UNIVERSITÀ DEGLI STUDI DI NAPOLI "FEDERICO II"
DOTTORATO DI RICERCA IN AGROBIOLOGIA E AGROCHIMICA – XXII CICLO
INDIRIZZO MIGLIORAMENTO GENETICO E ORTICOLTURA

GENOMICA AVANZATA PER LO STUDIO DEI GENI DI RESISTENZA A PATOGENI IN PIANTA

CANDIDATO
DOTT WALTER SANSEVERINO

RELATORE
DOTT.SSA MARIA RAFFAELLA ERCOLANO

CORRELATORE
CHIAR.MO PROF. LUIGI FRUSCIANTE

COORDINATORE
CHIAR.MO PROF. MATTEO LORITO

UNIVERSITÀ DEGLI STUDI DI NAPOLI "FEDERICO II"
DOTTORATO DI RICERCA IN AGROBIOLOGIA E AGROCHIMICA – XXII CICLO
INDIRIZZO MIGLIORAMENTO GENETICO E ORTICOLTURA

GENOMICA AVANZATA PER LO STUDIO DEI GENI DI RESISTENZA A PATOGENI IN PIANTA

CANDIDATO
DOTT WALTER SANSEVERINO

RELATORE
DOTT.SSA MARIA RAFFAELLA ERCOLANO

CORRELATORE
CHIAR.MO PROF. LUIGI FRUSCIANTE

COORDINATORE
CHIAR.MO PROF. MATTEO LORITO

1. Introduzione

1.1 Le piante e la loro risposta agli stress biotici

Come tutti gli organismi viventi, anche le piante, per sopravvivere, hanno sviluppato accurati e sofisticati sistema di difesa in grado di riconoscere e combattere le minacce esterne. Gli stress a cui le piante sono sottoposte, se indotti da altri organismi viventi quali batteri, virus, funghi, nematodi, insetti ed animali erbivori, sono definiti *biotici* (1). I meccanismi che regolano il rapporto tra una pianta ed il proprio ospite, definiti interazioni pianta-patogeno o relazioni antagoniste, sono complessi ed interessano svariati aspetti della biologia degli organismi. Le piante sono continuamente sottoposte all'attacco dei patogeni che si nutrono sia della loro parte aerea che dei tessuti interni, penetrando attraverso la cuticola delle foglie o attraverso l'apparato radicale (2).

Per difendersi, i vegetali hanno sviluppato nel tempo tutta una serie di meccanismi che tendono a riconoscere precocemente la presenza del patogeno e a porre in atto la sintesi di composti di difesa antimicrobici (fitoalessine) (3,4). Le piante sono resistenti alla maggior parte dei patogeni, per cui la comparsa dell'infezione è un'eccezione. La malattia è la manifestazione di un'interazione genica compatibile tra ospite e patogeno in cui il patogeno è definito come "virulento" e l'ospite come "suscettibile", cioè non possiede nessun gene di resistenza contro il patogeno. Una mancata infezione è invece il risultato di una interazione genica incompatibile tra ospite e patogeno in cui il patogeno è "avirulento" e l'ospite "resistente" (5).

Il riconoscimento dei patogeni nelle piante è spesso determinato da singoli geni R (di resistenza) (6), mentre il gene corrispondente nel patogeno è chiamato "gene avr " (di avirulenza). L'interazione "gene R/gene avr" è basata sulla specificità dell'interazione stessa e, in assenza di riconoscimento (assenza di gene R o avr), si manifesta la malattia. Affinché si verifichino resistenza o immunità, la pianta deve essere portatrice del gene R e il patogeno deve essere portatore del gene avr . La resistenza delle piante ai patogeni è quindi spesso determinata da singoli geni presenti in entrambi gli organismi interagenti. In questa interazione "gene per gene", la resistenza si verifica solo se il gene R della pianta interagisce con il corrispondente gene avr del patogeno (7). Da un punto di vista

biochimico, i geni R codificano proteine in grado di riconoscere prodotti diretti o indiretti di un gene avr del patogeno (chiamati generalmente "elicitori").
Le interazioni tra piante e patogeni possono essere spostate in favore della pianta o del patogeno da piccole variazioni ambientali, principalmente temperatura e stato nutrizionale della pianta, e non dipendono solo dai genomi dei due organismi coinvolti. Inoltre, anche lo stadio di sviluppo di una pianta può influenzare la resistenza o la suscettibilità ad un dato patogeno.
E' comunque importante notare che le infezioni dei patogeni non causano sempre una risposta netta nella pianta, ma spesso determinano reazioni intermedie chiamate "tipi di infezione" (pianta immune, altamente resistente, moderatamente resistente, moderatamente suscettibile, completamente suscettibile) (8).
Il prodotto di un gene R potrebbe essere una proteina recettore citoplasmatica che può legare diverse molecole (9). I geni avr del patogeno, invece, non agiscono da soli, ma sono spesso abbinati a delle proteine che inibiscono l'attività dei geni R.
Sebbene il prodotto del gene avr sia spesso chiamato "elicitore", perché provoca una risposta di difesa antimicrobica, ci sono molti altri elicitori non codificati da classici geni avr, come ad esempio prodotti del metabolismo batterico, frammenti di parete dei funghi, sotto-prodotti dei danni o delle risposte delle cellule vegetali, o ancora trattamenti chimici o stress, che imitano l'impatto degli elicitori sulle cellule vegetali (ad es. ozono o UV-C), anche in assenza di patogeni, aumentando, in tal modo, i sistemi di protezione naturale del vegetale (10).
Inoltre tutti gli elicitori sono proteine che derivano dai prodotti primari dei geni avr oppure enzimi coinvolti nella loro via di produzione ed in tutti i casi agiscono solo in piante portatrici del gene di resistenza (R) corrispondente.
Una risposta efficace e attiva spesso determina nella pianta l'insorgenza di una reazione in cui le cellule vicine al sito di infezione vanno incontro alla morte e bloccano la crescita del patogeno privandolo di sostanze nutritive e creando un ambiente altamente ossidante che danneggia le sue proteine e strutture cellulari (11). Tale reazione è innescata dall'elicitore, il quale attiva geni per proteine legate a fattori che codificano per proteine di resistenza, le quali agiscono direttamente o indirettamente sullo sviluppo del patogeno nella pianta.
I siti di infezione sono le zone principali in cui avviene l'attivazione di una serie di geni di difesa della pianta. La successiva sintesi di ulteriori metaboliti protettivi e di proteine

inibitorie intorno ai siti di infezione è considerata importante per bloccare l'invasione del patogeno.

Il riconoscimento pianta-patogeno avviene tra recettori posti sulla membrana della cellula vegetale e molecole "elicitori", derivanti ad esempio dall'idrolisi della propria parete, conseguente alla produzione di enzimi litici da parte del patogeno, o con sostanze che si liberano dal patogeno stesso; ciò induce le cellule, contigue al punto di invasione, ad attivare un enzima situato nella membrana cellulare, in grado di produrre molecole che innescano lo "scoppio ossidativo" e composti con forte potere ossidante, avviando una vera e propria morte cellulare programmata, detta risposta di ipersensibilità (HR), che tende a fare terra bruciata intorno al parassita, impedendone la moltiplicazione (11).

Ulteriori segnali, non ancora identificati, derivanti dalla risposta di ipersensibilità, si muovono in tutta la pianta e attivano la resistenza sistemica acquisita (SAR), con la quale anche le cellule che non sono venute direttamente in contatto con l'elicitore vanno incontro occasionalmente a morte programmata, ma, principalmente, sono indotte a produrre fitoalessine (12).

In poche ore dalla necrosi localizzata, la pianta comincia ad esprimere una serie di geni correlati alla patogenesi (PR, proteine di resistenza) sia localmente, nel punto di infezione, sia sistemicamente in tutto il resto della pianta, nei tessuti distali (13).

Tra gli enzimi principali implicati nella risposta di difesa della piante, le β-glucanasi e le chitinasi giocano un ruolo chiave. Le attività di questi enzimi sono correlate con la resistenza della pianta in quanto i loro substrati, β-1,3-glucano e chitina, sono i principali componenti della parete cellulare di molti funghi patogeni e vanno a degradare le giovani ife fungine durante la loro penetrazione nei tessuti dell'ospite (14).

1.2 Il sistema immunitario vegetale dei geni R

Conoscere il sistema immunitario vegetale è di fondamentale importanza sia dal punto di vista scientifico sia dal punto di vista socio-economico. Dal punto di vista scientifico sarebbe importante ed affascinante chiarire i meccanismi di difesa delle specie vegetali e comprendere i processi cellulari coinvolti. Dal punto di vista socio-economico la conoscenza del sistema immunitario vegetale consente e consentirà di coltivare specie vegetali resistenti che non necessitano di agenti chimici deputati alla difesa, con una

maggior salvaguardia dell'ambiente e della salute umana (15). Nonostante l'importanza di questo argomento e la dedizione di molti ricercatori, gli studi in questo campo procedono più lentamente rispetto ad altri a causa di una condizione necessaria per la sperimentazione: l'interazione della pianta con il patogeno e l'instaurarsi della malattia. Di grande impatto scientifico è stato il manoscritto di J. Jones, "*The plant immune system*" del 2006, nel quale si demarcano alcuni punti saldi nello studio del sistema immunitario vegetale (16). Tale articolo illustra il sistema immunitario diviso in due rami principali: il primo, con un'azione più generica definito PTI (PAMP-triggered immunity), esplica la sua funzione attraverso recettori vegetali ad ampio spettro che riconoscono un pattern generico del patogeno definito PAMP (Pathogen Associated Molecolar Pattern), ed il secondo, molto specifico, utilizza la classe di proteine definite di resistenza (R) che, riconoscendo proteine patogeniche (definite proteine di avirulenza), sviluppano un'interazione che attiva una risposta di difesa da parte della pianta.

Lo studio delle proteine R iniziò con Flor nel 1946 che teorizzò, per la prima volta, la possibilità per un patogeno di essere riconosciuto dalla pianta attraverso una singola interazione tra due proteine, una di pianta (proteina di resistenza o R) ed una del patogeno (proteina di avirulenza o avr). Questa interazione fu definita *"gene for gene interaction"* (interazione a singolo gene) (7) e, ad oggi, è ancora una valida ipotesi per lo studio del sistema immunitario vegetale.

Convenzionalmente, i geni correlati alla risposta di resistenza nei confronti del patogeno sono definiti geni di resistenza (17). Il panorama dei geni di resistenza è vasto e variegato ma diversi studi hanno dimostrato che tali geni hanno caratteristiche comuni:

- sono posizionati in specifiche regioni cromosomiche e sono strutturati in cluster (18);
- molti cluster di resistenza contengono geni omologhi posizionati in tandem (19);
- in molte specie i cluster di resistenza sono situati in regioni sinteniche (20);
- un singolo gene di resistenza può conferire resistenza a diversi patogeni (anche di classi diverse) o a varianti di un patogeno (21, 22);
- geni R isolati in una specie sono attivi anche in altre specie, in particolare tra quelle tassonomicamente più vicine (23).

I geni R sono in grado di svolgere la propria funzione poiché contengono dei domini funzionali tipici ed in base ad essi sono divisi in 5 classi (24, 25). Tutti i geni di resistenza caratterizzati, per svolgere la propria funzione, hanno bisogno di un sito di

legame (Nucleotide Binding Site - NBS) e di sequenze ripetute ricche di leucina (Leucine Rich Repeat - LRR). Tali domini possono essere presenti in un unico gene o in più geni tra essi associati, come nel caso dei geni *Pto* e *Prf* (26, 22). Il dominio LRR è un dominio variabile coinvolto nell'interazione proteina-proteina ed è il maggiore determinante del processo di riconoscimento pianta-patogeno (27).

Dal punto di vista evoluzionistico la selezione per una nuova e diversa capacità ricognitiva è un fenomeno molto complesso. Diversi studi hanno dimostrato che la famiglia dei geni R si evolve attraverso sistemi di selezione positiva o bilanciata (28) e che diversi meccanismi come l'inserzione nel cluster di elementi trasponibili, lo scambio di sequenze e la metilazione del DNA, sono correlati all'espressione ed all'evoluzione degli stessi (29, 30).

Il primo gene di resistenza funzionale isolato nelle *Solanaceae* è stato il gene *Pto* che conferisce resistenza a *Pseudomonas siringaea* in pomodoro (31). Nonostante la dimostrazione che *pto* sfrutti l'interazione a singolo gene, fu poi chiarito che un altro gene, *prf*, localizzato in prossimità di *pto*, era necessario per il corretto funzionamento del sistema d'interazione pianta-patogeno. Specie vegetali aventi nel proprio patrimonio genetico i geni *pto* e *prf* sono in grado di riconoscere il patogeno *Pseudomonas siringaea* attraverso due passaggi chiave: il riconoscimento, tramite *pto*, della proteina patogenica avr*pto* e la trasduzione di un segnale di difesa ad opera del prodotto genico di *prf*. Molti sono gli studi che si sono susseguiti dagli anni '90 ad oggi sui geni R migliorando moltissimo la conoscenza di questa famiglia genica. Questi studi non solo hanno messo in luce i passaggi che regolano l'interazione a singoli geni (32), ma ad esso hanno anche affiancato altri meccanismi che dimostrano la formazione di complessi proteici all'interno dell'organismo vegetale per il riconoscimento del patogeno (33). Fino ad ora solo una minima parte di tutti i geni R è stata esplorata e caratterizzata (34, 35); ad oggi sono stati isolati 73 geni R e, di questi, 32 afferiscono alla famiglia delle *Solanaceae* (9). La disponibilità di un'ampia collezione di sequenze correlate ai processi di difesa e la disponibilità delle sequenze provenienti dai progetti di sequenziamento offre nuove opportunità per chiarire i meccanismi di resistenza nelle piante.

