAAT CG 3 ′ and trn V/ndh C rev (R:5 ′ GGA GGG GTT TTT CTT GGT TG) were used to perform PCR reactions with 10 ng of genomic, 10 pmol primers, 1 U of Taq DNA polymerase Kit (Invitrogen, Carlsbad, CA, USA), 10 pmol dNTPs,and 2 mM MgCl2 in 25 µl reaction volumes. Amplification was performed using the following cycling conditions: 1 min at 94 °C, followed by 30 cycles of 1 min at 94 °C, 1 min 30 s at 60 °C and 2 min at 72 °C, with a final extension for 7 min at 72 °C. All positive controls were manipulated separately from the herbarium samples to avoid contamination using the same master mix Amplicons were separated by electrophoresis on agarose gel (1.5 %), and photographed by a Gel Doc system (Bio-Rad, Milan, Italy). Amplicons were sequenced using the BigDye Terminator Cycle Sequencing Kit (AppliedBiosystems, Foster City, CA, USA) and run on automated DNA sequencers (ABI PRISM 3100 DNASequencer, Applied Biosystems). Sequence data were aligned with corresponding reference sequences using clustalw (www.genome.jp/tools/clustalw).

Preparation of library and sequencing

The genome library for target sequencing was prepared using Illumina Nextera XT DNA sample prep kit (Illumina, San Diego, USA) according to the manufacturer's protocol. The gDNA was fragmented by random transposon integration. DNA adapters with sample-specific barcodes were added to each sample prior to PCR amplification. The library was size-selected using magnetic beads, and the sequencing of samples was conducted on the Illumina HiSeq 2500 (Illumina, San Diego, USA). The sequencing reads were processed in order to remove low quality reads.

Data processing and SNP calling (Analysis of variation)

The sequence data from sequencing were processed using Super-W pipeline (http://www.sequentiabiotech.com/sequentia-research-and-development/projects/) of SEQUENTIA BIOTECH SL (http://www.sequentiabiotech.com/). The pipeline was

divided into three steps: filtering, mapping and variation calling. After the filtering step, all the samples were mapped against a *Solanum lycopersicum* genome v.2.50 (https://solgenomics.net) with BWA (Li & Durbin 2009) using the bwa aln algorithm. The mapped files were filtered for removing PCR duplicates using Picard (http://broadinstitute.github.io/picard/), compressed in bam files, sorted and indexed (Li et al. 2009) creating as output a bam file and a statistical output with all the information about the trimming and the mapping. The variant calling (SNPs) and short deletion and insertion polymorphisms (DIPs) was performed with SAMtools (Li et al. 2009) through a double calling step. The first run of SAMtools was used to perform a multiple pileup (Mpileup), in which the all samples were used together to perform the SNP and DIP calling while a second run is used to call small variations independently for each sample. The final result of the two previous analyses were compared with the variant data (listed at the link: http://www.tomatogenome.net/accessions.html) of 82 tomato wild and cultivated varieties (Aflitos et al. 2014).

Construction of Phylogenetic three and PCA analysis

Data from ancient and cultivated varieties genomes (http://www.tomatogenome.net/accessions.html) were used for build up a phylogenetic three and to perform a PCA analysis. The variant calling results were converted into a binary format, i.e. for each observed SNP a 0 or a 1 was assigned to each genotype for absence or presence, respectively. The table obtained was imported in R and analysed with the package "ape" (Paradis et al. 2004) to produce a neighbour joining tree ("nj" command) and to perform a Principal Component Analysis, with the function "prcomp". The PCA was then plotted with the package "ggplot2" (Wickham 2011).

4.3 Results

Reconstruction of herbarium samples history

Two fragments of tomato plant samples (Figure 10) conserved in *Herbarium Porticense* were collected to perform further DNA analysis.

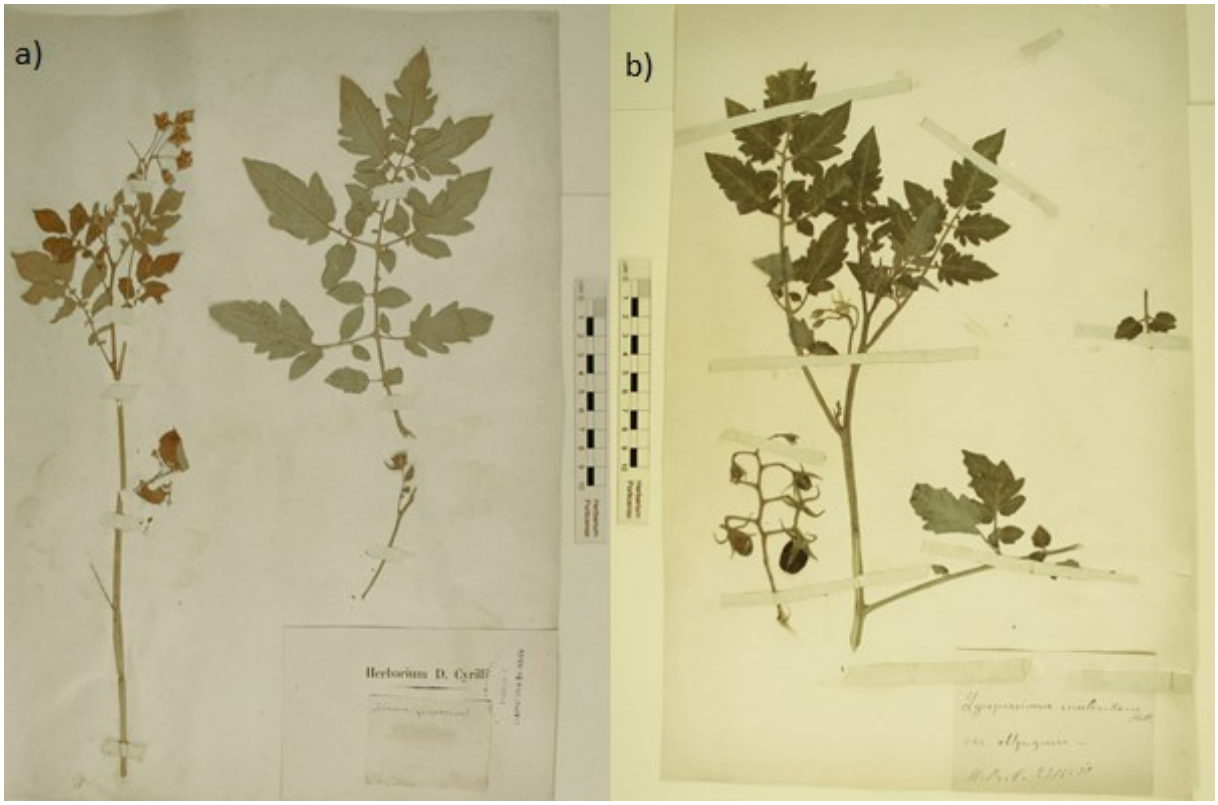


Figure 10. a) Picture of 18th herbarium from Cirillo collection. Scale bar (at right), 1 cm. b) Picture of 19th herbarium from Orazio collection. Scale bar (at left), 1 cm.

A tentative of visual identifications of plant material was conducted by a competent botanist through a careful examination of samples. Morphological diagnostic features, such as portion of shoot, flowers, or fruits as well as annotations made on the cards supported specimens assignment. Furthermore, since herbaria typically contain multiple specimens of the same area and of closely related species, the botanist assessed the extent of natural variability between plants of the same and separate species present in the same collections to support his conclusions.

Comparison of aDNA extraction methods

Three different DNA extraction protocols were tested for extract aDNA form herbarium samples: Table 10 reports the quantity of DNA obtained and the quality parameters.

Extraction Method	Herbarium ID	Collection	Collection date	Mean DNA yield ng/ml	280/260	260/230
Kit Qiagen	SET17	Cirillo	-	23.36	1.8	0.7
	LEO90	Comes	1890	44	1.5	0.41
Ames	SET17	Cirillo	-	174	1.87	0.81
	LEO90	Comes	1890	88.3	1.6	0.69
Yoshida	SET17	Cirillo	-	18.8	1.7	0.82
	LEO90	Comes	1890	1516	1.29	0.58

Table 10. The effects of different DNA extraction protocols on herbaria SET17 and LEO90.

The quantity of DNA extracted ranged from 23.36 ng/ml to 174 ng/ml with a 280/260 ratio ranging from 1.29 to 1.87 and a ratio 260/230 ranging from 0.41 to 0.82. In all cases, the DNA was highly fragmented and gave a smear on agarose gels, revealing mostly fragment sizes below 200 bp (Figure 11). DNA extracts were colourless to brownish, depending on the method used. Since the yield of the extracted aDNA with the three methods was a magnitude of several ng μl^{-1}, the amplification of plastid genes was performed resulting to the detection of PCR products. The aDNA extracted using Ames protocol gave clear DNA-amplicons, that sequenced confirmed the expected sequence. For this reason, aDNA extracted using the last protocol was delivered to a sequencing center.