1.3 La famiglia delle *Solanaceae* come modello di studio per le resistenze vegetali

La famiglia delle *Solanaceae* è importante sia da un punto di vista scientifico sia da un punto di vista sociale (36). Le *Solanaceae* devono la loro importanza a diversi fattori tra cui la loro adattabilità ambientale, la loro plasticità genetica, il grande numero di specie da cui è composta la famiglia ed il loro alto valore nutriceutico, salutistico ed economico. Originaria del centro America, la famiglia conta più di 2000 specie e, grazie alla propria adattabilità, ha conquistato ambienti inospitali come deserti, terreni rocciosi, regioni fredde o con grosse escursioni climatiche. Tale adattabilità è dovuta al loro genoma che oltre a permettere lo sviluppo negli ambienti più disparati ha permesso anche una grossa differenziazione in sotto-famiglie. Tale plasticità, oltre ad avere una grossa importanza genetica, ha permesso alle diverse specie di intraprendere percorsi evolutivi di differenziazione che oggi hanno portato ad un grosso utilizzo delle *Solanaceae* in ambito alimentare, officinale ed ornamentale (37). Dal punto di vista alimentare basti pensare che la famiglia delle *Solanaceae* comprende alcune delle più importanti specie vegetali agrarie quali la patata (*Solanum tuberosum*), il pomodoro (*Solanum lycopersicum*), la melanzana (*Solanum melongena*) ed il peperone (*Capsicum annuum*), oltre a piante d'interesse officinale come la belladonna, lo stramonio e la datura, sociale come il tabacco ed ornamentale come la petunia. L'importanza di questa famiglia è data non solo dagli aspetti genetici ed evolutivi degli organismi stessi ma anche da aspetti socio-economici che vedono le *Solanaceae* ai primi posti per le produzioni alimentari e per l'indotto economico presente in Italia e nel mondo. A tale proposito l'Italia è la sesta produttrice di pomodoro, con un giro economico di circa un miliardo e mezzo di dollari ed una quantità prodotta di circa 6 milioni di tonnellate annue, e la nona produttrice di tabacco con un giro economico di 200 milioni di dollari e circa 100 mila tonnellate annue prodotte (FAO-stat 2007).

Questi dati portano alla luce l'importanza delle *Solanaceae* nella nostra società ed è facile intuire come questa famiglia sia diventata un modello di studio per l'aumento della produzione e della resa delle colture, per lo studio delle patologie vegetali e della risposta alle malattie, per lo studio della genetica e genomica vegetale. Questa famiglia interigesce con molti patogeni, ha una grossa diversità naturale e quindi è una grossa fonte di nuove resistenze (si conoscono circa 100 geni di resistenza di cui molti clonati). Inoltre alcune specie sono facili da trasformare attraverso le tecnica mediata da

Agrobacterium ed attraverso i VIGS. In particolare, il genoma del pomodoro, per le sue caratteristiche intrinseche e le risorse genetiche ad esso associate, può facilitare molte scoperte nel campo della genetica vegetale. Per tutte queste motivazioni nel 2005 è iniziato il progetto internazionale "the Tomato Sequencing Project" per sequenziare il genoma del pomodoro (950 Mb) (38). L'isolamento di nuovi geni R, in particolare nella famiglia delle *Solanaceae,* ed il loro trasferimento, attraverso approcci classici di miglioramento genetico, può portare molti vantaggi in termini ecologici, economici e di salute nell'ottica di una sempre più diffusa agricoltura sostenibile. Esplorare la base molecolare della variazione genomica può essere utile per l'identificazione di nuovi geni di resistenza e per migliorare le nostre conoscenze sul loro meccanismo di funzionamento. Ad oggi le principali caratteristiche di tali geni sono state desunte proprio dallo studio delle piante appartenenti alla famiglia delle *Solanaceae*.

1.4 Strumenti per l'indagine e l'analisi dei geni R

Nel 1993 Gregory Martin, tramite un approccio di clonaggio posizionale, riuscì ad isolare in pomodoro il primo gene di resistenza *pto* (31). Per molti anni l'utilizzo di approcci posizionali ha segnato la ricerca nel campo delle resistenze vegetali portando all'isolamento di molti altri geni tramite l'utilizzo di marcatori molecolari ed analisi di linkage (39, 40, 9). Negli anni '90 sono state costruite molte mappe genetiche e sono stati identificati molti marcatori molecolari associati alle resistenze, aggiungendo numerose nuove informazioni su questa famiglia genica, tra cui il fatto che i geni R si trovano in regioni genomiche ben precise (hot spot), che sono strutturati in cluster e che sono conservati tra specie affini (41).

Queste novità hanno permesso di utilizzare approcci diversi per il clonaggio di nuove sequenze, andando ad utilizzare tecniche come il "chromosome walking" per l'analisi delle regioni adiacenti al gene già isolato ricche di geni omologhi, il "clonaggio funzionale" per il rapido clonaggio di sequenze funzionalmente simili ad altre già isolate e funzionali, ed il "candidato per posizione" per sfruttare l'ortologia genica tra specie affini (42, 43) . L'utilizzo di queste tecniche, che hanno trovato lo loro massima espressione nella ricerca dei geni di resistenza alla fine degli anni '90, ha permesso di isolare altri geni di resistenza funzionali, di approfondire la struttura dei cluster genici,

di portare alla luce moltissime sequenze omologhe ad i geni già isolati nella stessa specie o in specie affini.

Il nuovo millennio ha segnato l'avvento dell'era della genomica con la pubblicazione del primo genoma sequenziato e con l'inizio di molti progetti di sequenziamento di organismi vegetali (44, 45). L'approccio genomico alle resistenze da una parte ha chiarito moltissimi meccanismi molecolari ed evoluzionistici di questa famiglia genica, ma da un'altra parte non ha portato i benefici che ci si aspettava per quanto riguarda l'isolamento di nuovi geni di resistenza. I motivi che non hanno permesso l'utilizzo delle informazioni genomiche per lo scopo sopradetto sono diversi e se da una parte sono da imputare alla struttura stessa del sistema immunitario vegetale che, a causa delle sue caratteristiche, rende pressoché impossibile discernere geni di resistenza funzionali da geni omologhi, silenziati, o con diverse caratteristiche di riconoscimento, se non con approfonditi e lunghi studi di biologia molecolare ed ingegneria genetica, dall'altra è da imputare anche alla qualità delle informazioni stesse provenienti da esperimenti di genomica su larga scala che ad oggi sono ancora carenti di qualità, ordine e precisione. Alla situazione descritta bisogna aggiungere un dettaglio non trascurabile: i geni R funzionali si trovano in specie selvatiche e non domesticate, mentre gli esperimenti di genomica vengono effettuati su specie coltivate, prive o poco ricche di geni R funzionali (46).

Le problematiche descritte già da tempo sono discusse dalla comunità scientifica e il rapido evolversi delle tecnologie informatiche sta risolvendo velocemente molte delle questioni poste. La soluzione è chiaramente quella di utilizzare un approccio comparativo e multidisciplinare, approfondendo queste tematiche in diversi campi scientifici e sfruttando al meglio metodologie integrate. Partendo dall'informatica ed arrivando alla biologia molecolare attraverso la genomica è possibile chiarire le numerose caratteristiche dei geni di resistenza ed arrivare a risultati concreti e fruibili dalla comunità scientifica per il bene comune. Studi genomici, strutturali, funzionali devono essere effettuati attraverso una visione d'insieme della problematica, organizzando le informazioni grezze e indirizzandole abilmente allo scopo prefisso.

1.5 Metodologie innovative per lo studio dei geni R

Gli ultimi anni sono stati caratterizzati dall'avvento di nuove tecnologie, dalla possibilità di processare un numero sempre maggiore di dati e di ottenere sempre più informazioni. Tecniche di genomica e biologia molecolare, come sequenziamento, analisi comparative ed espressione differenziale, sono diventate di routine (47, 48). Queste metodologie ad elevata componente tecnologica hanno la caratteristica di produrre dati in quantità ed in informatività elevatissima. Tale caratteristica ha modificato completamente l'approccio dei genetisti alla materia ed il background che questi devono avere per portare a termine le proprie ricerche. Mentre fino a pochi anni fa era norma studiare set con un numero ridotto di geni, oggi è consuetudine studiare genomi interi o più genomi contemporaneamente. Questa modifica delle tecniche di supporto alla analisi biologica hanno permesso la nascita di nuovi campi di studio come la "biologia dei sistemi", la "biologia comparativa" e la bioinformatica (49, 50, 51).

La biologia dei sistemi, "system biology", è una disciplina biologica che studia gli organismi viventi in quanto sistemi che si evolvono nel tempo, ossia nell'interazione dinamica delle parti di cui sono composti. In particolare questo obiettivo viene conseguito tramite l'integrazione di modelli dinamici e dei risultati di differenti esperimenti ad alto rendimento, "high throughput", unendo nella pratica per esempio le conoscenze di genomica, proteomica, trascrittomica e teoria dei sistemi dinamici. La biologia dei sistemi parte quindi dalla conoscenza dei geni e delle proteine presenti nel corso del tempo in un organismo e utilizza tecniche di trascrittomica per determinare cambiamenti nell'espressione genica e, per valutare i cambiamenti dinamici derivati da una perturbazione del sistema. Lo scopo è quindi quello di arrivare a creare un modello sempre più completo del funzionamento dei sistemi biologici.

La biologia comparativa, "comparative biology", sfrutta invece un approccio multidisciplinare per lo studio della biodiversità e per mettere in luce le correlazioni filogenetiche tra gli organismi. Partendo dai geni, fino ad arrivare alle relazioni ecosistematiche tra gli organismi, la biologia comparativa ha come scopo quello di creare collegamenti tra le varie specie e permettere il trasferimento delle informazioni. Diramazione della biologia comparativa è la genomica comparativa che, partendo dalle mappe genetiche e dalle informazioni di sequenziamento, permette la comparazione degli organismi a livello genomico (52). Nella famiglia delle *Solanaceae* la comparazione dei genomi di patata e pomodoro ha mostrato che questi differiscono solo per 5

inversioni paracentriche (53), mentre il genoma del pomodoro differisce da quelli di peperone e melanzana per diversi riarrangiamenti (54, 55). Come si è potuto comprendere, studi di genomica comparativa hanno rilevato un alto livello di conservazione nel numero e nell'ordine dei geni tra le diverse specie di questa famiglia. L'abbattimento dei costi permette di progettare esperimenti in cui più genomi possono essere confrontati tra loro in modo esaustivo (56) e con tecniche comparative, come il "Sequence Capture", è possibile utilizzare un genoma di riferimento per sequenziare specifiche regioni su specie affini. Un approccio del genere può essere utilizzato per le specie selvatiche del genere *Solanum,* che continuano ad essere la maggior fonte di geni resistenza e pertanto vanno esplorate e caratterizzate per fronteggiare gli stress biotici.
Strumento indispensabile per lo studio delle discipline descritte e per l'interazione con le nuove tecnologie genomiche è la bioinformatica, che si occupa di fornire modelli validi per l'interpretazione dei dati provenienti da esperimenti di biologia molecolare e biochimica. In questo modo è possibile identificare tendenze e leggi numeriche, generare nuovi modelli e strumenti matematici per l'analisi di sequenze, al fine di creare un corpus di conoscenze relative alla loro evoluzione ed eventuale funzione. Queste discipline riescono quindi ad incanalare in modo appropriato le informazioni ottenute tramite i sistemi di analisi ad alto rendimento ed a collegare e confrontare le informazioni tra di loro. Da questo punto di vista le informazioni ritenute poco fruibili per lo studio dei geni di resistenza diventano invece un bene prezioso per il loro studio, la loro caratterizzazione ed il loro clonaggio. Grazie ad approcci genomici, comparativi e bioinformatici è possibile ripensare lo studio del sistema immunitario vegetale, partendo da dati grezzi fino all'ottenimento di specifici set di geni, selezionati per essere nuovi e funzionali candidati per la risposta agli stress biotici.

1.6 Risorse genomiche per lo studio dei geni R

Molte sono le risorse genomiche vegetali prodotte grazie alle tecnologie innovative sopra descritte. Ad oggi le banche dati contengono circa 1 mln di sequenze nucleotidiche, 800mila sequenze proteiche, 20 mln di sequenze ESTs ed infine 600mila sequenze UniGene (57), tutte non ridondanti. Inoltre sono stati completamente sequenziati 5 genomi vegetali, *Arabidopsis thaliana* (45), vite (*Vitis vinifera*) (58), riso (*Oryza sativa*) (59), pioppo (*Popolus tricocharpa*) (60), sorgo (*Sorgum bicolor*) (61), patata (*Solanum tuberosum*), cetriolo (*Cucumis sativus*) e molti altri sono in fase di

sequenziamento, tra cui mais (*Zea mais*), pomodoro (*Solanum lycopersicum*), e loto (*Lotus Japonica*). Le fonti principali da cui è possibile attingere queste informazioni sono il National center for Biotechnology Information (NCBI) (62), che contiene tutti i dati prodotti dalla comunità scientifica, e poi risorse specifiche per le piante strutturate in database organizzati, utili per specifiche esigenze. Tra i più importanti troviamo il Solanaceae Genomic Network (SOL) (63), Il TAIR (64), il PlantGDB (65), il TIRG (66) e le organizzazioni specifiche per i genomi parzialmente o completamente sequenziati.

Tramite una visione multidisciplinare, che punta ad un profondo studio del sistema immunitario vegetale, è possibile utilizzare queste risorse per progettare studi focalizzati alla comprensione del funzionamento dei geni di resistenza, alla loro caratterizzazione, alla loro evoluzione ed alla loro conservazione tra gli organismi.

Per il ruolo scientifico ed economico che rivestono, così come è già stato fatto per altre famiglie geniche, è fondamentale creare una risorsa specifica per i geni di resistenza in modo da poter organizzare le informazioni e renderle facilmente disponibili a tutti i ricercatori del settore ed a tutti coloro che desiderano approfondire lo studio di questa famiglia. Inoltre è possibile utilizzare tali risorse per progettare e costruire specifici sistemi di analisi per la catalogazione e la predizione di nuovi geni di resistenza.

1.7 Scopo della tesi

Un importante obiettivo della ricerca genetica è quindi quello di collezionare le informazioni disponibili per i geni di resistenza, studiarne i meccanismi di funzionamento, effettuare studi evoluzionistici e comparativi, comprendere i meccanismi di interazione pianta-patogeno e raggiungere il traguardo finale di isolare nuovi geni di resistenza per fronteggiare le malattie non ancora sconfitte e quelle future nel rispetto dell'uomo e dell'ambiente.

Lo scopo di questo lavoro è:

- La catalogazione di tutti i geni R vegetali
- La creazione di una risorsa specifica dedicata ai geni R
- La creazione di un'interfaccia web di facile utilizzo per la condivisione dei dati ottenuti con la comunità scientifica
- La creazione di un sistema di predizione genica specifico capace di processare sequenze ESTs

- La creazione di un sistema di predizione genica specifico capace di processare interi genomi vegetali
- Un approfondito studio dei geni predetti e la loro catalogazione tramite un approccio di genomica comparativa
- L'analisi e la caratterizzazione *in silico* delle sequenze predette
- La caratterizzazione molecolare di geni candidati R predetti
- La conferma attraverso tecniche di biologia molecolare dei dati predittivi ottenuti
- Lo sviluppo di sistemi per l'identificazione di geni putativamente funzionali come geni di resistenza
- L'identificazione di geni R omologhi in specie affini

2. Materiali e metodi

2.1 Ricerca bibliografica e catalogazione geni R

Le informazioni sui 73 geni R caratterizzati ad oggi, sono state raccolte attraverso una minuziosa ricerca bibliografica esplorando risorse cartacee e multimediali e raccogliendo in un'unica scheda informazioni provenienti da fonti diverse. Cinquantacinque geni R sono stati raccolti dal manoscritto di G.B. Martin "*Understanding the function of disease resistance proteins*" (22), mentre gli altri 18 geni dalle singole fonti che ne riportavano il clonaggio e la caratterizzazione funzionale (25, 67-82). Tramite interrogazioni in banca dati è stato possibile risalire alle sequenze nucleotidiche e proteiche depositate, alle quali sono state correlate informazioni sulla tassonomia, degli organismi coinvolti nel processo di resistenza attraverso il portale NCBI Taxonomy (83). Infine per ogni relazione pianta-patogeno-gene R, sono state ricercate le informazioni sulla malattia provocata, sul metodo di attacco del patogeno e sui geni avr coinvolti nel processo. Una volta reperite tutte le informazioni riguardanti i singoli geni R è stato creato un database apposito per la loro corretta catalogazione. Questo pool di geni caratterizzati e corredati di molte informazioni è stato definito "set dei geni R di referenza".

2.2 Raccolta dati da database pubblici

Oltre ai geni R di referenza lo studio condotto ha utilizzato molti set di dati provenienti da fonti diverse. In primo luogo è stato necessario raccogliere tutte le sequenze depositate in NCBI putativamente correlate ai processi di resistenza e poiché tale raccolta è in continuo aggiornamento è stato creato un *script* in PERL (84) per automatizzare il processo di recupero delle sequenze. Attraverso questo sistema lo *script* si collega ad NCBI, ricerca tutte le sequenze nucleotidiche correlate ai processi di resistenza in pianta attraverso la query "*plants AND ("disease resistance gene" OR "disease resistance protein") NOT bacteria NOT virus*" e ne recupera il risultato. Durante il lavoro di tesi è stato necessario reperire anche tutte le proteine del genere *Viridiplantae* e del genere *Solanum* provenienti da NCBI ei proteomi virtuali annotati da diversi genomi vegetali. I proteomi utilizzati sono stati:

- *Arabidopsis thaliana* versione TAIR8

- *Vitis vinifera* versione 1
- *Popolus trichocarpa* versione 1.1
- *Lotus japonica* versione 07/05/2006
- *Oryza sativa* versione 4
- *Solanum lycopersicum* versione ITAG13 (genoma parziale)
- *Sorghum bicolor* versione sbi1.4 (genoma parziale)
- *Zea Mais* versione 0.157 (genoma parziale)

2.3 PRGdb: il Plant Resistance Gene Database

Tutte le informazioni raccolte riguardo la famiglia dei geni R sono state inserite e catalogate tramite il PRGdb. I passi che hanno portato alla creazione del database sono stati:

- Progettazione del database ed inserimento dati
- Creazione di un'interfaccia grafica
- Strumenti di ricerca del PRG ed integrazione del BLAST
- Creazione di un interfaccia privata per il controllo dati

2.3.1 Progettazione database ed inserimento dati

Il PRG è stato progettato secondo lo schema dei database relazionali RDBMS (85), nei quali vengono create tabelle composte da righe a lunghezza fissa, collegate tra loro tramite relazioni singole o complesse. Questo sistema di memorizzazione dati oltre ad essere molto efficiente ha il grande vantaggio di poter effettuare ricerche dettagliate, *query*, tra le diverse tabelle. Il linguaggio utilizzato è SQL ed i programmi di supporto sono MySQL v5.1 ed il pacchetto software relativo (http://www.mysql.com). I programmi relativi al db sono stati tutti scritti in linguaggio Perl utilizzando i moduli Bioperl (84). PRGdb è stato pensato per contenere tipologie di informazioni molto diverse tra loro, ma correlate. Infatti, com'è possibile osservare in figura 1 sono due le sezioni principali del db, la prima relativa alle informazioni strettamente correlate ai geni di resistenza, la seconda relativa alle informazioni sulla natura dell'interazione pianta-patogeno. Il db è attualmente composto da 14 tabelle nelle quali sono stati inseriti tutti i dati raccolti tramite le ricerche manuali ed automatiche sopra descritte. Le informazioni sono state inserite in modo automatico e controllate manualmente eliminando gli eventuali errori di annotazione o di inserimento.

2.3.2 Creazione di un' interfaccia grafica

L'interfaccia grafica è stata costruita in linguaggio php (http://www.php.net) e in Apache web server, sfruttando modelli pre-esistenti disponibili (http://www.apache.org). Attraverso l'interfaccia in php è stato possibile creare un sito internet dinamico ed integrato con il database. L'interfaccia è stata pensata per essere semplice, chiara ma anche accattivante in modo da rendere la navigazione utile e piacevole.

2.3.3 Strumenti di ricerca del PRG ed integrazione del "BLAST"

Caratteristica fondamentale del portale d'interfaccia al db (www.prgdb.org) è la possibilità di interagire tramite l'interfaccia grafica con le informazioni contenute nel database. Attraverso la pagina principale è possibile eseguire le seguenti ricerche:

- Ricerca sul set dei geni R di referenza
- Ricerca avanzata su tutte le sequenze presenti in PRG
- Ricerca tramite specie vegetale o patogeno
- Ricerca tramite BLAST

In questo modo l'utente in una sola schermata ha a disposizione diverse scelte e può procedere velocemente verso il dato di interesse. L'interfaccia delle ricerche è stata collegata al database attraverso un sistema in php in grado di trasformare le preferenze degli utenti in *query* automatiche sottomesse al database. Tra le possibili ricerche c'è anche quella per omologia di sequenza, attraverso l'algoritmo BLAST (86). Per tale ricerca è stata creata una pagina apposita, utilizzando l'ultima versione del software disponibile gratuitamente in rete.

2.3.4 Creazione di un'interfaccia privata per il controllo dati

Importante ai fini del controllo dei dati inseriti è l'interfaccia privata creata in concomitanza con quella pubblica. Attraverso il sito internet pubblico prgdb.org è possibile accedere alla sezione protetta per l'inserimento ed il controllo dei dati. Anch'essa sviluppata in php permette ai curatori del portale di modificare in tempo reale le informazioni presenti nella sezione pubblica andando ad interagire direttamente con la base dati. Attraverso questa sezione non solo è possibile modificare, inserire o cancellare i dati, ma anche creare o eliminare le relazioni tra di essi. Inoltre, tutti i geni R di referenza e tutte le informazioni correlate ad essi sono stati inseriti manualmente attraverso questa interfaccia.

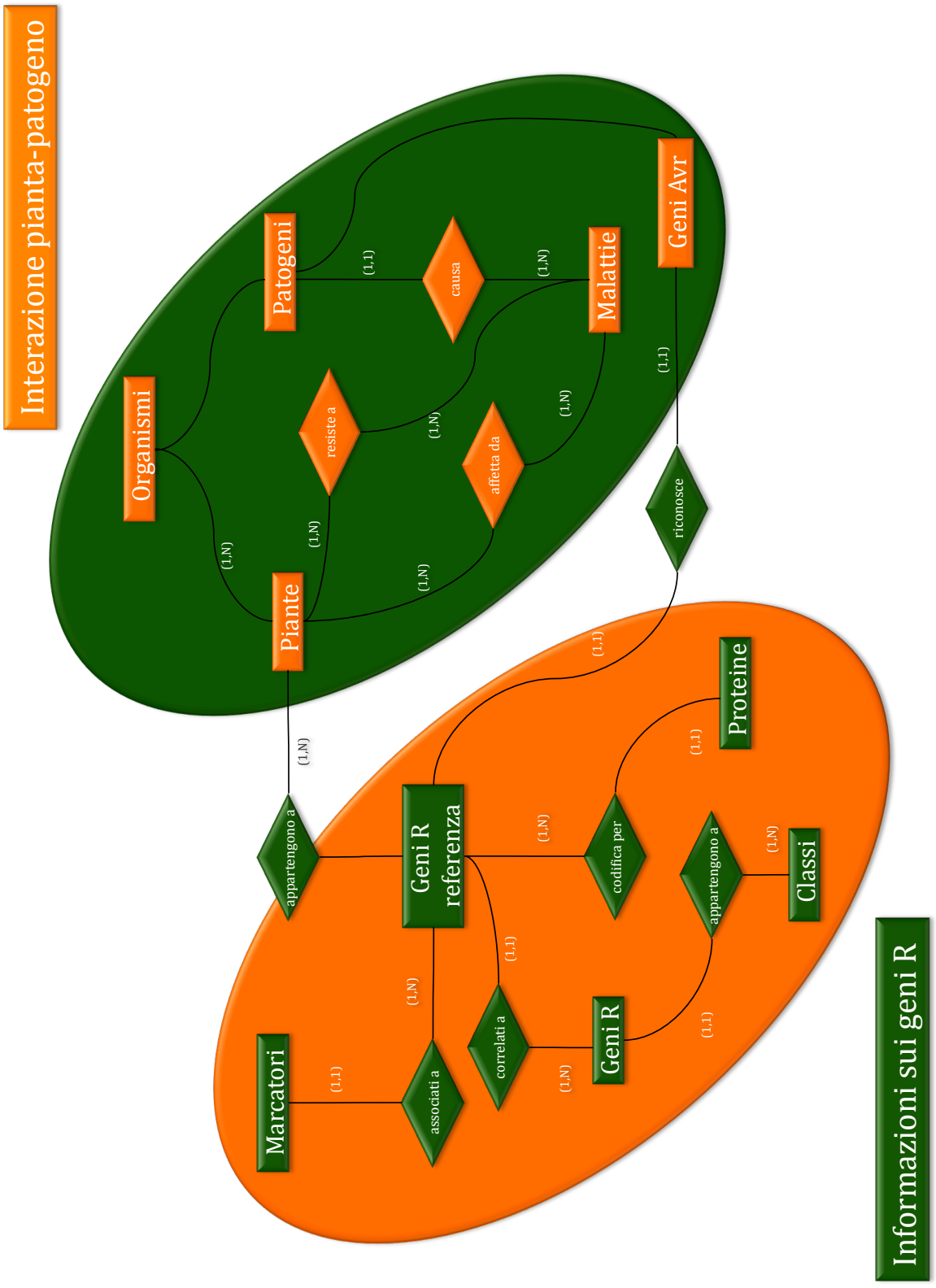


Figura 1. Schema logico del database PRG

2.4 Progettazione della "pipeline" di predizione DRAGO

DRAGO, acronimo di **D**isease **R**esistance **A**nalysis and **G**ene **O**rthology, è uno strumento bioinformatico, sviluppato appositamente per la predizione di geni, putativi per la funzione di resistenza, da set di sequenze ESTs. Per evitare la ridondanza delle informazioni e predire geni completi e non frammentati è stato scelto come set di partenza quello proveniente dalla sezione UniGene di NCBI. Da qui sono stati raccolti 604981 UniGene non ridondanti provenienti da 33 organismi vegetali differenti e tradotti in proteine attraverso il software ESTscan, versione 3.0.2 (87) utilizzando impostazioni predefinite, "codon usage" e matrice probabilistica di *Arabidopsis thaliana.* Le sequenze tradotte correttamente sono state 488250, le quali sono state poi comparate, tramite l'algoritmo BLAST, con i geni R di referenza. Tutte le sequenze risultate simili ai geni R con un E-value minore di 1*10-15 sono state raccolte. Le sequenze cosi ottenute sono state poi analizzate tramite InterProScan (88) versione 3.0.2 aggiornato con l'ultimo database disponibile, e divise, in base ai loro domini conservati, secondo le 5 già conosciute classi di resistenza. Il numero totale di sequenze, definite "putativi geni R predetti da UniGene", annotate tramite questo sistema è di 10463 (figura 2). La pipeline è utilizzata sfruttando un server Linux con distribuzione Gentoo (http://www.gentoo.org) ed il sistema di regolamentazione del lavoro sui cluster PBS (http://www.openpbs.org).

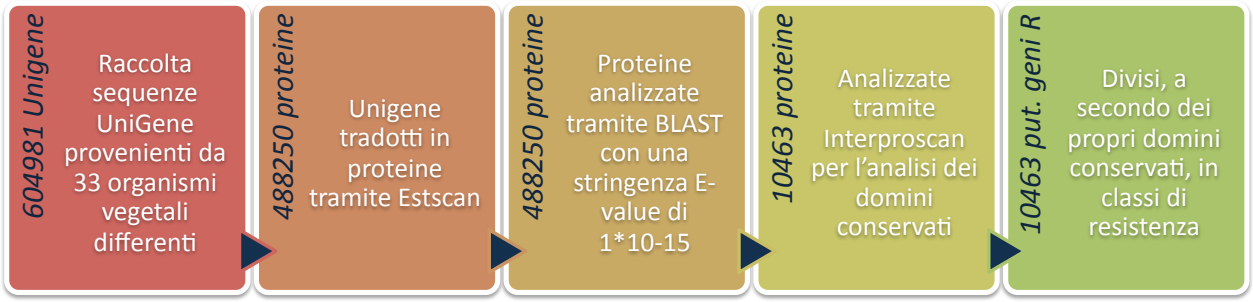


Figura 2. Sviluppo della "pipeline" di predizione ed annotazione DRAGO

2.4.1 Validazione del predittore DRAGO

Per verificare l'efficienza ed il margine d'errore del sistema di predizione DRAGO, la pipeline è stata utilizzata su un set di sequenze contenente un numero di geni di resistenza noti. Tramite quest'analisi è stato confermato che il sistema ha una efficienza del 100% con errore nella discriminazione dei geni di resistenza pari a 0.