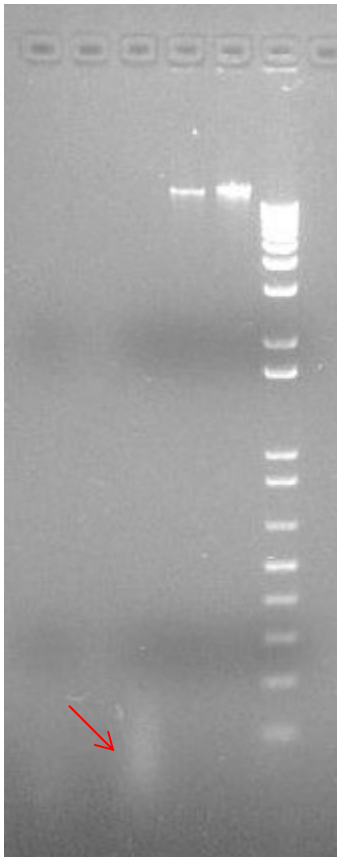


Figure 11. Agarose gel visualization of aDNA extracted with Ames modified method. The red arrow indicates the band of extracted and amplified DNA.

Sequencing and mapping to reference genome

SET17 and LEO90 samples were sequenced following a pair-end sequencing strategy. A total 83,941,779 of aDNA short reads were extracted from SET17 sample whilst 34,300,900 pair-end reads were sequenced from LEO90 with a mean read length of 92.6 for the first sample and 80.5 for the second. Considering that the tomato genome expected size of about 900 Mb (SL2.50), an average coverage of about 8x and 4x genome equivalent was obtained, respectively for SET17 and LEO90. Subsequently, a quality check of raw sequencing data was performed in order to remove adapters and low quality portions, while preserving the longest high quality part of the NGS read (Table 11).

Sample ID	Mapped reads n.	Reads after removing duplicates n.	Reads after removing low-qual reads n.
LEO90	58.184.724 (84.8%)	8.429.533 (12.2%)	3.743.080 (5.4%)
SET17	143.825.677 (85.6%)	9.639.051 (5.7%)	5.005.751 (2.9%)

Table 11. Number and percentage of reads mapped on reference genome obtained after duplicates and low quality reads removal.

More that the 80% of LEO90 (58,184,724) and Set 17 (143,825,677) reads were mapped on the reference genome of *Solanum lycopersicum* (genome assembly SL2.50). However some adjustments were made in order to perform further analysis. In particular PCR duplicates and reads having a mapping quality below 30 were removed. After this effort about 3,743,080 (5.4%) reads for LEO90 and 5,005,751 (2.9%) were available for genome analysis (table 11). The distribution of filtered reads was also assessed and plotted along each chromosome (Figure 12).

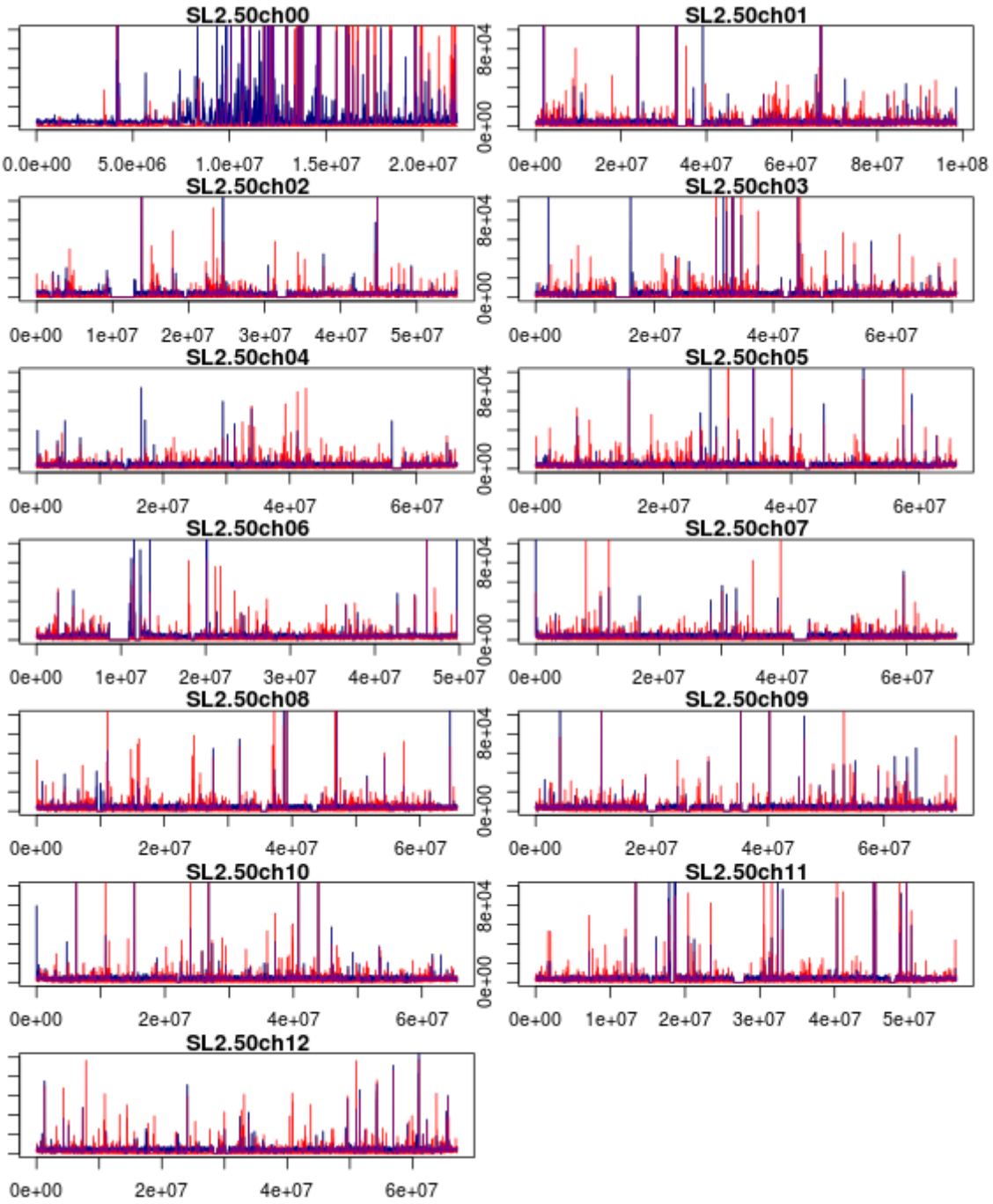


Figure 12. Coverage read distribution along chromosomes subdivided in bins of 10k bp for genotype SET17 (blue) and LEO90 (red). X-coordinate shows chromosome length. Y-coordinate shows coverage (range 0-100000 in all plots).

The histogram, revealed a major coverage peak on chromosome 0. This value is likely let to underestimate the real genome size as suggested by the presence of cytoplasmic DNA. However, the reads albeit with a low coverage were present along all chromosomes.

DNA variant annotation

Overall, the sequencing of SET17 revealed the presence of 274,229 SNPs and 5,993 small Indels. LEO90 showed a number of variants almost 10 times greater than SET with 2,484,966 SNPs/ and 33,063 indels.

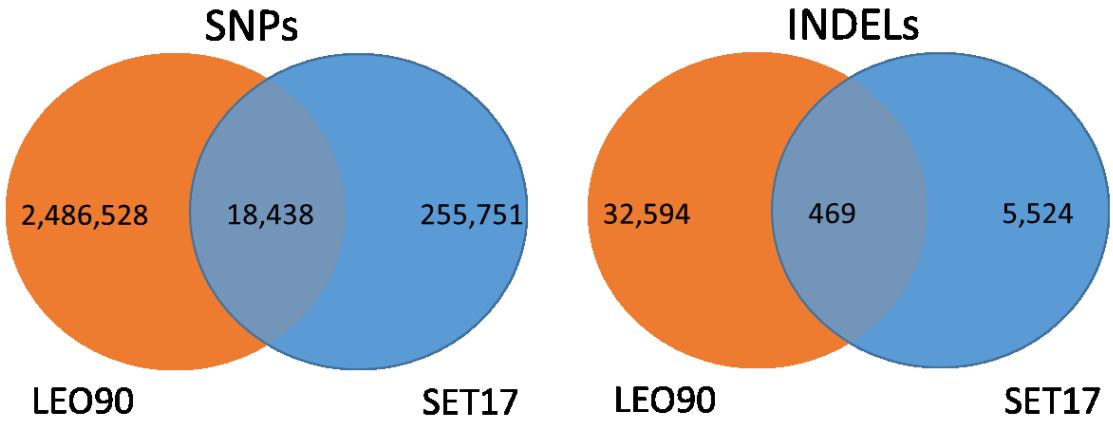


Figure13. Common and private SNPs and INDELs between LEO90 and SET17.

A small number of variants is shared between the genotypes corresponding to 18,438 SNPs and 469 Indels.

The variants were also filtered for genome location in order to identify important regions harbouring genes putatively involved in the tomato domestication and improvement sweeps (Lin et al. 2014). Also in such regions the number of SET17 variants (15.561) was lower than that LEO90 (254.386).

		LEO90	SET17	Common
Domestication sweeps	variants in region	254,386	15,561	938
	variants in genes	167,659	5,585	549
	genes with variants	4,823	2,823	214
	shared variants with landrace genotypes from 82 TGRP*	11,489	4,632	343
Improvement sweeps	variants in region	225,245	10,775	766
	variants in genes	132,797	3,684	478
	genes with variants	3,722	2,127	153
	shared variants with landrace genotypes from 82 TGRP*	9,128	2,636	324

Table12. Number of variants in domestication and improvement sweeps. Note: variants lying within 2.000bp upstream and downstream of genes are considered gene variants . 82 TGRP=genotypes listed at http://www.tomatogenome.net/accessions.html.*

In total, we identified 5,585 variants in 2,823 Set 17 genes and 167,659 variants in 4.823 LEO90 genes involved in domestication sweeps as well as 3.684 variants in 2.127 SET17 genes and 132,797 variants in 3,722 LEO90 genes covering improvement sweeps.