2.5 Integrazione dei dati di predizione prodotti da DRAGO con il PRGdb

Per creare una risorsa completa sui geni R, i dati di predizione ottenuti tramite DRAGO sono stati inseriti ed intergrati nel database PRG. L'integrazione è avvenuta utilizzando MySQL per la creazione di nuove tabelle e per l'inserimento dei riferimenti dei putativi geni R. Di particolare importanza è stata l'integrazione con l'interfaccia grafica, avvenuta tramite php, attraverso la quale è attualmente possibile effettuare ricerche sul set di geni R putativi. Caratteristica di questo sistema è la possibilità di effettuare la ricerca per domini proteici o per classi di resistenza.

2.6 Sviluppo del sistema di predizione MATRIX

Il sistema di predizione ad alta efficienza MATRIX è stato sviluppato per analizzare grandi set di dati al fine di identificare le sequenze putative per la funzione di resistenza. Primo passaggio necessario alla creazione di questo nuovo sistema di analisi è la suddivisione tramite filogenesi (vedi materiali e metodi paragrafo filogenesi) dei geni di resistenza appartenenti alla famiglia delle *Solanaceae.* Le sequenze proteiche dei geni così suddivisi sono state raccolte in unico file multi fasta ed allineate tramite il programma MUSCLE (89). Gli allineamenti sono stati poi ripuliti manualmente ed utilizzati come base per la creazione di moduli HMM (90). Le regioni ad alta omologia dei gruppi di geni allineati sono state estrapolate attraverso programmi scritti in perl (vedi risultati paragrafo MATRIX). Attraverso *script* ad hoc sono state estrapolate le regioni altamente conservate dei singoli gruppi di proteine R e tramite il pacchetto HMMER versione 2.3.2 (91) sono stati creati i profili HMM attraverso *hmmbuilt* e calibrati tramite *hmmcalibrate*. I profili così costruiti sono poi stati allineati, utilizzando *hmmalign,* con le proteine sconosciute, per identificare putativi geni di resistenza. Per automatizzare il processo di allineamento dei profili HMM con le proteine da analizzare è stato scritto, in perl, un programma in grado di leggere le sequenze proteiche in modo sequenziale da file multi fasta, allineare tutti i profili HMM, uno alla volta, sulla proteina e, basandosi sulla matrice di sostituzione BLOSUM62, produrre un valore pari all'omologia tra il singolo profilo e la proteina. Questo processo, una volta che il software avrà analizzato tutte le proteine del set sottomesso, produrrà una matrice numerica che rappresenta l'omologia di tutte le proteine con tutti i profili HMM. Poiché il sistema sopra descritto, nominato MATRIX per la produzione di matrici numeriche come risultato finale, è influenzato dalla grandezza e dalla specificità del set di dati sottoposto

e dall'ordine dei moduli HMM, sono stati creati dei parametri configurativi opzionali attraverso i quali è possibile scegliere le impostazione adeguate al proprio set di dati prima dell'utilizzo del sistema di predizione. Nonostante il software effettui operazioni numeriche complesse ed a molte cifre, il sistema costruito è stato ottimizzato riducendo notevolmente il tempo computazionale e processando un genoma intero composto da circa 30000 proteine in circa 3-5 minuti.

Anche la matrice numerica ottenuta viene ottimizzata eliminando, ancor prima di essere mostrata, tutte le sequenze che non hanno allineato neanche con un modulo HMM riducendone di molto la dimensione. Inoltre, nel caso si desideri rendere l'analisi ancor più stringente, è possibile chiedere al software di eliminare anche le sequenze con un numero basso di allineamenti con i profili HMM. Nel caso della sperimentazione con i geni R è stato scelto di eliminare tutte le proteine che contenevano meno di 8 moduli HMM (vedi discussioni).

Per rendere il prodotto di Matrix piacevole alla vista e facile da interpretare anche per i meno esperti del mestiere è stato scelto di visualizzare le matrici numeriche attraverso i tanti software di raggruppamento presenti. Questa tipologia di software semplicemente trasforma il numero della matrice in una gradazione di colore in funzione del livello di omologia. Numeri più alti significano omologie più alte e quindi colori più accesi. Per rendere le matrici prodotte leggibili tramite questi software, è stato necessario implementare con un'ultima opzione lo *script* di MATRIX, trasformando tutti i valori ottenuti dagli allineamenti. Infatti le matrici matrix posso avere valori da $-\infty$ a $+\infty$ mentre affinché il risultato sia visualizzabile è stato necessario trasformare i valori delle matrici da 0 a $+\infty$. Da un punto di vista pratico viene preso il valore più basso della matrice ottenuta e portato a 0, aumentando poi tutti gli altri numeri della egual differenza. Tale piccolo accorgimento ha reso possibile la visualizzazione della matrice utilizzando un solo colore più o meno vivo a seconda dell'omologia delle proteine con i profili HMM e utilizzando il nero lì dove non v'era omologia.

Per le analisi trattate in questo lavoro il software utilizzato per la visualizzazione delle matrici è stato un programma scritto in java (http://www.java.com) chiamato Genesis versione 1.7 (92).

2.7 Creazione della mappa R di pomodoro

Tramite il predittore MATRIX è stato possibile esplorare il genoma di pomodoro sequenziato per la ricerca di putativi geni R. I geni predetti sono stati raccolti in un database SQL e catalogati insieme a tutte le informazioni disponibili sulle loro caratteristiche e sulla loro localizzazione. Grazie alla raccolta di queste informazioni è stato possibile scrivere un programma che, collegandosi al db, trasformasse le informazioni presenti in dati di grafica vettoriale SVG (93). Attraverso questo sistema è stato disegnato ogni singolo contig e su ogni contig sono stati posizionati i geni putativi per la funzione di resistenza e tutti i marcatori COSII. I marcatori sono stati raccolti dalla collezione pubblica del SOL, "Solanaceae Genomic Network", e localizzati creando un database BLAST in locale contenente tutti i contig del genoma del pomodoro ed estrapolando i dati di omologia degli stessi contro i marcatori.
In questo modo è stato possibile ottenere una mappa fisica del genoma di pomodoro nella quale sono disegnati tutti i contig in grandezza naturale (un pixel corrisponde ad un nucleotide) e sui quali è possibile osservare la posizione dei putativi geni R, i marcatori associati e la loro classe di resistenza.

2.8 Analisi filogenetiche

Tutte le filogenesi prodotte in questo lavoro sono state prodotte tramite il software PHYML (94) con impostazioni predefinite, bootstrap 100 ed utilizzando come matrice di sostituzione amminoacidica la LG (95). La visualizzazione dei risultati è stata prodotta tramite il software FigTree versione 3.1.2 e le annotazioni dei geni sono state composte automaticamente tramite uno *script* in perl creato appositamente per l'annotazione automatica delle informazioni di sequenza, che ha permesso di dettagliare e discriminare le sequenze presenti negli alberi prodotti.

2.9 Ottenimento dei dati presentati

Tutti dati presentati sono stati ottenuti utilizzando i sistemi di predizione DRAGO e MATRIX su diversi set di dati. Le informazioni sono state estrapolate:

- manualmente
- utilizzando le ricerche del PRGdb
- effettuando ricerche specifiche all'interno di database privati e pubblici

L'interfacciamento ai db è avvenuto tramite riga di comando o tramite il software MAMP (http://www.mamp.info).

Tutte le analisi di sequenza sono state effettuate utilizzando i software:

- BLAST per la ricerca tramite omologia di sequenza
- CLustalW, clustalX, MUSCLE, M-Coffee per gli allineamenti (96, 97)
- Interproscan per l'analisi delle sequenze proteiche
- Phobius per l'identificazione dei domini trans membrana (98)
- Geneious per la gestione dei dati e le operazioni di modifica delle sequenze (http://www.geneious.com)
- PHYML per le filogenesi
- FigTree per la visualizzazione degli alberi filogenetici (http://tree.bio.ed.ac.uk)
- Artemis per l'analisi dei BAC e dei Contig (99)

Per tutti i software è stata utilizzata l'ultima versione disponibile a Novembre 2009.

2.10 Analisi molecolari

2.10.1 Materiale vegetale

Gli esperimenti molecolari sono stati condotti su sei genotipi differenti afferenti al genere *Solanum*. In particolare è stata utilizzata la varietà coltivata Heinz (*Solanum lycopersicum)*, 4 specie selvatiche: *Solanum pimpinnellifolium*, *Solanum pennellii, Solanum Hirsutum, Solanum peruvianum* e la linea di introgressione 7-3 data dall'incrocio tra *Solanum lycopersicum* e *Solanum pennellii.*

La semina è avvenuta utilizzando per ogni varietà 0.5 grammi di semi che sono stati sterilizzati in falcon da 50 ml e lavati una prima volta con etanolo al 70% per 5 minuti, ed una seconda volta con etanolo al 70% per 1 minuto. Eliminato l'etanolo, in ambiente controllato i semi sono stati sterilizzati con una soluzione di ipoclorito di sodio al 10% (cloro attivo al 4%) e SDS al 0.1% per 10 minuti, ripetendo il passaggio due volte. Per eliminare i residui della soluzione di sterilizzazione i semi sono stati lavati 5 volte con acqua distillata sterile, ed infine distribuiti in barattoli Magenta sterili contenenti 50 ml di substrato solido MS30. La germinazione e la crescita del materiale vegetale è avvenuta in camere di crescita alla temperatura costante di 24 C°, in fotoperiodo 16/8 h luce/buio ed intensità luminosa 100 µE m^{-2} s^{-1}.

2.10.1 Estrazione acidi nucleici e retro trascrizione

Il DNA genomico è stato isolato e purificato attraverso il sistema DNasy Plant mini Kit (Quiagen) o attraverso la procedura CTAB (100). L'RNA è stato estratto dalle rispettive varietà di pomodoro coltivato e selvatico con l'ausilio del kit Rneasy mini (Quiagen).

La sintesi del cDNA a partire dal RNA estratto è avvenuta attraverso l'ausilio del kit "First-Strand cDNA Synthesis Super*Script* III".

All'RNA totale sono stai aggiunti:

Oligo dt	50 µm
Primer	2 pmol

La soluzione è stata incubata a 65° C per 5 min e velocemente trasferita in ghiaccio per 1 min. Sono stati poi aggiunti

Buffer First-Strand	5x
DTT	0.1 M
Super*Script* III RT	1 µl (200 unità/µl)

La soluzione è stata incubata a 50° C per 60 min ed a 70° C per 10 min

2.10.2 Quantificazione del DNA genomico e visualizzazione del RNA su gel di agarosio

Il DNA genomico è stato separato e quantizzato mediante elettroforesi su gel di agarosio al 0,8-1% in TAE (10 mM Tris-Cl pH 7.8, 5 mM Acetato di Sodio, 0.5 mM EDTA, pH 7.8 con Acido Acetico). Al gel di agarosio liquido sono stati aggiunti 0,03 µl/ml di Bromuro di etidio. I campioni da visualizzare su gel sono stati preparati aggiungendo a 2 µl di DNA genomico 2 µl di colorante 1 (4M Urea, 50% Saccarosio, 50 mM EDTA pH 8.0, 0.1% Bromofenolo Blu) e 8 µl di acqua. Gli standard di riferimento sono stati ottenuti utilizzando quantità note di DNA derivate dal Fago λ. Il gel é stato visualizzato e fotografato sotto lampada UV mediante apparato fotografico GelDOC (Biorad). Inoltre il DNA estratto è stato quantizzato anche tramite spettrofotometro.

I frammenti di RNA sono stati separati mediante elettroforesi su gel di agarosio al 1 % in TBE (Tris 0.89M, EDTA-Na_2 0.02M, Acido Borico 0.89 M pH 8.3). Al gel di agarosio liquido sono stati aggiunti 0,03 µl /ml di Bromuro di Etidio. Tutte le soluzioni sono state preparate con acqua DEPC e gli strumenti utilizzati sono stati sterilizzati con NaOH e lavati con acqua DEPC per eliminare qualsiasi traccia di RNA. Il gel é stato fotografato mediante apparato GelDOC generante luce UV.

2.10.3 Analisi di amplificazione tramite PCR ed RT-PCR

La tecnica PCR è stata utilizzata per l'amplificazione di frammenti di lunghezza variabile e per esperimenti di retrotrascrizione. Per le reazioni di amplificazione sono stati utilizzate polimerasi della Invitrogen, Promega e Takara seguendo le istruzioni dei rispettivi protocolli, e sono state condotte con termociclizzatori Biometra ed Eppendorf.
I frammenti di DNA sono stati separati mediante elettroforesi su gel di agarosio al 1.5 % in TAE (10 mM Tris-Cl pH 7.8, 5 mM Acetato di Sodio, 0.5 mM EDTA, pH 7.8 con Acido Acetico). Al gel di agarosio liquido sono stati aggiunti 0,03 µl/ml di Bromuro di etidio. I campioni da visualizzare su gel sono stati preparati aggiungendo 25 µl di reazione PCR e 4 µl di colorante. Il gel é stato fotografato mediante apparato GelDOC generante luce UV.

2.10.4 Clonaggio del putativo gene R I3

Per il clonaggio del gene 4701 predetto da MATRIX sul genoma di pomodoro è stata effettuata una prima amplificazione utilizzando la tecnica della long-PCR per l'isolamento della regione genica d'interesse ed in seguito è stata effettuata un nested-PCR per l'amplificazione del gene a partire dal suo ATG fino ad alcuni nucleotidi dopo il terminatore. Tutte le amplificazioni sono avvenute utilizzando polimerasi ad alta fedeltà con attività esonucleasica. Il frammento ottenuto è stato amplificato tramite l'utilizzo di primer specifici della lunghezza di 50bp per l'attacco degli adattatori di clonaggio. La tecnica di clonaggio utilizzata è stata quella gateway (101) con l'inserzione del frammento d'interesse nel vettore pDONR Zeo.

Risultati

3.1 Ricerca bibliografica e catalogazione dei geni R

È stata condotta un'analisi bibliografica per collezionare tutti i geni di resistenza isolati in pianta e tutte le informazioni ad essi correlate. In tutto sono stati individuati 73 geni R funzionali. Tali geni conferiscono resistenza a 33 patogeni differenti e sono stati isolati rispettivamente da 22 specie vegetali non solo della famiglia delle *Solanaceae* (33 geni) (25, 67), ma anche da *Arabidopsis thaliana* (21 R-geni) (68), *Oryza sativa* (riso, 4 R-geni) (69, 70), *Phaseulus vulgaris* (fogiolo, 1 R-gene) (71), *Glicine max* (soia, 2 R-geni) (72), *Zea mais* (mais, 2 R-geni) (73), *Hordeum vulgare* (orzo, 3 R-geni) (74, 75, 76), *Cucumis melo* (mellone, 2 R-geni) (77), *Lactuca sativa* (lattuga, 1 R-gene) (78), *Beta vulgaris* (bietola, 1 R-gene) (79), *Linum usatissimum* (lino, 3 R-geni) (80, 81, 82). Tale ricerca ha permesso di creare una raccolta affidabile di informazioni e di conoscere il numero reale di geni R isolati fino ad oggi. Grazie a questa raccolta meticolosa di geni R funzionali, è stato possibile effettuare un'accurata catalogazione delle sequenze, aggiungendo importanti informazioni, provenienti da fonti diverse, in modo da creare una solida struttura informativa. Per ogni gene è stata creata una scheda con le informazioni sulle sequenze nucleotidiche e proteiche, sulla classe di appartenenza, sulla specie vegetale da cui il gene proviene, sui marcatori molecolari associati, sul gene di avirulenza riconosciuto dal gene, sulla malattia bloccata, sul patogeno a cui resiste e sulla malattia ad essa correlata. A tutte queste informazioni sono poi stati aggiunti i dati tassonomici sulle specie e i dati presenti in database pubblici come NCBI (figura 3).

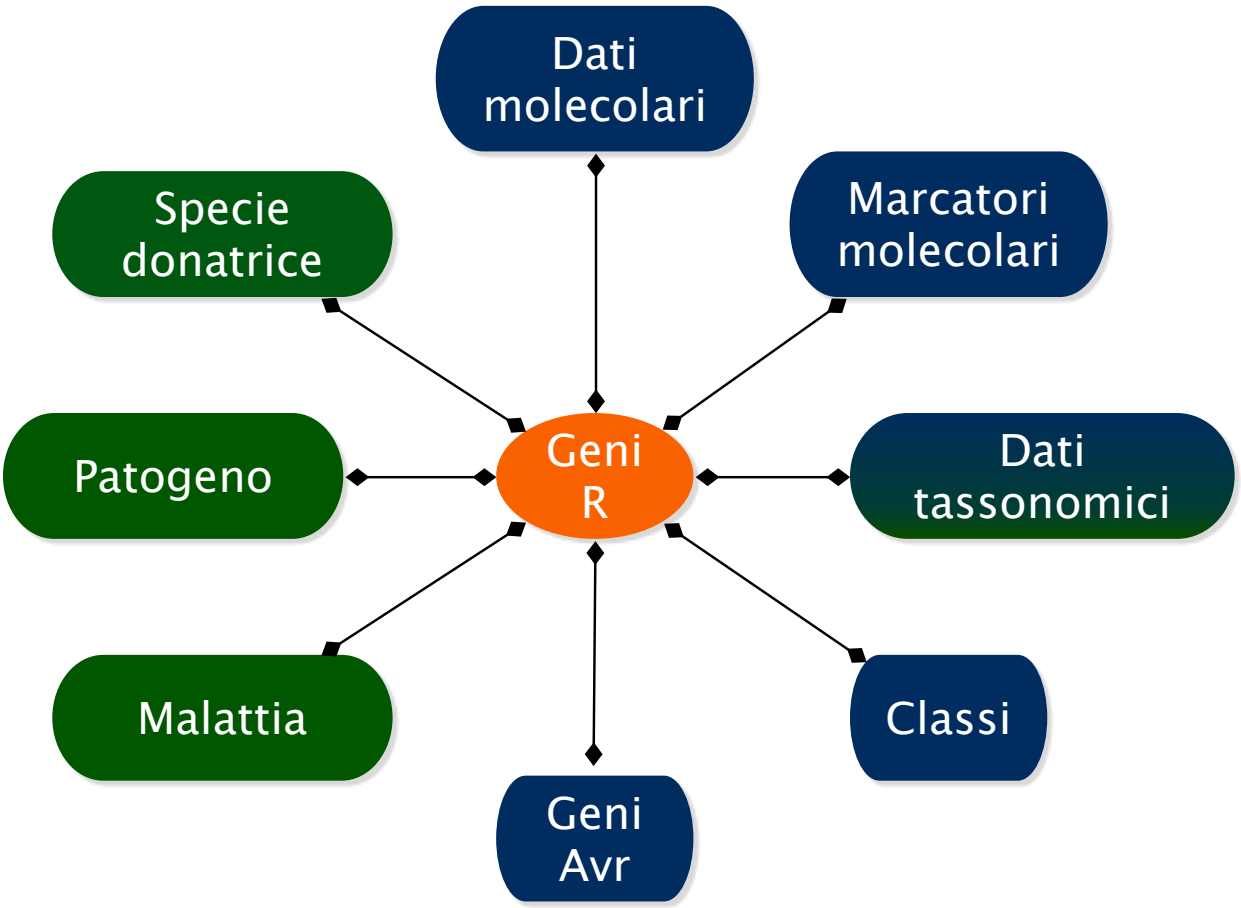


Figura 3. Integrazione dei dati correlati ai geni R provenienti da fonti diverse

Questi geni sono stati catalogati come "geni di referenza" in quanto hanno, a differenza delle altre sequenze presenti nei database pubblici, un supporto scientifico e bibliografico a conferma della loro funzionalità e possono essere facilmente distinti dai loro omologhi non funzionali. In tabella 1 è possibile osservare le informazioni relative ai geni R di referenza.

Nome del gene R	Organismo donatore	Malattia correlata	Patogeno
Asc1	*Solanum lycopersicum*	Alternaria Stem Canker	*Alternaria alternata*
At1	*Cucumis melo*	Downy mildew mellon	*Pseudoperonospora cubensis*
At2	*Cucumis melo*	Downy mildew mellon	*Pseudoperonospora cubensis*
Bs2	*Capsicum chacoense*	Bacterial Spot	*Xanthomonas campestris pv. vesicatoria str. 85-10*
Bs3	*Capsicum annuum*	Bacterial Spot	*Xanthomonas campestris pv. vesicatoria str. 85-10*
Bs3-E	*Capsicum annuum*	Bacterial Spot	*Xanthomonas campestris pv. vesicatoria str. 85-10*
Bs4	*Solanum lycopersicum*	Bacterial Spot	*Xanthomonas campestris*
Cf2	*Solanum pimpinellifolium*	Leaf Mould	*Passalora fulva*
Cf4	*Solanum habrochaites*	Leaf Mould	*Passalora fulva*
Cf4A	*Solanum habrochaites*	Leaf Mould	*Passalora fulva*
Cf5	*Solanum lycopersicum var. cerasiforme*	Leaf Mould	*Passalora fulva*
Cf9	*Solanum pimpinellifolium*	Leaf Mould	*Passalora fulva*
Cf9B	*Solanum pimpinellifolium*	Leaf Mould	*Passalora fulva*
Dm-3	*Lactica sativa*	Downy milldew	*Bremia lactucae*
EFR	*Arabidopsis thaliana*	Eliciting Bacteria	*Bacteria with flagellum*
ER-Erecta	*Arabidopsis thaliana*	Bacterial Wilt (Arabidopsis)	*Ralstonia solanacearum*
FLS2	*Arabidopsis thaliana*	Eliciting Bacteria	*Bacteria with flagellum*
Gpa2	*Solanum tuberosum*	Yellow potato cyst nematode	*Globodera*
Gro1.4	*Solanum tuberosum*	Late Blight Potato	*Phytophthora infestans*
Hero	*Solanum lycopersicum*	Yellow potato cyst nematode	*Globodera*
Hm1	*Zea mays*	Leaf Spot	*Bipolaris zeicola*
Hm2	*Zea mays*	Leaf Spot	*Bipolaris zeicola*
HRT	*Arabidopsis thaliana*	Turnip crinkle virus	*Turnip crinkle virus*
Hs1	*Beta procumbens*	Beet cyst nematode	*Heterodera schachtii*

I2	*Solanum lycopersicum*	Fusarium Wilt	*Fusarium oxysporum*
L6	*Linum usitatissimum*	Flax rust	*Melampsora lini*
LeEIX1	*Solanum lycopersicum*	Eliciting Fungus	*Fungal ethylene-inducing xylanase*
LeEIX2	*Solanum lycopersicum*	Eliciting Fungus	*Fungal ethylene-inducing xylanase*
M	*Linum usitatissimum*	Flax rust	*Melampsora lini*
Mi1.2	*Solanum lycopersicum*	Root-knot nematode	*Meloidogyne, Paratrichodorus minor*
MLA10	*Hordeum vulgare*	Powdery mildew (barley)	*Blumeria graminis*
Mlo	*Hordeum vulgare*	Powdery mildew (barley)	*Blumeria graminis*
N	*Nicotiana glutinosa*	Tobacco Mosaic Virus	*Tobacco mosaic virus*
P2	*Linum usitatissimum*	Flax rust	*Melampsora lini*
PEPR1	*Arabidopsis thaliana*	Damping Off	*Pythium*
PGIP	*Phaseolus vulgaris*	Eliciting Fungus	*Fungus producing polygalacturonases*
Pi33	*Oryza sativa*	Rice blast disease	*Magnaporthe grisea*
Pi-ta	*Oryza sativa Japonica Group*	Rice blast disease	*Magnaporthe grisea*
Prf	*Solanum pimpinellifolium*	Bacterial Speck	*Pseudomonas syringae*
Pto	*Solanum pimpinellifolium*	Bacterial Speck	*Pseudomonas syringae*
R1	*Solanum demissum*	Late Blight Tomato	*Phytophthora infestans*
R3a	*Solanum tuberosum*	Late Blight Tomato	*Phytophthora infestans*
RCY1	*Arabidopsis thaliana*	Cucumber Mosaic Virus	*Cucumber mosaic virus*
RFO1	*Arabidopsis thaliana*	Fusarium Wilt	*Fusarium oxysporum*
Rmd-c	*Glycine max*	Powdery mildew	*Microsphaera sparsa*
RPG1	*Hordeum vulgare*	Stem Rust	*Puccinia Graminis*
Rpi-blb1	*Solanum bulbocastanum*	Late Blight Tomato	*Phytophthora infestans*
Rpi-blb2	*Solanum bulbocastanum*	Late Blight Tomato	*Phytophthora infestans*
RPM1	*Arabidopsis thaliana*	Bacterial Blight	*Pseudomonas syringae*
RPP13nd	*Arabidopsis thaliana*	Downy mildew	*Hyaloperonospora parasitica*

RPP4	*Arabidopsis thaliana*	Downy milldew	*Peronospora parasitica*
RPP5	*Arabidopsis thaliana*	Downy mildew	*Hyaloperonospora parasitica*
RPP8	*Arabidopsis thaliana*	Downy mildew	*Hyaloperonospora parasitica*
Rps1-k-1	*Glycine max*	Phytophthora root	*Phytophthora sojae*
Rps1-k-2	*Glycine max*	Phytophthora root	*Phytophthora sojae*
Rps2	*Arabidopsis thaliana*	Bacterial Blight	*Pseudomonas syringae*
Rps4	*Arabidopsis thaliana*	Bacterial Blight	*Pseudomonas syringae*
RPS5	*Arabidopsis thaliana*	Bacterial Blight	*Pseudomonas syringae*
RPW8.1	*Arabidopsis thaliana*	Powdery mildew	*Golovinomyces cichoracearum*
RPW8.2	*Arabidopsis thaliana*	Powdery mildew	*Golovinomyces cichoracearum*
RRS1	*Arabidopsis thaliana*	Bacterial wilt	*Ralstonia solanacearum*
RTM1	*Arabidopsis thaliana*	Synergistic disease syndromes	*Tobacco etch virus*
RTM2	*Arabidopsis thaliana*	Synergistic disease syndromes	*Tobacco etch virus*
Rx	*Solanum tuberosum*	Latent Mosaic	*Potato virus X*
Rx2	*Solanum acaule*	Latent Mosaic	*Potato virus X*
RY1	*Solanum tuberosum subsp andigena*	Potato virus Y	*Potato virus Y*
Sw5	*Solanum lycopersicum*	Tomato Spotted Wilt	*Tomato spotted wilt virus*
Tm2	*Solanum lycopersicum*	Tobacco Mosaic Virus	*Tobacco mosaic virus*
Tm2a	*Solanum lycopersicum*	Tobacco Mosaic Virus	*Tobacco mosaic virus*
Ve1	*Solanum lycopersicum*	Verticillium wilt Potato	*Verticillium*
Ve2	*Solanum lycopersicum*	Verticillium wilt Potato	*Verticillium*
Xa1	*Oryza sativa*	Bacterial blight	*Xanthomonas oryzae*
Xa21	*Oryza sativa Indica Group*	Bacterial Blight	*Xanthomonas oryzae*

Tabella 1. Catalogazione dei geni R caratterizzati ed associazione con i di resistenza

Attraverso l'analisi della struttura proteica ogni gene è stato caratterizzato anche per i propri domini conservati. L'analisi condotta ha permesso di dividere le sequenze in cinque classi diverse, confermando i dati bibliografici.

- CNL 32
- TNL 9
- RLP 12
- RLK 7
- Other 13 geni

Se per le 4 classi contenenti domini proteici molto conservati la classificazione risulta immediata, per le 13 sequenze che esplicano la loro azione mediante sistemi diversi, la classificazione risulta difficile se non impossibile. Per questo motivo i 13 geni, che produco proteine di resistenza funzionali utlizzando altri meccanismi oltre a quelli "classici", sono stati inseriti in una classe virtuale definita "other". I geni all'interno di tale classe non hanno caratteristiche comuni tra di essi e pertanto devono essere studiati e caratterizzati singolarmente.