Comparing variants identified with TGRP (Tomato Genome Resequencing Project) genotypes (containing tomato cultivated cultivars and related wild species) was possible highlight variations present in one or more genotypes. Jointly, the domestication and improvement shared gene sweep variants were 9,128 in LEO90 and 2,636 in SET17. Overall, SET17 showed a lower number of variants despite its greater number of mapping reads than LEO90 (Table 12).

Nucleotide changes at target loci

Genes that could influence important agronomic traits, with particular attention to those related to the fruit quality traits were investigated further. More than 100 loci involved in the determination of the morphological traits of tomato (Sacco et al. 2015) were assessed for polymorphism. A high percentage of genes belonging to all investigated classes showed variants. However, the total number of varied genes is not indicative of specificity of variants for LEO90 or SET17 genes. The two genotypes share only 21 variants. On average 290 variations for trait, ranging from 0 to 689, have been identified in LEO90, and 7 ranging from 1 to12 in SET17 (Table 13).

Trait	Loci analysed for trait n.	Common Variants n.	Private SET17 variants n	Private LEO90 variants n.
Fruit color	18	5	12	689
Fruit shape/size	72	16	15	670
Fruit weight	2	0	1	69
Pericarp tichkness	2	0	2	0
Plant architecture	8	0	6	29

tot.	102	21	36	1457

Table 13. Number of variants identified in tomato loci related to morphological traits. Note: variants lying within 2.000bp upstream and downstream of genes are considered variants in genes.

Interestingly, three genes involved in fruit shape determination (Ovate, FAS and LC) varied only in LEO90. In LEO90 was also conspicuous the number (69) of variants in fruit size genes (Fw2.2 and *Fw2.3)*.

Principal component and phylogenetic analysis

To explore the LEO90 and SET17 pedigree history, we compared the genomes with a panel of wild and cultivated tomato previously sequenced genomes (Aflitos et al. 2014; http://www.tomatogenome.net/accessions.html) using a Principal-component analysis (PCA) approach and a neighbour joining tree algorithm. In particular, the pattern of genetic structure within the collection was detected performing a global PCA with the common variants (Figure 14).

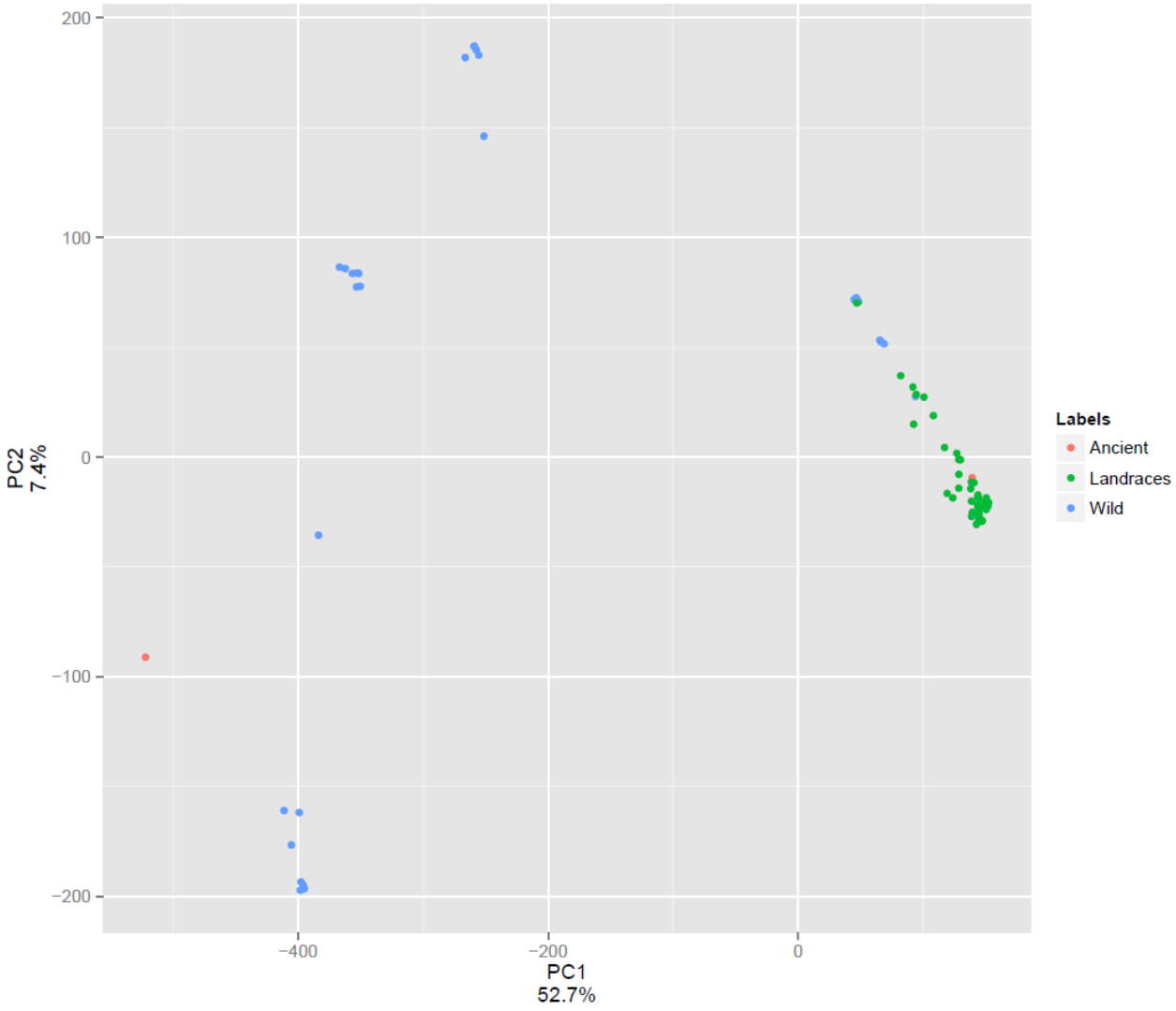


Figure 14. PCA showing ancient samples projected onto the PCA axes with other 82 tomato accessions.

The graphical pattern of the first two principal components (PCs) suggests an arced structure with a clear edge corresponding to cultivated species and a less dense edge corresponding mainly to wild species. Most wild accessions are differentiated only along PC2, forming the right edge (negative PC1, distributed PC2). Additionally, few wild genotypes appeared close to main cultivate group, this wild accessions belong to the species *S. pimpinellifolium*, *S. galapagense* and *S. cheesmaniae*. Principal-component analysis (PCA) confirms the fundamental patterns of population structure across present-day accessions and wild species reported in previous studies. Moreover, the PCA plot revealed that LEO90 is not closely related affiliated with any particular group of

cultivated varieties (Figure 14), rather its genome close to a *S. pennellii* accession (up in the Figure 14) and a group of accessions belonging to *S. habrochaites* species (down). SET17 clustered with the cultivated landraces despite it is the oldest sample. The history of herbarium collections was completely different and tomato samples preserved may came from different geographic areas. By contrast, a phylogenetic analysis on the same dataset of PCA was performed to detect the nearest accession to herbarial samples (Figure 15).

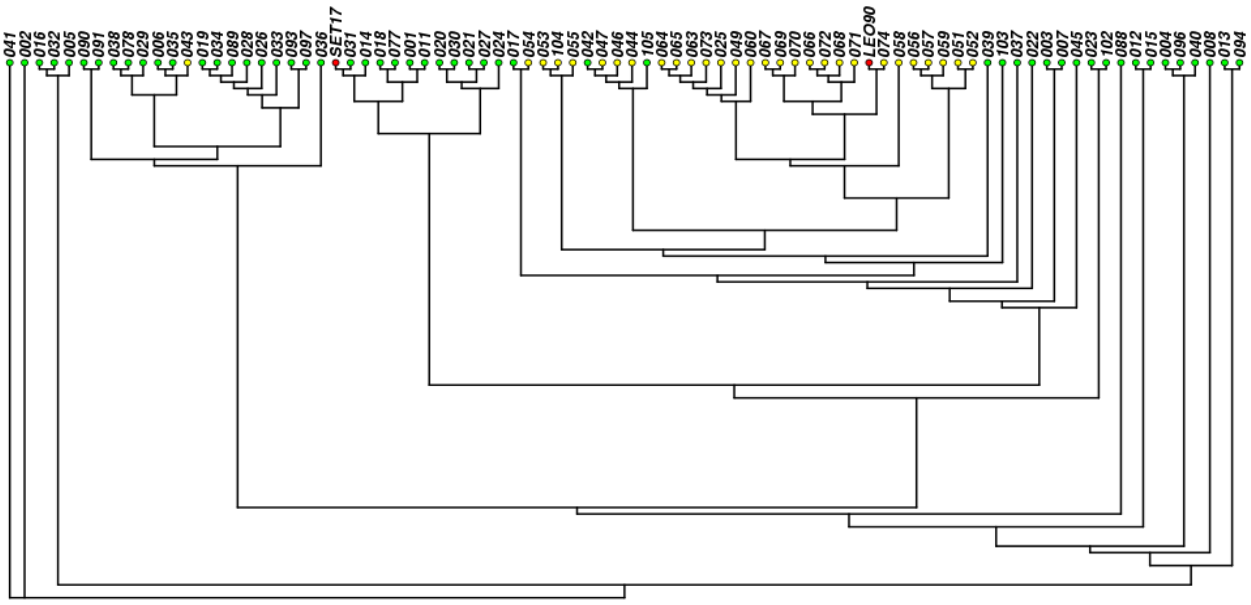


Figure 15. The phylogenetic analysis on the same dataset of PCA.

SET17 is closed related to an accession labelled with number 031, this is yellow fruits variety, collected at beginning of last century in the Caserta area, indicating that more than 100 years of cultivation not modified the basic tomato genome makeup. The most related tomato to LEO90 was a *S. pennellii* accession (074). This accession is a wild tomato species endemic to Andean regions in South America, it was recently sequenced (Bolger et al. 2014) and used in tomato breeding programs.

4.4 Discussions

To better understand the history of tomato spread in Europe, the draft genomes of two *Lycopersicon esculentum* samples belonging to the Herbarium Porticense (PORUN) were investigated. A detailed analysis of Herbarium card, including handwritten labels and notes, allowed to shed light on the significance of the PORUN collection samples within the historical and scientific context. The oldest sequenced sample belongs to Cirillo Collection (XVIII Century), a famous private herbarium of the eighteenth century. Personal studies of Cirillo focused both on wild plants (Ricciardi & Castellano 2014b) or on plants grown in his own garden. The herbarium bearing down an author's autograph card (de Natale & Cellinese 2009), confirms that such sample is one of the oldest tomato preserved herbarium samples collected in Italy. The tomato sample, however, does not possess other indication unless the taxonomic data, even if the leaf traits resulted similar to traits of cultivated traditional varieties (De Natale personal communication). The date is unknown, but certainly attributable to the second half of the eighteenth century, as the activities of Cirillo took place during this period. The second sample, used for the DNA analysis belongs to the Comes Collection (XIX-XX Centuries), but currently is conserved into the General Herbarium Collection (De Natale 2007). The dried herbarium sample realized by Comes is accompanied by an autograph card with annotation of the species (*Lycopersicum esculentum* Mill. Var. *Oblungum*), the location of collection (H.B.P.= Hortus Botanicus Porticense) and the date of collection, June 23, 1890. Such sample showed leave, traces of flowers and elongated fruits supported specimens assignment.

Accordingly the aDNA literature some issues related to DNA sample preservation and DNA isolation, need to be addressed before begin aDNA sequencing project (Rizzi et al. 2012). The aDNA extraction from dried herbarium tomato leave was done using three different methods, namely the Ames method, a protocol based on proteinase K digestion (Yoshida) and the commercial kit QIAGEN DNAesy for DNA extraction. All procedures allowed to obtain DNA from plant herbarium samples. Technical adjustments of DNA isolation protocols, including the increasing the tissue disruption with tissue lyser and the elimination of proteins, improved the quality of aDNA obtained. The aDNA extracted by

Ames protocol, showed the best performance and therefore was sequenced using an Illumina platform.

aDNA sequencing allowed to obtain NGS reads of good quality even if most of them map on chromosome 0. Such chromosome is an artefact that grouped unmapped scaffold and cytoplasmic DNA, supporting the finding that mitochondrial or plastid DNA are more easily retrieved in ancient specimens than nuclear DNA (Rizzi et al. 2012). Moreover, the hydrolytic and oxidative damage degrade aDNA to short fragments no longer than 200 nucleotides (Gugerli et al. 2005) and targeting such short fragments is difficult, compared with those typically employed with contemporary material. Interestingly, a high number of variants in such material were discovered in genes related to agronomic traits and most of them were supported by variants found in modern accessions. Such data resulted useful to address questions related to routes of tomato migration and improvement.

At this end, a suite of ancient samples collected from several repositories has been investigated. In literature the majority of the tomato diversity is explained by the derived alleles of the *Fw2*, *Fw3*, *FAS*, *SUN*, *OVATE* and *LC* genes (Aflitos et al. 2014). Interestingly, the herbarial sample LEO90 classified in 1890, as oblungum variety, showed several private variants in loci implicated on fruit shape determination. By contrast, the detected variants in SET17 are more similar to which found in modern cultivated varieties. Moreover, the PCA analysis conducted on same dataset highlighted an high similarity of SET 17 with a tomato landraces deriving from same area of herbarium collection. The Italians were considered early leaders in the development of new cultivars across 18th and 19th (Stevens & Rick 1986) and such finding suggests that important agronomic traits were already improved in the eighteenth century. A high selection pressure in fruit morphological traits occurred during the domestication and improvement of tomato (Lin et al. 2014). Accession 031 (http://www.tomatogenome.net/accessions.html) harvested in Campania region at begin of last century showed also a yellow fruits supporting the hypothesis that it derives for "pomi d'oro" (golden apple), which cultivation is documented in Europe in 18th (Stevens & Rick 1986). Instead, LEO90 was more close to a *S. pennellii* accession and to a group of a *S. habrochaites* accessions that belongs the "Hirsutum" group (Pease et al. 2016)

suggesting that the tomato elongated fruit was originate from a cross between a local landrace and a wild ancestor.

5. CONCLUSIONS AND PERSPECTIVES

In this work, the tomato diversification at genomic level of agronomical traits was investigated. For this purpose, it was necessary to acquire knowledge on genes and genomic regions involved in determination of specific agronomical traits, collecting a large amount of data from numerous databases and extracting genomic information from different kind of samples. The complexity of agronomical traits was examined using multi-omic approaches, ranging from the latest-generation sequencing technology to modern *in silico* methods. The genomic investigations were useful to clarify the history of diversification of agronomical traits, and, at the same time, to pave the way for the implementation of a more sustainable agriculture.

In detail, the three main aims pursued in this thesis allow to perform:

- **A reconstruction of evolutionary history of NLR-like gene family in Metaphyta kingdom**. More than 34,000 NLR-like genes from 102 organism were identified and classified. A great diversification was observed in NLR genes highlighting specific dynamic in each botanical taxa. A tendency to duplicate and cluster only a specific gene member in species belonging to the same family or taxa was envisage in orthologue gene groups. Such finding suggests that a basic R gene structure is selected duplicated and diversified in taxa/species in order to trigger the best plant immunity response. The expansions/contraction across the entire land plant lineages and the specific recombination events provide the most appropriate arsenal for each plant species.

- **A comparison of Solanaceae orthologous pathogen recognition gene-rich regions**. A complete catalogue of *Solanum melongena* and *Capsicum annum* nucleotide-binding site (NBS), receptor-like protein (RLP) and receptor-like kinase (RLK) gene repertories was generated. The results confirm that most Solanaceae PRGs tend to be physically clustered and that clustering play a leading role in new functional resistance gene generation, since it represents a localized island of genetic variability in the plant genome. Targeted capture sequencing allowed the detection of allelic variation in important resistance loci and comparative analysis of PRG loci showed a high level of genome rearrangement.

- **An investigation of European tomato improvement history through aDNA sequencing**. The sequencing of two ancient genomes sourced from a museum Herbaria collection allowed to access to a vast archive of genome data. A proper aDNA extraction method was set up to conduct further sequencing. Detailed investigations on ancient, wild and cultivated genome varieties permitted to discovery site of selective pressure. A high level of improvement in agronomic traits in the eighteenth century was discovered and, ultimately, a cross event that originate to elongated shape fruit varieties, cultivated in the begin of last century in Campania region, was hypothesized.

The results confirm the importance of omics approaches both to study and improve agronomic important traits.

6. REFERENCES

Aflitos, S. et al., 2014. Exploring genetic variation in the tomato (Solanum section Lycopersicon) clade by whole-genome sequencing. *The Plant journal : for cell and molecular biology*, (January), pp.136–148. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25039268 [Accessed August 28, 2014].

Alexeyenko, A. et al., 2006. Automatic clustering of orthologs and inparalogs shared by multiple proteomes. In *Bioinformatics*.

Alvarez, M.E., Nota, F. & Cambiagno, D. a., 2010. Epigenetic control of plant immunity. *Molecular Plant Pathology*, 11(4), pp.563–576.

Ames, M. & Spooner, D.M., 2008. DNA from herbarium specimens settles a controversy about origins of the European potato. *American Journal of Botany*, 95(2), pp.252–257. Available at: http://www.amjbot.org/cgi/doi/10.3732/ajb.95.2.252.

Anderson, S. et al., 1981. Sequence and organization of the human mitochondrial genome. *Nature*, 290(5806), pp.457–465. Available at: http://www.ncbi.nlm.nih.gov/pubmed/7219534%5Cnhttp://www.nature.com.libproxy.ucl.ac.uk/nature/journal/v290/n5806/pdf/290457a0.pdf.

Andolfo, G. et al., 2014. Defining the full tomato NB-LRR resistance gene repertoire using genomic and cDNA RenSeq. *BMC plant biology*, 14(1), p.120. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4036795&tool=pmcentrez&rendertype=abstract [Accessed July 11, 2014].

Andolfo, G. et al., 2013. Genome-wide identification and analysis of candidate genes for disease resistance in tomato. *Molecular Breeding*, 33(1), pp.227–233. Available at: http://link.springer.com/10.1007/s11032-013-9928-7 [Accessed July 11, 2014].