Al fine di dare una visione completa dei geni R, è stata effettuata un'analisi filogenetica utile per la classificazione delle sequenze e per effettuare gli studi di omologia e speciazione. A cause della loro non conformità i geni della classe other non sono stati sottoposti ad analisi filogenetica (figura 4). Da tale analisi è facilmente osservabile come i geni si siano raggruppati in base alle loro caratteristiche genetiche e tassonomiche in rami distinti e robusti (numero di "bootstrap" alto). Il ramo più vicino alla radice è quello dei geni CNL. In seguito, l'albero si divide in due rami: il primo che porta alla formazione della classe TNL e il secondo che, dividendosi a sua volta, porta alla formazione delle classi RLP ed RLK. I geni della classe TNL sono raggruppati in un unico insieme e all'interno del loro ramo è possibile osservare un'ulteriore divisione: ad un primo livello, tra geni provenienti dalla famiglia delle Solanaceae ed *Arabidopsis thaliana* e ad un secondo livello, tra geni provenienti da specie dicotiledoni e monocotiledoni. In blu ed azzurro sono rappresentati i geni affferenti rispettivamente alle classi RLK e RLP. Le caratteristiche della loro suddivisione, anche in questo caso, rappresentano le loro caratteristiche molecolari. Infatti entrambe le classi contengono geni che codificano per proteine di membrana molto simili tra loro e che differiscono per le caratteristiche del loro dominio chinasico. La classe CNL, la cui caratteristica più importante è un'elevata conservazione delle sequenze, confermata dalla loro posizione in prossimità della radice

dell'albero, è stata divisa in due rami pricipali le cui differenze però non intaccano le strutture secondarie e terziarie delle proteine, ma solo la composizione della sequenza, lasciando perceperire che, nonostante ci possa essere variabilità sulla struttura primaria, resta comunque una forte conservazione dei domini tipici dei geni R. I geni *RPG1, Pto e RFO*, tutti e tre appartenenti alla RLK, a causa della loro forte variabilità rispetto alla struttura tipica degli RLK presentano un braccio molto lungo ed un bootstrap basso che non gli permettono di essere adeguatamente classificati.

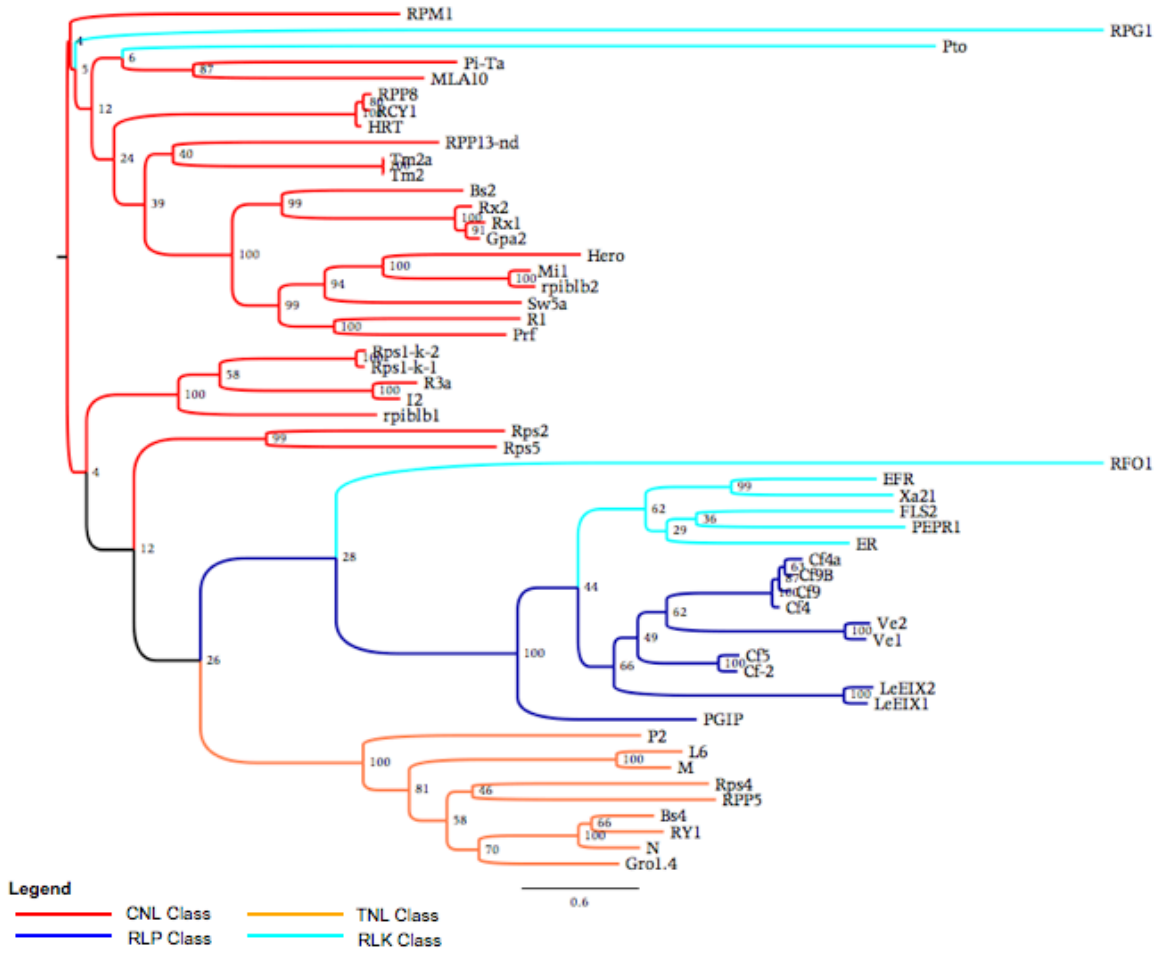


Figura 4. Filogenesi del gruppo dei geni R di referenza

3.2 Automazione del processo di catalogazione delle sequenze provenienti da database pubblici

Per avere una visione esaustiva dei geni R, insieme al set composto dalle 73 sequenze di referenza, ne è stato creato un altro composto da tutte le sequenze omologhe ai geni R presenti nei database pubblici. Nonostante per tali sequenze non siano stati effettuati esperimenti di complementazione per verificarne la reale funzionalità, esse sono molto importanti ai fini scientifici per gli studi di omologia, evoluzione di genomica funzionale e strutturale. Per tale motivo è stato creato uno *script ad hoc* per ottenere automaticamente tutte le sequenze correlate ai geni di resistenza presenti in rete.

Le sequenze ottenute tramite questo processo automatizzato sono state raccolte ed analizzate per l'eliminazione di sequenze ridondanti, falsi positivi, sequenze contaminate da vettori o erroneamente catalogate come putativi geni R. Inoltre, le sequenze così prodotte sono state sottoposte ad uno studio a livello proteico per la caratterizzazione dei loro domini e catalogate in base alla classe genica di appartenenza. Poiché le sequenze all'interno dei db pubblici aumentano continuamente, tale sistema è stato strutturato in modo da raccogliere ad intervalli regolari le nuove sequenze depositate in banca dati. Attualmente il numero di sequenze collezionate tramite questo approccio è pari a 6308 (figura 5).

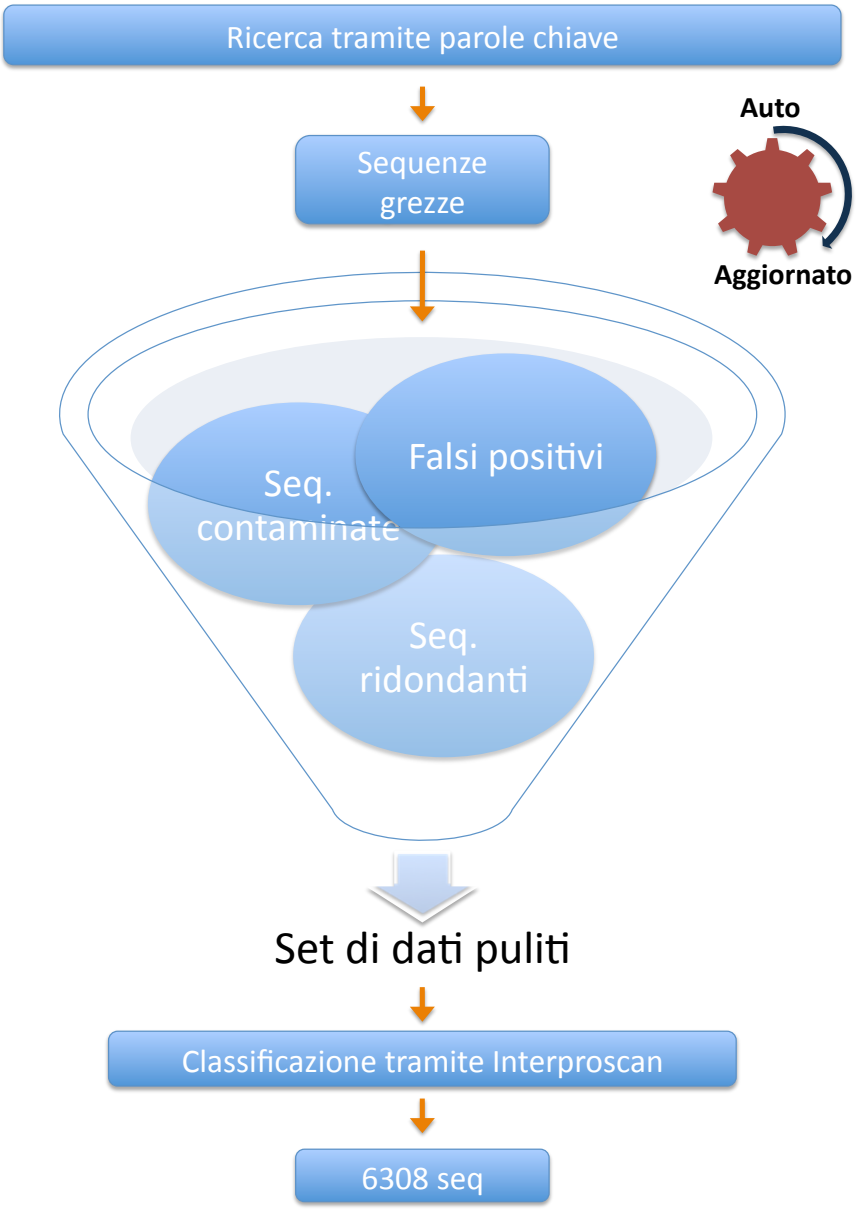


Figura 5. Approccio per la raccolta automatica dei dati relativi alle resistenze vegetali

3.3 PRGdb: il Plant Resistance Gene database

Al fine di fornire un catalogo completo dei dati raccolti e di creare una risorsa unica, accessibile e duratura per i geni di resistenza, è stato progettato un database specifico: il **P**lant **R**esistance **G**ene (www.prgdb.org). Il PRG è, ad oggi, la prima risorsa specifica per lo studio e la catalogazione dei geni di resistenza. Al suo interno sono raccolti 73 geni R di referenza e 6308 sequenze correlate e provenienti da database pubblici. Le informazioni sono state collegate tra di loro tramite un sistema relazionale secondo lo schema riportato in figura 1. I dati, inseriti in modo semi-automatizzato, sono stati controllati manualmente al fine di produrre un risorsa accurata e dettagliata.

3.3.1 I Dati del PRGdb

Il cuore del database sono i 73 geni R catalogati in modo manuale ed accurato. Per ogni sequenza sono state inserite e convalidate le seguenti informazioni:

- Nome del gene
- Grafico per informazioni di sequenza e per i domini conservati presenti
- Sinonimi del gene
- Descrizione della sequenza
- Classe di appartenenza
- Cromosoma di appartenenza
- Link a NCBI nucleotide
- Link a NCBI protein
- Specie da cui è stato isolato il gene
- Specie originaria da cui deriva il gene
- Link alle informazioni tassonomiche delle due specie
- Marcatore molecolare associato
- Referenza bibliografica del marcatore molecolare
- Categoria del gene
- Nome del prodotto genico
- Referenza bibliografica del gene
- Malattia bloccata dal gene
- Descrizione della malattia e dei suoi sintomi
- Gene di virulenza riconosciuto
- Link alle informazioni tassonomiche sul patogeno
- Sequenza CDS
- Sequenza proteica
- Sequenze proteiche dei domini conservati
- Link ai database dei domini conservati Pfam e Panther

Per i dati ricavati da NCBI, nonostante la struttura informativa sia la stessa, le informazioni sui marcatori molecolari, sulle malattie e sui geni di avirulenza non possono essere presenti, in quanto per le sequenze in questione non sono stati effettuati esperimenti di complementazione che ne confermino la funzionalità. Attualmente le sequenze sono correlate a 115 patogeni, 192 specie vegetali e 21 geni avr ed il db è strutturato in modo tale da essere facilmente espandibile per contenere nuove informazioni. Interessante è l'integrazione automatica tra i dati dei diversi set attraverso cui è possibile visualizzare i geni R di referenza ed i propri omologhi provenienti da altre specie o varietà.

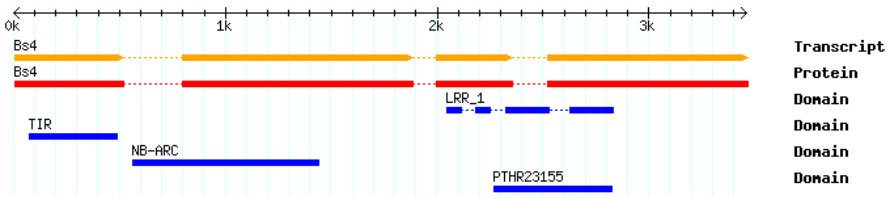
3.3.2 Interfaccia grafica

Il PRGdb, dotato di una semplice ed intuitiva interfaccia grafica, è accessibile attraverso l'indirizzo www.prgdb.org. Attraverso i link ben visibili, i colori e le didascalie concise ma esaustive, lo stile grafico del PRG permette a tutti gli utilizzatori una facile navigazione ed un facile reperimento dei dati. La struttura grafica creata permette di accedere velocemente alle ricerche semplici o avanzate ed i dati di output sono visualizzati in maniera esaustiva. Le ricerche sono posizionate in home page, sono ben visibili ed i tool aggiuntivi possono essere raggiunti tramite il menù d'integrazione posizionato nella parte superiore dell'home page (Figura 7).

Figura 6. PRGdb "Home page"

Le informazioni di sequenza sono visualizzate attraverso un format comune in tutto il database, in modo da rendere la lettura delle informazioni più rapida e ,per ogni gene inserito, è stato creato un apposito "*script*" in grado di trasformare le informazioni di sequenza (cds, sequenza proteica, domini conservati) in un'accattivante ed informativa immagine per la visualizzazione degli elementi genici (figura 7).

Gene Bs4 (Reference)



Synonyms	-
Description	Lycopersicon esculentum bacterial spot disease resistance protein 4 (Bs4) gene, Bs4-MM allele, complete cds.
Class	**TNL**
Chromosome	5
Entrez Nucleotide	**AY438027**
Entrez Protein	**AAR21295**
Species	 **Solanum lycopersicum**
Donor Species	 **Solanum pennellii**
Markers	**SGN-M3925**
Markers Reference	SOL Genomic Network
Category	Reference R-Genes, manually curated
Product	bacterial spot disease resistance protein 4
References	Schornack,S., Ballvora,A., Gurlebeck,D., Peart,J., Ganal,M., Baker,B., Bonas,U. and Lahaye,T., The tomato resistance protein Bs4 is a predicted non-nuclear TIR-NB-LRR protein that mediates defense responses to severely truncated derivatives of AvrBs4 and overexpressed AvrBs3, Plant J. 37 (1), 46-60 (2004) - **PUBMED:14675431**

Disease	**Pathogen**	**Avirulance Gene**
Bacterial Spot	 **Xanthomonas campestris**	**AvrBs4**

Bacterial spot develops on seedlings and mature plants. On seedlings, infections may cause severe defoliation. On older plants, infections occur primarily on older leaves and appear as water-soaked areas. Leaf spots turn from yellow or light green to black or dark brown. Older spots are black, slightly raised, superficial and measure up to 0.3 inch (7.5 mm) in diameter. Larger leaf blotches may also occur, especially on the margins of leaves. Symptoms on immature fruit are at

Figura 7. Esempio di visualizzazione di una pagina informativa su un gene R

3.3.3 Consultazione e strumenti di ricerca

Il PRGdb offre numerose possibilità di consultazione, utilizzando le sue molteplici modalità di ricerca. Le ricerche posso essere utilizzate in modo singolo o incrociato in base ai bisogni dell'utente e possono essere divise in base al set sul quale si voglio ricercare le informazioni.

3.3.3.1 Ricerca sul set di referenza

Per il più dettagliato set di referenza è possibile effettuare ricerche singole o incrociate scegliendo sei diverse opzioni:

- Nome del gene
- Specie vegetale
- Patogeno
- Malattia
- Gene di virulenza
- Classe

Tramite questo semplice sistema è possibile filtrare le informazioni a seconda dei dati che si desidera consultare o che si conoscono, così da ottenere le informazioni d'interesse in modo rapido ed efficace. Inoltre, se si desidera visualizzare il set dei geni di referenza nella sua integrità, basta effettuare una ricerca lasciando in tutti i campi la dicitura "all". Nel caso si voglia conoscere la provenienza dei dati mostrati, un link all'interno del quadro di ricerca condurrà l'utente nella sezione delle referenze bibliografiche (figura 8).

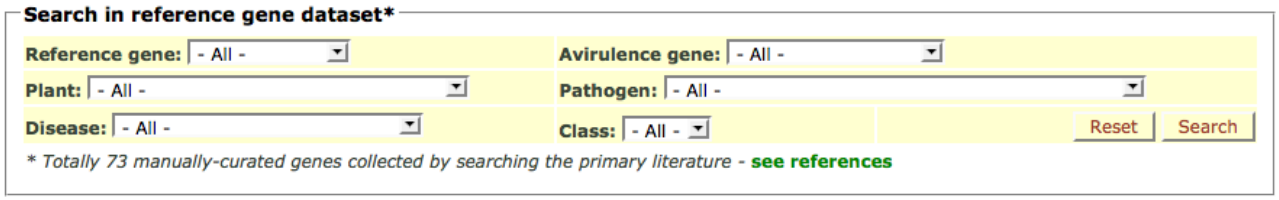


Figura 8. Finestra di ricerca per il set dei geni R funzionali

3.3.3.2 Ricerca sull'intero sistema

La ricerca più completa del PRG è quella che può essere effettuata sull'intero set di dati. Questo tipo di ricerca, a differenza della precedente basata sull'uso di parole chiave, sfrutta un approccio misto basato sia sull'uso di parole chiave, filtrando le informazioni tramite pianta e/o patogeno, sia utilizzando le informazioni prodotte da Interproscan sui domini conservati presenti all'interno delle sequenze. Tramite questa tipologia di ricerca avanzata l'utente potrà scegliere in primo luogo su quale set di dati effettuare la ricerca, poi quali e quanti domini selezionare ed infine se filtrare le informazioni per

tipologia di pianta o patogeno. Grazie a questo sistema le combinazioni di ricerca sono molteplici e possono rispondere in modo esaustivo anche alle richieste molto specifiche di un singolo utente (figura 9). Ad esempio, si potrebbe richiedere al database di ricercare, nel set di dati raccolti da NCBI, tutte le sequenze che contengono il dominio LRR associato ai domini TIR e NBS, per la specie *Solanum lycopersicum*. Lasciando invariati tutti i campi, la ricerca produrrà un foglio con tutte le sequenze disponibili nel PRG.

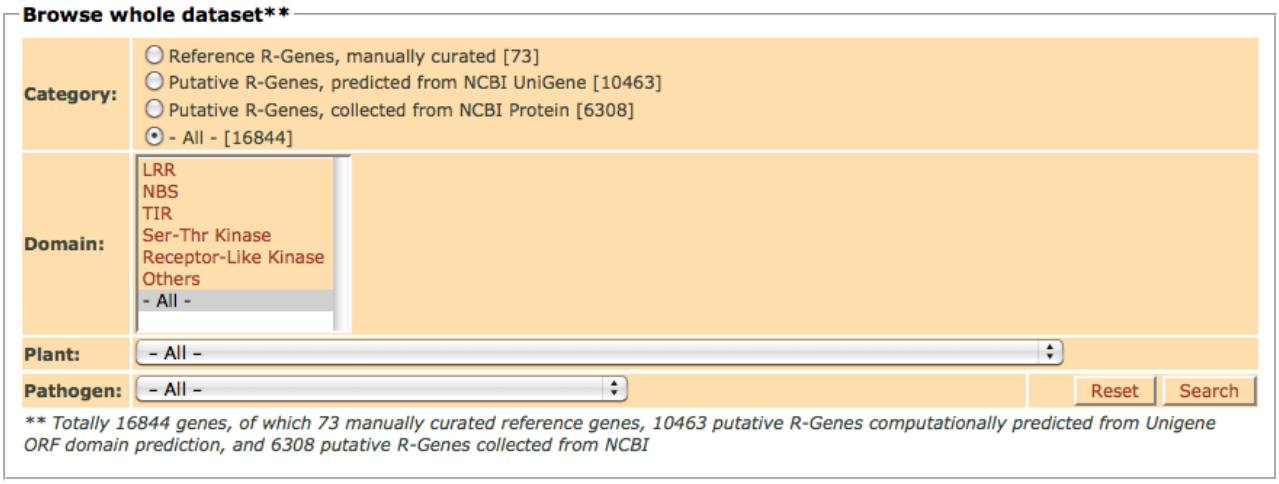


Figura 9. Finestra di ricerca avanzata sui dati presenti in PRGdb

3.3.3.3. Ricerca filtrata per organismi

Per chi lavora nel campo delle resistenze vegetali è spesso di fondamentale importanza capire qual è il profilo di resistenze di un organismo vegetale. Allo stesso modo, è importante sapere velocemente cosa si conosce di un patogeno vegetale, se è riconosciuto da organismi vegetali o se i suoi geni di avirulenza sono riconosciuti da specifici geni R. A questo scopo sono state create due schermate dove sono collezionati tutti gli organismi vegetali che contengono geni R putativi o funzionali e dove sono collezionati tutti i patogeni di cui si abbiano informazioni di interazione con i geni R. Le sezioni nominate rispettivamente “Plant search” e “Pathogen search” sono accessibili tramite “home page” e sono caratterizzate da fotografie che ritraggono tutti gli organismi vegetali e patogeni presenti nel database (figura 10).

Cliccando sui i singoli nomi degli organismi o sulle loro rappresentazioni, il PRG ricercherà automaticamente tutti i geni presenti in quella specie, permettendo all’utilizzatore di avere velocemente una visione chiara delle caratteristiche di ogni singola specie.

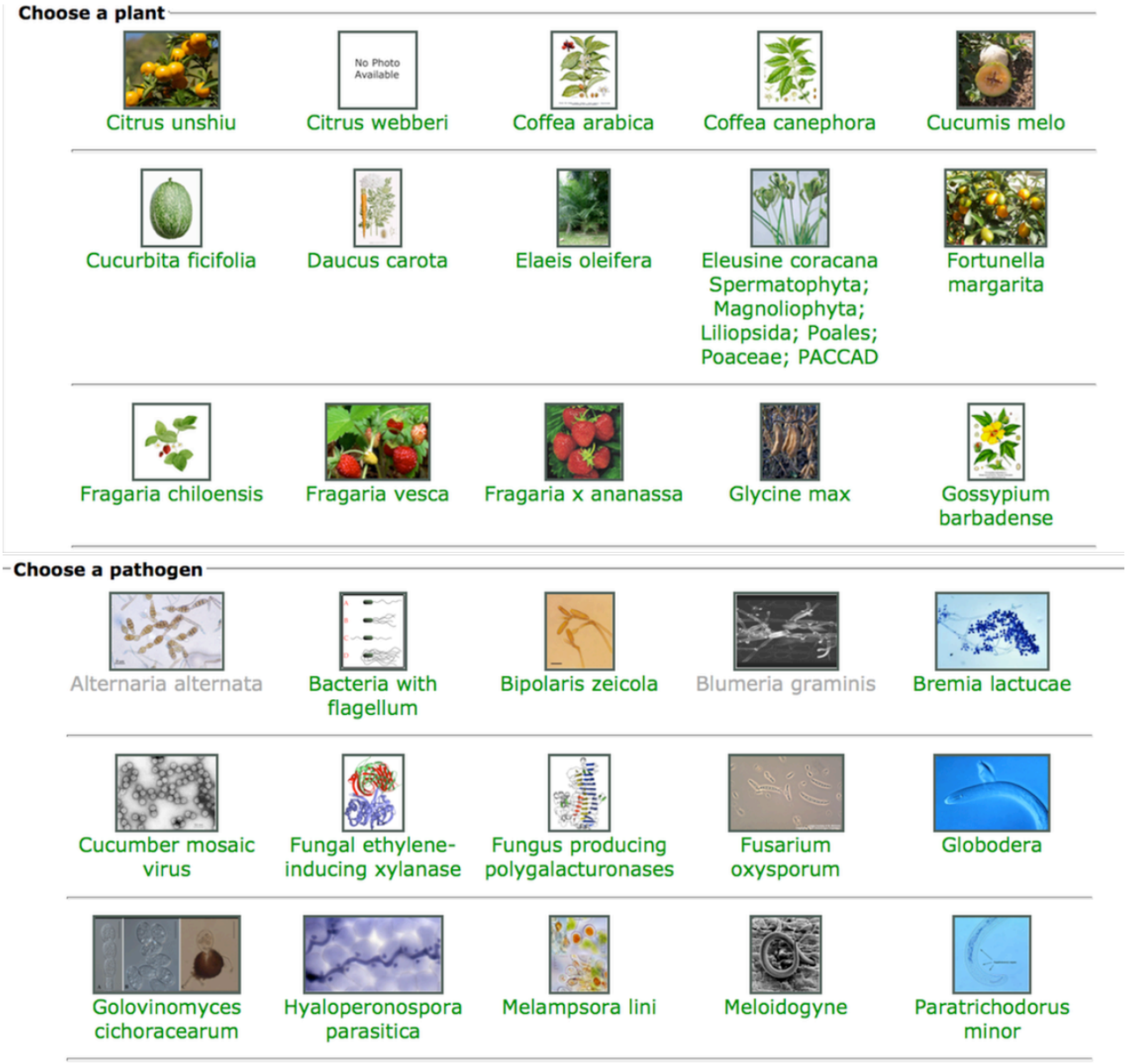


Figura 10. Finestra di ricerca per specie

3.3.3.4 Ricerca attraverso il BLAST

Di fondamentale importanza è anche la possibilità di ricercare informazioni di interesse tramite l'utilizzo dell'algoritmo di omologia BLAST. Attraverso l'apposita sezione, accessibile tramite "home page", cliccando sull'indicazione "BLAST search" è possibile entrare nella sezione del database deputata alla ricerca delle sequenze tramite omologia (figura 11). L'utilizzatore può quindi inserire una o più sequenze di cui desidera conoscere l'omologia con i geni presenti nel PRG ed il tool automaticamente produrrà una schermata nella quale saranno presenti tutti i geni omologhi alla sequenza inserita (figura 12). Nella schermata principale del BLAST è possibile scegliere il livello di stringenza della ricerca, modificando il parametro "E-value", e decidere se ricercare

sequenze nucleotidiche o proteiche. Il tool può essere utilizzato copiando le sequenze d'interesse nell'apposita schermata o caricando direttamente un file FASTA o Multi-FASTA tramite l'apposito link.