Andolfo, G. et al., 2013. Overview of tomato (Solanum lycopersicum) candidate pathogen recognition genes reveals important Solanum R locus dynamics. *The New phytologist*, 197(1), pp.223–37. Available at:

http://www.ncbi.nlm.nih.gov/pubmed/23163550.

Andolfo, G. & Ercolano, M.R., 2015. Plant Innate Immunity Multicomponent Model. *Frontiers in Plant Science*, 6(November), p.987. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4643146&tool=pmcentrez&rendertype=abstract.

Aversano, R. et al., 2015. *The* Solanum commersonii *Genome Sequence Provides Insights into Adaptation to Stress Conditions and Genome Evolution of Wild Potato Relatives*, Available at: http://www.plantcell.org/lookup/doi/10.1105/tpc.114.135954.

Bai, Y. & Lindhout, P., 2007. Domestication and Breeding of Tomatoes: What have We Gained and What Can We Gain in the Future? *Annals of Botany*, 100(5), pp.1085–1094. Available at: https://academic.oup.com/aob/article-lookup/doi/10.1093/aob/mcm150.

Bar-Or, C., Czosnek, H. & Koltai, H., 2007. Cross-species microarray hybridizations: a developing tool for studying species diversity. *Trends in Genetics*, 23(4), pp.200–207.

Bateman, A. et al., 2002. The Pfam Protein Families Database. , 30(1), pp.276–280.

Baumgarten, A. et al., 2003. Genome-level evolution of resistance genes in Arabidopsis thaliana. *Genetics*, 165(1), pp.309–319.

Bevan, M.W. & Uauy, C., 2013. Genomics reveals new landscapes for crop improvement. *Genome biology*, 14, p.206. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3706852&tool=pmcentrez&rendertype=abstract.

Blanca, J. et al., 2015. Genomic variation in tomato, from wild ancestors to contemporary breeding accessions. *BMC genomics*, 16(1), p.257. Available at: http://www.biomedcentral.com/1471-2164/16/257%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4404671&tool=pmcentrez&rendertype=abstract.

Blanca, J. et al., 2012. Variation Revealed by SNP Genotyping and Morphology Provides Insight into the Origin of the Tomato W. Yan, ed. *PLoS ONE*, 7(10),

p.e48198. Available at: http://dx.plos.org/10.1371/journal.pone.0048198.

Bolger, A. et al., 2014. The genome of the stress-tolerant wild tomato species Solanum pennellii. *Nature Genetics*, 46(9), pp.1034–1038. Available at: http://www.nature.com/doifinder/10.1038/ng.3046 [Accessed July 28, 2014].

Boyko, A. et al., 2007. Transgenerational changes in the genome stability and methylation in pathogen-infected plants: (Virus-induced plant genome instability). *Nucleic Acids Research*, 35(5), pp.1714–1725.

Camacho, C. et al., 2009. BLAST plus: architecture and applications. *BMC Bioinformatics*, 10(421), p.1.

Caranta, C., Lefebvre, V. & Palloix, A., 1997. Polygenic Resistance of Pepper to Potyviruses Consists of a Combination of Isolate-Specific and Broad-Spectrum Quantitative Trait Loci. *Molecular Plant-Microbe Interactions*, 10(7), pp.872–878.

Chakrabarti, M. et al., 2013. A cytochrome P450 regulates a domestication trait in cultivated tomato. *Proceedings of the National Academy of Sciences of the United States of America*, 110(42), pp.17125–17130. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3801035&tool=pmcentrez&rendertype=abstract.

Chase, M.W. et al., 2016. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Botanical Journal of the Linnean Society*, 181(1), pp.1–20.

Cingolani, P. et al., 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w 1118; iso-2; iso-3. *Fly*, 6(2), pp.80–92.

Clark, M.J. et al., 2011. Performance comparison of exome DNA sequencing technologies. *Nature Biotechnology*, 29(10), pp.908–914. Available at: http://dx.doi.org/10.1038/nbt.1975.

Collier, S.M., Hamel, L.-P. & Moffett, P., 2011. Cell death mediated by the N-terminal domains of a unique and highly conserved class of NB-LRR protein. *Molecular plant-microbe interactions*, 24(8), pp.918–931.

Coppe, A., Danieli, G.A. & Bortoluzzi, S., 2006. REEF: searching REgionally Enriched Features in genomes. *BMC bioinformatics*, 7(1), p.453. Available at: http://www.biomedcentral.com/1471-2105/7/453 [Accessed July 16, 2014].

Dangl, J.L. & Jones, J.D.G., 2001. Plant pathogens and integrated defence responses to infection. *Nature*, 411(6839), pp.826–833. Available at: http://www.nature.com/doifinder/10.1038/35081161.

Destefanis, M. et al., 2015. A disease resistance locus on potato and tomato chromosome 4 exhibits a conserved multipartite structure displaying different rates of evolution in different lineages. *BMC plant biology*, 15, p.255. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4619397&tool=pmcentrez&rendertype=abstract.

Djian-Caporalino, C. et al., 2007. Root-knot nematode (Meloidogyne spp.) Me resistance genes in pepper (Capsicum annuum L.) are clustered on the P9 chromosome. *Theoretical and Applied Genetics*, 114(3), pp.473–486.

Dunin-Horkawicz, S., Kopec, K.O. & Lupas, A.N., 2014. Prokaryotic ancestry of eukaryotic protein networks mediating innate immunity and apoptosis. *Journal of Molecular Biology*, 426(7), pp.1568–1582.

Ercolano, M.R. et al., 2012. Genetic and genomic approaches for R-gene mediated disease resistance in tomato: Retrospects and prospects. *Plant Cell Reports*, 31, pp.973–985.

Fernandez-Pozo, N. et al., 2015. The Sol Genomics Network (SGN)-from genotype to phenotype to breeding. *Nucleic Acids Research*, 43(D1), pp.D1036–D1041.

Finn, R.D., Clements, J. & Eddy, S.R., 2011. HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Research*, 39(SUPPL. 2).

Frary, A., 2000. fw2.2: A Quantitative Trait Locus Key to the Evolution of Tomato Fruit Size. *Science*, 289(5476), pp.85–88. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10884229.

Gabriëls, S.H.E.J. et al., 2007. An NB-LRR protein required for HR signalling mediated by both extra- and intracellular resistance proteins. *Plant Journal*, 50(1), pp.14–28.

Garcia-Mas, J. et al., 2012. The genome of melon (Cucumis melo L.). *Proceedings of the National Academy of Sciences*, 109(29), pp.11872–7. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3406823&tool=pmcentrez&rendertype=abstract.

Gasc, C., Peyretaillade, E. & Peyret, P., 2016. Sequence capture by hybridization to explore modern and ancient genomic diversity in model and nonmodel organisms. *Nucleic acids research*, (8), p.gkw309-. Available at: http://nar.oxfordjournals.org/content/early/2016/04/21/nar.gkw309.full.

González, V.M. et al., 2013. High presence/absence gene variability in defense-related gene clusters of Cucumis melo. *BMC genomics*, 14, p.782. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3845527&tool=pmcentrez&rendertype=abstract.

Graham, E.A.M., 2007. DNA reviews: Ancient DNA. *Forensic Science, Medicine, and Pathology*, 3(3), pp.221–225.

Grover, C.E., Salmon, A. & Wendel, J.F., 2012. Targeted sequence capture as a powerful tool for evolutionary analysis. *American Journal of Botany*, 99(2), pp.312–319.

Grube, R.C., Radwanski, E.R. & Jahn, M., 2000. Comparative genetics of disease resistance within the solanaceae. *Genetics*, 155(2), pp.873–887.

Gugerli, F., Parducci, L. & Petit, R.J., 2005. Ancient plant DNA: Review and prospects. *New Phytologist*, 166(2), pp.409–418.

Hallauer, A.R., 2011. Evolution of plant breeding. *Crop Breeding and Applied Biotechnology*, 11(3), pp.197–206. Available at: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1984-70332011000300001&lng=en&nrm=iso&tlng=en.

Hammer, G. et al., 2006. Models for navigating biological complexity in breeding improved crop plants. *Trends in Plant Science*, 11(12), pp.587–593.

Hardison, R.C., 2003. Comparative genomics. *PLoS Biology*, 1(2).

Hayashi, N. et al., 2010. Durable panicle blast-resistance gene Pb1 encodes an atypical

CC-NBS-LRR protein and was generated by acquiring a promoter through local genome duplication. *Plant Journal*, 64(3), pp.498–510.

Higuchi, R. et al., 1984. DNA sequences from the quagga, an extinct member of the horse family. *Nature*, 312(5991), pp.282–284.

Hirakawa, H. et al., 2014. Draft Genome Sequence of Eggplant (Solanum melongena L.): the Representative Solanum Species Indigenous to the Old World. *DNA research : an international journal for rapid publication of reports on genes and genomes*, pp.1–12. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25233906.

Hoberman, R. & Durand, D., 2005. The Incompatible Desiderata of Gene Cluster Properties. *Proc. of RECOMB 2005 International Workshop on Comparative Genomics, RCG 2005*, 3678, pp.73–87.

Hoberman, R., Sankoff, D. & Durand, D., 2005. The statistical significance of max-gap clusters. *Comparative Genomics*, pp.55–71. Available at: http://www.springerlink.com/index/VQQN6C6ECN9T0NM0.pdf.

Holub, E.B., 2001. The arms race is ancient history in Arabidopsis, the wildflower. *Nature reviews. Genetics*, 2(7), pp.516–527.

Jacob, F., Vernaldi, S. & Maekawa, T., 2013. Evolution and conservation of plant NLR functions. *Frontiers in Immunology*, 4(SEP).

Jia, Y. et al., 2015. Extreme expansion of NBS-encoding genes in Rosaceae. *BMC genetics*, 16, p.48. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4417205&tool=pmcentrez&rendertype=abstract.

Jones, P. et al., 2014. InterProScan 5: Genome-scale protein function classification. *Bioinformatics*, 30(9), pp.1236–1240.

Joshi, R.K. & Nayak, S., 2013. Perspectives of genomic diversification and molecular recombination towards R-gene evolution in plants. *Physiology and Molecular Biology of Plants*, 19(1), pp.1–9.

Jupe, F. et al., 2012. Identification and localisation of the NB-LRR gene family within the potato genome. *BMC genomics*, 13, p.75. Available at:

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3297505&tool=pmcentrez&rendertype=abstract.

Katoh, K. & Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, 30(4), pp.772–780.

Kawchuk, L.M. et al., 2001. Tomato Ve disease resistance genes encode cell surface-like receptors. *Proceedings of the National Academy of Sciences of the United States of America*, 98(11), pp.6511–5. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=33499&tool=pmcentrez&rendertype=abstract.

Kearse, M. et al., 2012. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28(12), pp.1647–1649.

Kiialainen, A. et al., 2011. Performance of microarray and liquid based capture methods for target enrichment for massively parallel sequencing and SNP discovery. *PloS one*, 6(2), p.e16486. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3036585&tool=pmcentrez&rendertype=abstract [Accessed November 24, 2014].

Kim, S. et al., 2014a. Genome sequence of the hot pepper provides insights into the evolution of pungency in Capsicum species. *Nature genetics*, 46(3), pp.270–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24441736 [Accessed July 21, 2014].

Kim, S. et al., 2014b. Genome sequence of the hot pepper provides insights into the evolution of pungency in Capsicum species. *Nature genetics*, 46(3), pp.270–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24441736.

Krzywinski, M. et al., 2009. Circos: An information aesthetic for comparative genomics. *Genome Research*, 19(9), pp.1639–1645.

Lees, J.G. et al., 2016. Functional innovation from changes in protein domains and their combinations. *Current Opinion in Structural Biology*, 38, pp.44–52.

Li, H. et al., 2009. The Sequence Alignment/Map format and SAMtools.

Bioinformatics, 25(16), pp.2078–2079.

Li, H. & Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14), pp.1754–1760.

Li, J. et al., 2010. Unique evolutionary pattern of numbers of gramineous NBS-LRR genes. *Molecular Genetics and Genomics*, 283(5), pp.427–438.

Lin, T. et al., 2014. Genomic analyses provide insights into the history of tomato breeding. *Nature genetics*, (October). Available at: http://www.ncbi.nlm.nih.gov/pubmed/25305757 [Accessed October 13, 2014].

Liu, J. et al., 2002. A new class of regulatory genes underlying the cause of pear-shaped tomato fruit. *Proceedings of the National Academy of Sciences*, 99(20), pp.13302–13306. Available at: http://www.pnas.org/content/99/20/13302.short.

Lu, Y., Huggins, P. & Bar-Joseph, Z., 2009. Cross species analysis of microarray expression data. *Bioinformatics*, 25(12), pp.1476–1483. Available at: http://bioinformatics.oxfordjournals.org/content/25/12/1476.full.

Luo, S. et al., 2012. Dynamic Nucleotide-Binding Site and Leucine-Rich Repeat-Encoding Genes in the Grass Family. *Plant Physiology*, 159(1), pp.197–210.

Marone, D. et al., 2013. Plant Nucleotide Binding Site–Leucine-Rich Repeat (NBS-LRR) Genes: Active Guardians in Host Defense Responses. *International Journal of Molecular Sciences*, 14(4), pp.7302–7326. Available at: http://www.mdpi.com/1422-0067/14/4/7302/.

Mascher, M. et al., 2016. Genomic analysis of 6,000-year-old cultivated grain illuminates the domestication history of barley. *Nature Genetics*, (July). Available at: http://www.nature.com/doifinder/10.1038/ng.3611.

Maxam, a M. & Gilbert, W., 1977. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 74(2), pp.560–4. Available at: http://www.ncbi.nlm.nih.gov/pubmed/265521%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC392330.

McHale, L. et al., 2006. Plant NBS-LRR proteins: adaptable guards. *Genome biology*, 7,

p.212.

McHale, L.K. et al., 2012. Structural variants in the soybean genome localize to clusters of biotic stress-response genes. *Plant physiology*, 159(4), pp.1295–308. Available at: http://www.plantphysiol.org/content/159/4/1295.

Meyers, B.C. et al., 2003. Genome-wide analysis of NBS-LRR – encoding genes in Arabidopsis. *The Plant Cell*, 15(April), pp.809–834. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=152331&tool=pmcentrez&rendertype=abstract.

Meyers, B.C. et al., 1998. The major resistance gene cluster in lettuce is highly duplicated and spans several megabases. *The Plant cell*, 10(11), pp.1817–1832.

Moghe, G.D. et al., 2014. Consequences of Whole-Genome Triplication as Revealed by Comparative Genomic Analyses of the Wild Radish Raphanus raphanistrum and Three Other Brassicaceae Species. *The Plant cell*, 26(5), pp.1925–1937. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4079359&tool=pmcentrez&rendertype=abstract [Accessed July 10, 2014].

Molinier, J. et al., 2006. Transgeneration memory of stress in plants. *Nature*, 442(7106), pp.1046–1049.

Nandety, R.S. et al., 2013. The role of TIR-NBS and TIR-X proteins in plant basal defense responses. *Plant physiology*, 162(3), pp.1459–72. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3707564&tool=pmcentrez&rendertype=abstract.

De Natale, A., 2007. Herbarium Porticense. Available at: http://www.herbariumporticense.unina.it/doc/pdf/Erbario/Herbarium_Porticense.pdf.

de Natale, A. & Cellinese, N., 2009. Imperato, Cirillo, and a series of unfortunate events: A novel approach to assess the unknown provenance of historical herbarium specimens. *Taxon*, 58(3), pp.963–970.

Nazar, R.N. et al., 2010. DNA chip analysis in diverse organisms with unsequenced genomes. *Molecular Biotechnology*, 44(1), pp.8–13.

Neves, L.G. et al., 2013. Whole-exome targeted sequencing of the uncharacterized pine genome. *The Plant journal : for cell and molecular biology*, 75(1), pp.146–56. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23551702 [Accessed November 14, 2014].

Olsen, J.L. et al., 2016. The genome of the seagrass Zostera marina reveals angiosperm adaptation to the sea. *Nature*, 530(7590), pp.331–335. Available at: http://dx.doi.org/10.1038/nature16548%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/26814964%5Cnhttp://www.nature.com/doifinder/10.1038/nature16548.

Paradis, E., Claude, J. & Strimmer, K., 2004. APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20(2), pp.289–290.

Peart, J.R. et al., 2005. NRG1, a CC-NB-LRR protein, together with N, a TIR-NB-LRR protein, mediates resistance against tobacco mosaic virus. *Current Biology*, 15(10), pp.968–973.

Pease, J.B. et al., 2016. Phylogenomics Reveals Three Sources of Adaptive Variation during a Rapid Radiation. *PLoS Biology*, 14(2), pp.1–24. Available at: http://dx.doi.org/10.1371/journal.pbio.1002379.

Peele, H.M. et al., 2014. Loss and retention of resistance genes in five species of the Brassicaceae family. *BMC plant biology*, 14(Cc), p.298. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4232680&tool=pmcentrez&rendertype=abstract.

Peng, J. et al., 1999. “Green revolution” genes encode mutant gibberellin response modulators. *Nature*, 400(July), pp.256–261.

Prohens, J., 2011. Plant Breeding: A Success Story to be Continued Thanks to the Advances in Genomics. *Frontiers in Plant Science*, 2(September), pp.1–3.

Qin, C. et al., 2014. Whole-genome sequencing of cultivated and wild peppers provides insights into Capsicum domestication and specialization. *Proceedings of the National Academy of Sciences of the United States of America*, 111(14), pp.5135–40. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3986200&tool=pmcentrez&rendertype=abstract.

Quinlan, A.R. & Hall, I.M., 2010. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6), pp.841–842.

Ramos-Madrigal, J. et al., 2016. Genome Sequence of a 5,310-Year-Old Maize Cob Provides Insights into the Early Stages of Maize Domestication. *Current Biology*, pp.1–7. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0960982216311204.

Rehrig, W.Z. et al., 2014. *CaDMR1* Cosegregates with QTL *Pc5.1* for Resistance to in Pepper (*Capsicum annuum*). *The Plant Genome*, 7(2), pp.1–12. Available at: https://www.crops.org/publications/tpg/abstracts/7/2/plantgenome2014.03.0011.

Remm, M., Storm, C.E. & Sonnhammer, E.L., 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *Journal of molecular biology*, 314(5), pp.1041–1052.

Ricciardi, M. & Castellano, M.L., 2014a. Domenico Cirillo's Collections. *Nuncius*, 29(2), pp.499–530. Available at: http://booksandjournals.brillonline.com/content/journals/10.1163/18253911-02902008.

Ricciardi, M. & Castellano, M.L., 2014b. Domenico Cirillo's Collections. *Nuncius*, 29(2), pp.499–530.

Richly, E., Kurth, J. & Leister, D., 2002. Mode of amplification and reorganization of resistance genes during recent Arabidopsis thaliana evolution. *Molecular biology and evolution*, 19(1), pp.76–84.

Rizzi, E. et al., 2012. Ancient DNA studies: new perspectives on old samples. *Genetics, selection, evolution : GSE*, 44(1), p.21. Available at: http://www.gsejournal.org/content/44/1/21.

Rodriguez, G.R. et al., 2011. Distribution of SUN, OVATE, LC, and FAS in the Tomato Germplasm and the Relationship to Fruit Shape Diversity. *PLANT PHYSIOLOGY*, 156(1), pp.275–285. Available at: http://www.plantphysiol.org/cgi/doi/10.1104/pp.110.167577.

Sacco, A. et al., 2015. Exploring a tomato landraces collection for fruit-related traits by the aid of a high-throughput genomic platform. *PLoS ONE*, 10(9), pp.1–20.

Sanseverino, W. et al., 2009. PRGdb: A bioinformatics platform for plant resistance gene analysis. *Nucleic Acids Research*, 38(November 2009), pp.814–821.

Sanseverino, W. & Ercolano, M.R., 2012. In silico approach to predict candidate R proteins and to define their domain architecture. *BMC Research Notes*, 5(1), p.678. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3532234&tool=pmcentrez&rendertype=abstract.

Der Sarkissian, C. et al., 2015. Ancient genomics. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 370(1660), p.20130387. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4275894&tool=pmcentrez&rendertype=abstract.

Sarris, P.F. et al., 2016. Comparative analysis of plant immune receptor architectures uncovers host proteins likely targeted by pathogens. *BMC Biology*, 14(1), p.8. Available at: http://www.biomedcentral.com/1741-7007/14/8.

Schultz, J. et al., 1998. SMART, a simple modular architecture research tool: identification of signaling domains. *Proceedings of the National Academy of Sciences of the United States of America*, 95(11), pp.5857–64. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=34487&tool=pmcentrez&rendertype=abstract.

Sekhwal, M. et al., 2015. Disease Resistance Gene Analogs (RGAs) in Plants. *International Journal of Molecular Sciences*, 16(8), pp.19248–19290. Available at: http://www.mdpi.com/1422-0067/16/8/19248/.

Seo, E. et al., 2016. Genome-wide Comparative Analyses Reveal the Dynamic Evolution of Nucleotide-Binding Leucine-Rich Repeat Gene Family among Solanaceae Plants. *Frontiers in Plant Science*, 7(August), p.1205.

Shao, Z.-Q., Xue, J.-Y., et al., 2016. Large-Scale Analyses of Angiosperm Nucleotide-Binding Site-Leucine-Rich Repeat Genes Reveal Three Anciently Diverged Classes with Distinct Evolutionary Patterns. *Plant Physiology*, 170(4), pp.2095–2109. Available at: http://www.plantphysiol.org/lookup/doi/10.1104/pp.15.01487.

Shao, Z.-Q. et al., 2014. Long-Term Evolution of Nucleotide-Binding Site-Leucine-Rich Repeat (NBS-LRR) Genes: Understandings Gained From and Beyond the Legume Family. *Plant physiology*, 166(September), pp.217–234. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25052854.

Shao, Z.-Q., Wang, B. & Chen, J.-Q., 2016. Tracking ancestral lineages and recent expansions of NBS-LRR genes in angiosperms. *Plant Signaling & Behavior*, 2324(June), pp.00–00. Available at: http://www.tandfonline.com/doi/full/10.1080/15592324.2016.1197470.

Sim, S.-C. et al., 2011. Population structure and genetic differentiation associated with breeding history and selection in tomato (Solanum lycopersicum L.). *Heredity*, 106(6), pp.927–935. Available at: http://dx.doi.org/10.1038/hdy.2010.139.

Spoel, S.H. & Dong, X., 2012. How do plants achieve immunity? Defence without specialized immune cells. *Nature Reviews Immunology*, 12(2), pp.89–100. Available at: http://www.nature.com/doifinder/10.1038/nri3141.

Steuernagel, B. et al., 2016. Rapid cloning of disease-resistance genes in plants using mutagenesis and sequence capture. *Nature Biotechnology*, (August 2015). Available at: http://www.nature.com/doifinder/10.1038/nbt.3543.

Stevens, M.A. & Rick, C.M., 1986. Genetics and breeding. In *The tomato crop*. Springer, pp. 35–109.

Takeda, S. & Matsuoka, M., 2008. Genetic approaches to crop improvement: responding to environmental and population changes. *Nature reviews. Genetics*, 9(6), pp.444–57. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18475268.

Tamura, K. et al., 2013. MEGA6: Molecular evolutionary genetics analysis version 6.0. *Molecular Biology and Evolution*, 30(12), pp.2725–2729.

Tarr, D.E.K. & Alexander, H.M., 2009. TIR-NBS-LRR genes are rare in monocots: evidence from diverse monocot orders. *BMC research notes*, 2, p.197.

Tian, D. et al., 2003. Fitness costs of R-gene-mediated resistance in Arabidopsis thaliana. *Nature*, 423(6935), pp.74–7. Available at: http://dx.doi.org/10.1038/nature01588.

Tilman, D. et al., 2011. Global food demand and the sustainable intensification of agriculture. *Proceedings of the National Academy of Sciences of the United States of Americaof the United States of America*, 108(50), pp.20260–20264. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3250154&tool=pmcentrez&rendertype=abstract.

Tomato, T. & Consortium, G., 2012. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, 485(7400), pp.635–41. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3378239&tool=pmcentrez&rendertype=abstract.

Vossen, J.H., Jo, K.R. & Vosman, B., 2014. Mining the genus solanum for increasing disease resistance. In *Genomics of Plant Genetic Resources: Volume 2. Crop Productivity, Food Security and Nutritional Quality*. Springer, pp. 27–46.

Wang, G.-L. et al., 1998. Xa21D Encodes a Receptor-Like Molecule with a Leucine-Rich Repeat Domain That Determines Race-Specific Recognition and Is Subject to Adaptive Evolution. *The Plant Cell*, 10(5), p.765. Available at: http://www.jstor.org/stable/10.2307/3870663?origin=crossref.

Wang, Y. et al., 2008. Sequencing and comparative analysis of a conserved syntenic segment in the solanaceae. *Genetics*, 180(1), pp.391–408.

Wei, C., Chen, J. & Kuang, H., 2016. Dramatic Number Variation of R Genes in Solanaceae Species Accounted for by a Few R Gene Subfamilies. *Plos One*, 11(2), p.e0148708. Available at: http://dx.plos.org/10.1371/journal.pone.0148708.

Whitham, S. et al., 1994. The product of the tobacco mosaic virus resistance gene N: similarity to toll and the interleukin-1 receptor [published erratum appears in Cell 1995 May 5;81(3):466]. *Cell*, 78(Department of Plant Pathology, University of California, Berkeley 94720.), pp.1101–1115.

Wickham, H., 2011. ggplot2. *Wiley Interdisciplinary Reviews: Computational Statistics*, 3(2), pp.180–185. Available at: http://doi.wiley.com/10.1002/wics.147.

Witek, K. et al., 2016. Accelerated cloning of a potato late blight–resistance gene using RenSeq and SMRT sequencing. *Nature Biotechnology*, (August 2015), pp.1–8.

Available at: http://www.nature.com/doifinder/10.1038/nbt.3540.

Wu, C.-H. et al., 2016. The NLR helper protein NRC3 but not NRC1 is required for Pto-mediated cell death in Nicotiana benthamiana. *New Phytologist*, 209, pp.1344–1352.

Xiao, H. et al., 2008. A Retrotransposon-Mediated Gene Duplication Underlies Morphological Variation of Tomato Fruit. *Science*, 319(2008), pp.1527–1530.

Xu, X. et al., 2011. Genome sequence and analysis of the tuber crop potato. *Nature*, 475(7355), pp.189–95. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21743474%5Cnhttp://dx.doi.org/10.1038/nature10158.

Yeaman, S., 2013. Genomic rearrangements and the evolution of clusters of locally adaptive loci. *Proceedings of the National Academy of Sciences of the United States of America*, 110(19), pp.E1743-51. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3651494&tool=pmcentrez&rendertype=abstract.

Yoshida, K. et al., 2013. The rise and fall of the Phytophthora infestans lineage that triggered the Irish potato famine. *eLife*, 2(2), pp.1–25. Available at: http://elifesciences.org/lookup/doi/10.7554/eLife.00731.

Yue, J.-X. et al., 2012. Tracing the origin and evolutionary history of plant nucleotide-binding site-leucine-rich repeat (NBS-LRR) genes. *New Phytologist*, 193(4), pp.1049–1063. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22212278.

Zhang, R. et al., 2014. Paleo-evolutionary plasticity of plant disease resistance genes. *BMC genomics*, 15(1), p.187. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24617999 [Accessed July 11, 2014].

Zhang, Y. et al., 2004. Expression of RPS4 in tobacco induces an AvrRps4-independent HR that requires EDS1, SGT1 and HSP90. *The Plant journal : for cell and molecular biology*, 40(2), pp.213–24. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15447648.

Zhang, Y. et al., 2016. The Diversification of Plant NBS-LRR Defense Genes Directs the Evolution of MicroRNAs That Target Them. *Molecular Biology and Evolution*,

p.msw154. Available at: http://mbe.oxfordjournals.org/lookup/doi/10.1093/molbev/msw154.

Zhang, Y.-M. et al., 2016. Uncovering the dynamic evolution of nucleotide-binding site-leucine-rich repeat (NBS-LRR) genes in Brassicaceae. *Journal of Integrative Plant Biology*, 58(2), pp.165–177. Available at: http://doi.wiley.com/10.1111/jipb.12365.

Zhong, Y. & Cheng, Z.-M. (Max), 2016. A unique RPW8-encoding class of genes that originated in early land plants and evolved through domain fission, fusion, and duplication. *Scientific Reports*, 6(August), p.32923. Available at: http://www.nature.com/articles/srep32923.

SUPPLEMENTARY DATA

Organism	**Family**	**Other Taxonomic info**	**Source of data**
Ananas cosmus	Bromeliaceae	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Aquilegia coerulea	Ranunculaceae	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Arabidopsis halleri	Brasicaceaee	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Arabidopsis lyrata	Brasicaceaee	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Arabidopsis thaliana	Brasicaceaee	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Beta vulgaris	Chenopodiaceae	Viridiplantae	http://bvseq.molgen.mpg.de/Genome/Download/RefBeet-1.2/
Boechera stricta	Brasicaceaee	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Beta vulgaris	Chenopodiaceae	Viridiplantae	http://bvseq.molgen.mpg.de/Genome/Download/RefBeet-1.2/
Boechera stricta	Brasicaceaee	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Beta vulgaris	Chenopodiaceae	Viridiplantae	http://bvseq.molgen.mpg.de/Genome/Download/RefBeet-1.2/
Boechera stricta	Brasicaceaee	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Brachypodium distachyon	Poaceae	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Brachypodium stacie	Poaceae	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Bradyrhizobium diazoefficiens	Bradyrhizobiaceae	Bacteria	https://phytozome.jgi.doe.gov/pz/portal.html
Brassica rapa	Brasicaceaee	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Capsella grandiflora	Brasicaceaee	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Capsella rubella	Brasicaceaee	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Capsicum annuum Zunla	Solanaceae	Viridiplantae	ftp://ftp.solgenomics.net
Capsicum anuum CM334	Solanaceae	Viridiplantae	ftp://ftp.solgenomics.net

Carica papaya	Caricaceae	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Chlamydomonas reinhardtii	Chlamydomonadaceae	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Chloroflexus aurantiacus	Chloroflexaceae	Bacteria	ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/bacteria/Chloroflexus_aurantiacus/
Chloroherpeton thalassium	Ignavibacteriaceae	Bacteria	ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/bacteria/Chloroherpeton_thalassium
Chondrus crispus	Gigartinaceae	Red alga	http://ftp.gramene.org/CURRENT_RELEASE/data/fasta/Viridiplantaes_rhodophyta1_collection/
Citrullus lanatus	Cucurbitaceae	Viridiplantae	ftp://www.icugi.org/pub/genome/watermelon/97103/v1/
Citrus clementina	Rutaceae	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Citrus sinensis	Rutaceae	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Clostridium cellulovorans	Clostridiaceae	Bacteria	ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/bacteria/Clostridium_cellulovorans/
Coccomyxa subellipsoidea	Coccomyxaceae	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Coffea canephora	Rubiaceae	Viridiplantae	http://coffee-genome.org/download
Cucumis melo	Cucurbitaceae	Viridiplantae	https://melonomics.net/files/Genome/Melon_genome_v3.5_Garcia-Mas_et_al_2012/
Cucumis sativus	Cucurbitaceae	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Cyanophora paradoxa	Glaucocystaceae	//	http://cyanophora.rutgers.edu/cyanophora/Cyanophora_CLC_112010.fasta
Dacus carota	Apiaceae	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Dunaliella salina	Dunaliellaceae	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Ectocarpus siliculosus	Ectocarpaceae	Brown Alga	https://bioinformatics.psb.ugent.be/gdb/ectocarpus/
Eragrostis tef	Poaceae	Viridiplantae	http://130.92.252.158/tef/version1/
Eucalyptus grandis	Myrtaceae	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Eutrema salsugineum	Brassicaceae	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Fragaria vesca	Rosaceae	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Galdieria sulphuraria	Galdieriaceae	Red alga	http://ftp.gramene.org/CURRENT_RELEASE/data/fasta/Viridiplantaes_rhodophyta1_collection/
Gloeobacter violaceus	Gviolaceus	Bacteria	http://www.uniprot.org/taxonomy/251221
Glycine max	Fabaceae	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Gossypium raimondii	Malvaceae	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Hordeum vulgare	Poaceae	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Kadua laxiflora	Rubiaceae	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Kalanchoe marnieriana	Crassulaceae	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Lactuca sativa	Asteraceae	Viridiplantae	http://gviewer.gc.ucdavis.edu/fgb2/gbrowse/lechuga_version_1_2/
Linum usitatissimum	Linaceae	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Lotus japonicus	Fabaceae	Viridiplantae	ftp://ftp.kazusa.or.jp/pub/lotus/lotus_r3.0/
Malus domestica	Rosaceae	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Manihot esculenta	Euphorbiaceae	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Marchantia polymorpha	Marchantiaceae	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Medicago truncatula	Fabaceae	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Methanosarcina mazei go1	Methanosarcinaceae	Bacteria	ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/bacteria/Methanosarcina_mazei

Micromonas pusilla	Mamiellaceae	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Micromonas sp.RCC299	Mamiellaceae	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Mimulus guttatus	Phrymaceae	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Musa acuminata	Musaceae	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Nicotiana benthamiana	Solanaceae	Viridiplantae	ftp://ftp.solgenomics.net
Nicotiana sylvestris	Solanaceae	Viridiplantae	ftp://ftp.solgenomics.net
Nicotiana tabacum	Solanaceae	Viridiplantae	ftp://ftp.solgenomics.net
Nostoc punctiforme PCC 73102	Nostocaceae	Bacteria	ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/bacteria/Nostoc punctiforme
Olea europea	Oleaceae	Viridiplantae	ftp://climb.genomics.cn/pub/10.5524/100001_101000/100201/
Oropetium thomaeum	Poaceae	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Oryza sativa	Poaceae	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Ostreococcus lucimarinus	Bathycoccaceae	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Panicum hallii	Poaceae	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Panicum virgatum	Poaceae	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Petunia axilaris	Solanaceae	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Petunia infilata	Solanaceae	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Phaseolus vulgaris	Fabaceae	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Phoenix dactylifera	Arecaceae	Viridiplantae	ftp://ftp.ncbi.nlm.nih.gov/genomes/Phoenix_dactylifera/GFF
Phyllostachys heterocycla	Poaceae	Viridiplantae	http://202.127.18.221/bamboo/down.php
Physcomitrella patens	Funariaceae	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Picea abies	Pinaceae	Viridiplantae	http://congenie.org/start
Pinus taeda	Pinaceae	Viridiplantae	http://dendrome.ucdavis.edu/ftp/Genome_Data/genome/pinerefseq/Pita/v1.01/gene_models/
Populus trichocarpa	Salicaceae	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Prunus Persica	Rosaceae	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Rhodopirellula baltica	Planctomycetaceae	Bacteria	http://www.genome.jp/dbget-bin/get_linkdb?uniprot:Q7UEH8_RHOBA
Ricinus communis	Euphorbiaceae	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Salix purpurea	Salicaceae	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Selaginella moellendorffii	Selaginellaceae	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Sesamum indicum	Pedaliaceae	Viridiplantae	http://ocri-genomics.org/Sinbase/login.htm
Setaria italica	Poaceae	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Setaria viridis	Poaceae	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Solanum lycopersicum	Solanaceae	Viridiplantae	ftp://ftp.solgenomics.net
Solanum melongena	Solanaceae	Viridiplantae	ftp://ftp.solgenomics.net
Solanum pennellii	Solanaceae	Viridiplantae	ftp://ftp.solgenomics.net
Solanum pimpinellifolium	Solanaceae	Viridiplantae	ftp://ftp.solgenomics.net
Solanum tuberosum phureja	Solanaceae	Viridiplantae	ftp://ftp.solgenomics.net
Sorghum bicolor	Poaceae	Viridiplantae	ftp://ftp.solgenomics.net

Spirodela polyrhiza	Aracaceae	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Thauera aminoaromatica S2	Rhodocyclacea e	Bacteria	ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/bacteria/Thauera_aminoaromatica
Theobroma cacao	Malvaceae	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Trifolium pratense	Fabaceae	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Triticum aestivum	Poaceae	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Triticum urartu	Poaceae	Viridiplantae	ftp://ftp.ensemblgenomes.org/pub/Viridiplantaes/release-33/fasta/triticum_urartu/
Vitis vinifera	Vitaceae	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Volvox carteri	Volvocaceae	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Zea mays	Poaceae	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html
Zostera marina	Zosteraceae	Viridiplantae	https://phytozome.jgi.doe.gov/pz/portal.html

Tables S1. The 102 sequenced genomes used for identification of NLR-like genes and their download sources.