Figura 11. Finestra per effettuare l'analisi BLAST

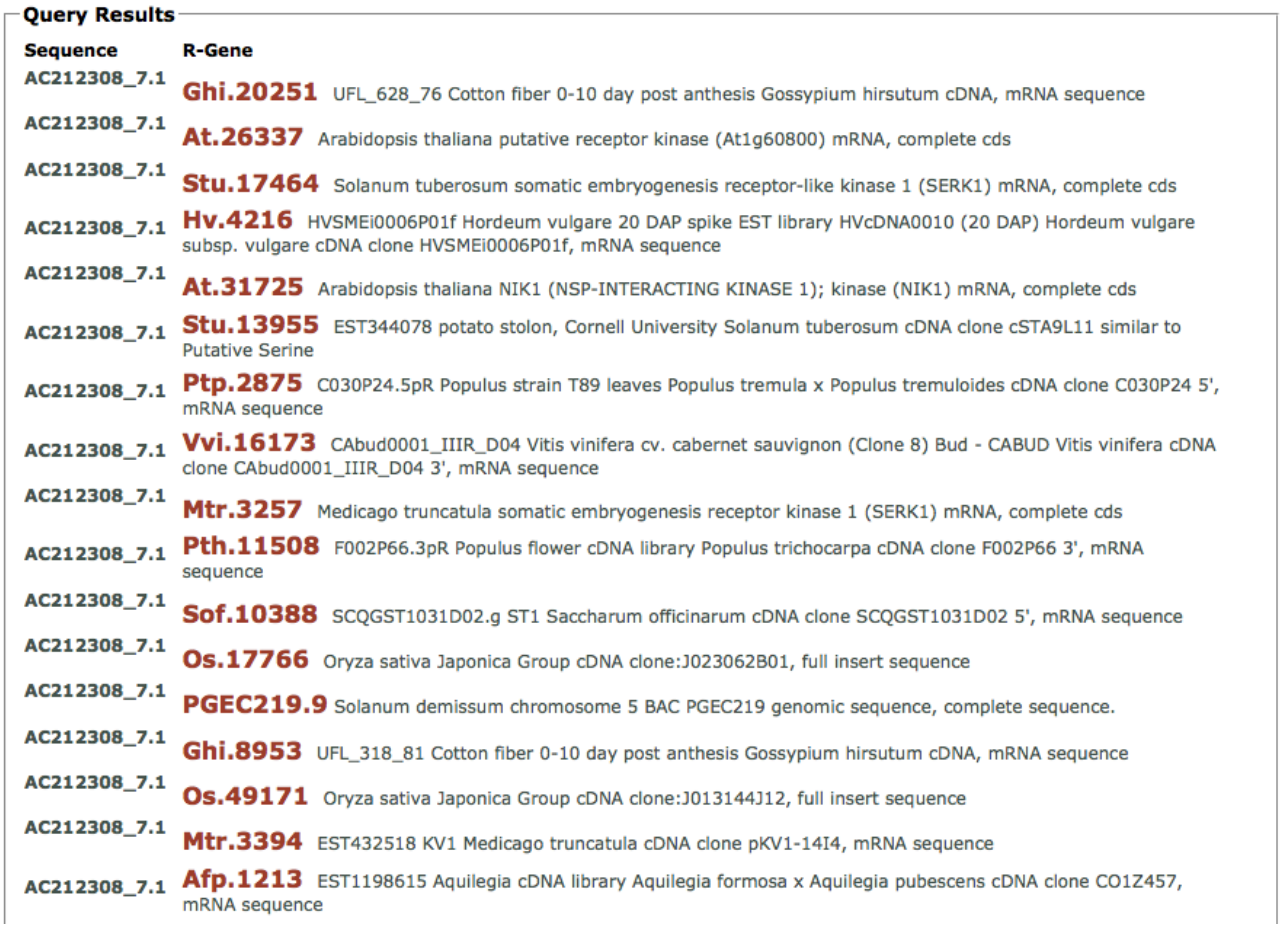


Figura 12. Esempio di risultato dell'analisi BLAST

3.3.4 L'interfaccia privata per il controllo dei dati

Al fine di creare un sistema per:

- Avere il controllo diretto sui dati inseriti
- Arricchire manualmente le sequenze geniche con dati nuovi
- Controllare l'operato degli *script* di manutenzione ed aggiornamento
- Aggiungere o eliminare velocemente informazioni
- Inserire nuovi geni di resistenza quando isolati

è stata creata un interfaccia privata, accessibile via Web tramite il link "curator site", attraverso la quale è possibile avere il completo controllo del database da qualsiasi postazione ed in qualsiasi momento. Tale sistema permette di controllare ed aggiornare PRG in tempo reale quando necessario.

3.4. DRAGO: uno strumento bioinformatico per la predizione di nuovi geni R

Obiettivo fondamentale per la ricerca nel campo delle resistenze vegetali è la caratterizzazione di nuove sequenze geniche che svolgono la funzione di resistenza a specifici patogeni. Per questo motivo, è stato ritenuto di primaria importanza la creazione di un sistema in grado di identificare con grande specificità i geni R in set di sequenze molto grandi. La progettazione di un predittore specifico per i geni R è stata molto complessa. **DRAGO**, "**D**isease **R**esistance **A**nalysis & **G**ene **O**rthology", un sistema di predizione creato appositamente per la ricerca di nuovi geni R, è stato sviluppato a partire dalle sequenze dei domini R conservati. Tali domini sono serviti come punto di partenza per effettuare ricerche su tutto il trascrittoma vegetale presente in banca dati. Manualmente, poi, sono state attribuite le combinazioni di domini conservati presenti, permettendo di classificare i geni nelle 5 classi conosciute e di evidenziarne di nuove.

3.4.1 Predizione di nuovi geni R e scelta del dataset di partenza

Al fine di creare un predittore che abbia alte probabilità di portare alla luce nuovi geni R, è stato ritenuto oppurtuno creare un sistema specifico in grado di funzionare su sequenze certamente espresse come le ESTs (Expressed Sequence Tag), in modo da trascurare volontariamente tutte le altre sequenze che hanno omologia con i geni R, ma che non si rinvengono nei trascrittomi degli organismi. Inoltre, per superare la problematica della frammentazione del gene, della ridondanza e della scarsa lunghezza delle sequenze nelle ESTs, si è scelto di utilizzare come dati di partenza il set UniGene di NCBI, le cui specifiche recitano così: "Ogni Unigene è un insieme di sequenze di

trascrizione che sembrano provenire dallo stesso locus di trascrizione (gene o pseudo-gene espresso)".

Le caratteristiche di tale set sono altamente idonee per lo sviluppo di un sistema, capace di individuare nuovi geni espressi omologhi ai già funzionali geni R. Per questo motivo sono stati raccolti 630mila UniGene provenienti da 33 organismi vegetali diversi ed analizzati tramite DRAGO.

3.4.2 Analisi delle sequenze UniGene ed integrazione dei risultati con PRGdb

Delle 630mila sequenze analizzate tramite DRAGO, 10463 sono risultate altamente omologhe ad uno o più geni R di referenza (Tabella 2). Così come per i dati provenienti da NCBI, anche queste sequenze sono state analizzate a livello proteico per la determinazione dei domini conservati e classificate secondo questo criterio. Le nuove sequenze sono state integrate con il database PRG ed il set di dati è stato definito *"putative R-Genes computationally predicted from Unigene ORF domain prediction"*. Per ogni nuovo putativo gene R è stato creato un foglio informativo secondo lo standard PRG e, come per gli altri set di dati, tutte le informazioni correlate alla sequenza sono state inserite. Grazie all'analisi dei domini proteici, per ogni sequenza è indicata la putativa classe di appartenenza che, integrata con la ricerca avanzata del PRG, permette di discernere le sequenze in base all'associazione di uno o più domini.

3.4.3 Analisi dei domini proteici dei geni predetti

L'analisi proteica dei geni predetti ha mostrato nuove associazioni di domini oltre alle 4 classi geniche in cui i geni R sono attualmente suddivisi. Grazie all'integrazione dei dati predetti con il PRG, è stato possibile non solo avere una visione d'insieme di tutte le associazioni proteiche trovate, ma anche di sapere esattamente il numero di sequenze presenti in ogni classe. Questo dato risulta ancora più interessante sapendo che tutte le sequenze analizzate sono espresse e quindi sono dotate di una propria funzionalità. In tabella 2 sono raccolti tutti i dati delle sequenze predette dal set UniGene, ordinati per specie di appartenenza.

SPECIE	N° UniGene
Oryza sativa	1834
Arabidopsis thaliana	1168
Triticum aestivum	639
Zea mays	600
Hordeum vulgare	466
Glycine max	424
Gossypium hirsutum	421
Pinus taeda	349
Medicago truncatula	331
Solanum tuberosum	326
Solanum lycopersicum	319
Malus x domestica	317
Brassica napus	303
Vitis vinifera	293
Picea sitchensis	274
Saccharum officinarum	245
Populus trichocarpa	239
Brassica rapa	224
Picea glauca	196
Nicotiana tabacum	189
Sorghum bicolor	176
Aquilegia formosa x Aquilegia pubescens	162
Lactuca sativa	147
Physcomitrella patens	129
Citrus sinensis	110
Lotus japonicus	101
Citrus clementina	86
Chlamydomonas reinhardtii	82
Helianthus annuus	78
Populus tremula x Populus tremuloides	71
Brassica oleracea	59
Prunus persica	59
Gossypium raimondii	46

Tabella 2. Visualizzazione di tutte le sequenze UniGene presenti nel PRG, divise per specie di appartenenza.

3.5 Analisi delle sequenze contenute nel database PRG

L'integrazione dei dati raccolti in modo manuale o automatico con i dati provenienti dalla pipeline di predizione DRAGO ha permesso di ottenere lo straordinario numero di 16844 sequenze, provenienti da 194 piante, coinvolte nelle risposte di difesa vegetali o altamente omologhe ad esse. Al fine di visualizzare al meglio quest'importante risultato, sono stati estrapolati i dati per produrre diverse tabelle e grafici informativi. In tabella 3, tutte le sequenze sono state divise a seconda della specie vegetale alla quale appartengono. Come è possibile notare, per *Arabidopsis* abbiamo più di 4000 sequenze putative per la funzione di resistenza, a seguire le specie vegetali i cui genomi sono stati sequenziati o sono in fase di sequenziamento ed al decimo posto iniziano a comparire le prime specie afferenti al genere *Solanum*. Interessanti sono anche le sequenze provenienti dalle specie selvatiche, dalle piante arboree e dalle specie afferenti al genere *Cytrus*. In figura 13, per evidenziare i risultati ottenuti dalla studio dei domini proteici, è stato creato un Venn i cui insiemi sono composti dai cinque domini specifici dei geni R e

le cui intersezioni corrispondono alle diverse associazioni di domini riscontrate nelle proteine.

SPECIE	N° geni
Arabidopsis thaliana	4211
Oryza sativa	1850
Triticum aestivum	700
Oryza sativa Japonica Group	685
Zea mays	620
Malus x domestica	519
Hordeum vulgare	514
Glycine max	500
Gossypium hirsutum	422
Solanum tuberosum	391
Solanum lycopersicum	370
Pinus taeda	352
Medicago truncatula	348
Brassica napus	340
Populus trichocarpa	315
Arabidopsis lyrata	298
Vitis vinifera	293
Picea sitchensis	274
Saccharum officinarum	257
Brassica rapa	236
Nicotiana tabacum	231
Sorghum bicolor	202
Picea glauca	196
Lactuca sativa	165
Aquilegia formosa x Aquilegia pubescens	162
Physcomitrella patens	133
Citrus sinensis	111
Malus floribunda	105
Lotus japonicus	101
Solanum demissum	92
Humulus lupulus	91
Citrus clementina	86
Chlamydomonas reinhardtii	82
Helianthus annuus	79
Brassica oleracea	74
Populus tremula x Populus tremuloides	71
Ipomoea batatas	67
Populus tremula	66
Prunus persica	60
Lactuca serriola	51
Solanum melongena	50
Gossypium raimondii	46
Capsicum annuum	46
Vicia faba	44
Musa acuminata	37
Arachis hypogaea	32
Pyrus pyrifolia	30
Pyrus communis	29
Oryza sativa Indica Group	27
Oryza rufipogon	26
Eleusine coracana	26
Phaseolus vulgaris	25
Cicer arietinum	25
Citrus grandis x Poncirus trifoliata	22
Solanum peruvianum	20
Gossypium barbadense	19
Hordeum vulgare subsp. vulgare	19
Beta vulgaris	18
Pyrus sinkiangensis	18
Musa balbisiana	18
Solanum pimpinellifolium	16
Malus prunifolia	16
Musa acuminata AAA Group	14

Pyrus x bretschneideri	14	*var. truncata*	
Solanum chilense	14	*Oryza longistaminata*	4
Musa ABB Group	14	*Musa banksii*	3
Coffea arabica	13	*Musa ornata*	3
Solanum sucrense	13	*Solanum chmielewskii*	3
Fragaria x ananassa	13	*Vigna unguiculata*	3
Musa AAB Group	13	*Cucumis melo*	3
Solanum habrochaites	12	*Lactuca saligna*	3
Musa acuminata subsp. malaccensis	12	*Citrus aurantium*	3
Pyrus hybrid cultivar	11	*Hordeum vulgare subsp. spontaneum*	3
Musa textilis	10	*Musa velutina*	3
Pisum sativum	9	*Musa acuminata subsp. burmannicoides*	3
Solanum bulbocastanum	9	*Citrus reticulata*	3
Musa schizocarpa	9	*Vitis hybrid cultivar*	3
Theobroma cacao	9	*Solanum vernei*	3
Ipomoea trifida	8	*Zingiber officinale*	3
Rosa roxburghii	8	*Solanum lycopersicum var. cerasiforme*	3
Zizania latifolia	7	*Musa textilis x Musa ABB Group*	3
Solanum berthaultii	7	*Aegilops tauschii*	3
Lolium perenne	7	*Citrus longispina*	2
Coffea canephora	7	*Citrus webberi*	2
Musa balbisiana x Musa textilis	6	*Solanum aethiopicum*	2
Musa acuminata subsp. errans	6	*Citrus halimii*	2
Musa acuminata subsp. microcarpa	6	*Medicago sativa*	2
Oryza meyeriana	6	*Nicotiana debneyi*	2
Citrus medica	5	*Oryza meridionalis*	2
Pyrus ussuriensis	5	*Citrus hanaju*	2
Malus baccata	5	*Solanum neorickii*	2
Brassica rapa subsp. oleifera	5	*Musa acuminata subsp. siamea*	2
Populus deltoides	5	*Camelina sativa*	2
Arabidopsis arenosa	4	*Solanum torvum*	2
Linum usitatissimum	4	*Solanum lycopersicoides*	2
Pyrus betulifolia	4	*Citrus nippokoreana*	2
Oryza officinalis	4	*Fragaria chiloensis*	2
Citrus maxima	4	*Avena sativa*	2
(Populus tomentosa x P. bolleana) x P. tomentosa	4	*Citrus amblycarpa*	2

Solanum tuberosum subsp andigena	2
Elaeis oleifera	2
(Populus tomentosa x P. bolleana) x P. tomentosa	2
Citrus aurantiifolia	2
Nicotiana glutinosa	1
Microcitrus australasica	1
Aegilops peregrina	1
Cucurbita ficifolia	1
Fortunella margarita	1
Vigna vexillata	1
Populus balsamifera	1
Solanum 52rnesi x Solanum hondelmannii	1
Poncirus trifoliata	1
Populus alba	1
Atalantia ceylanica	1
Citrus ichangensis	1
Arabidopsis halleri	1
Solanum acaule	1
Daucus carota	1
Capsicum chacoense	1
Capsicum chinense	1
Citrus nobilis	1
Brassica rapa subsp. pekinensis	1
Brassica nigra	1
Triticum turgidum	1
Sesbania rostrata	1
Begonia hybrid cultivar	1
Capsella rubella	1
Fragaria vesca	1
Setaria italica	1
Carex blanda	1
Prunus dulcis	1
Citrus limon	1
Mangifera indica	1
Pinus radiata	1
Pinus sylvestris	1
Citrus unshiu	1
Citrus limettioides	1

Tabella 3. Visualizzazione di tutti i geni presenti nel database PRG divisi per specie di appartenenza

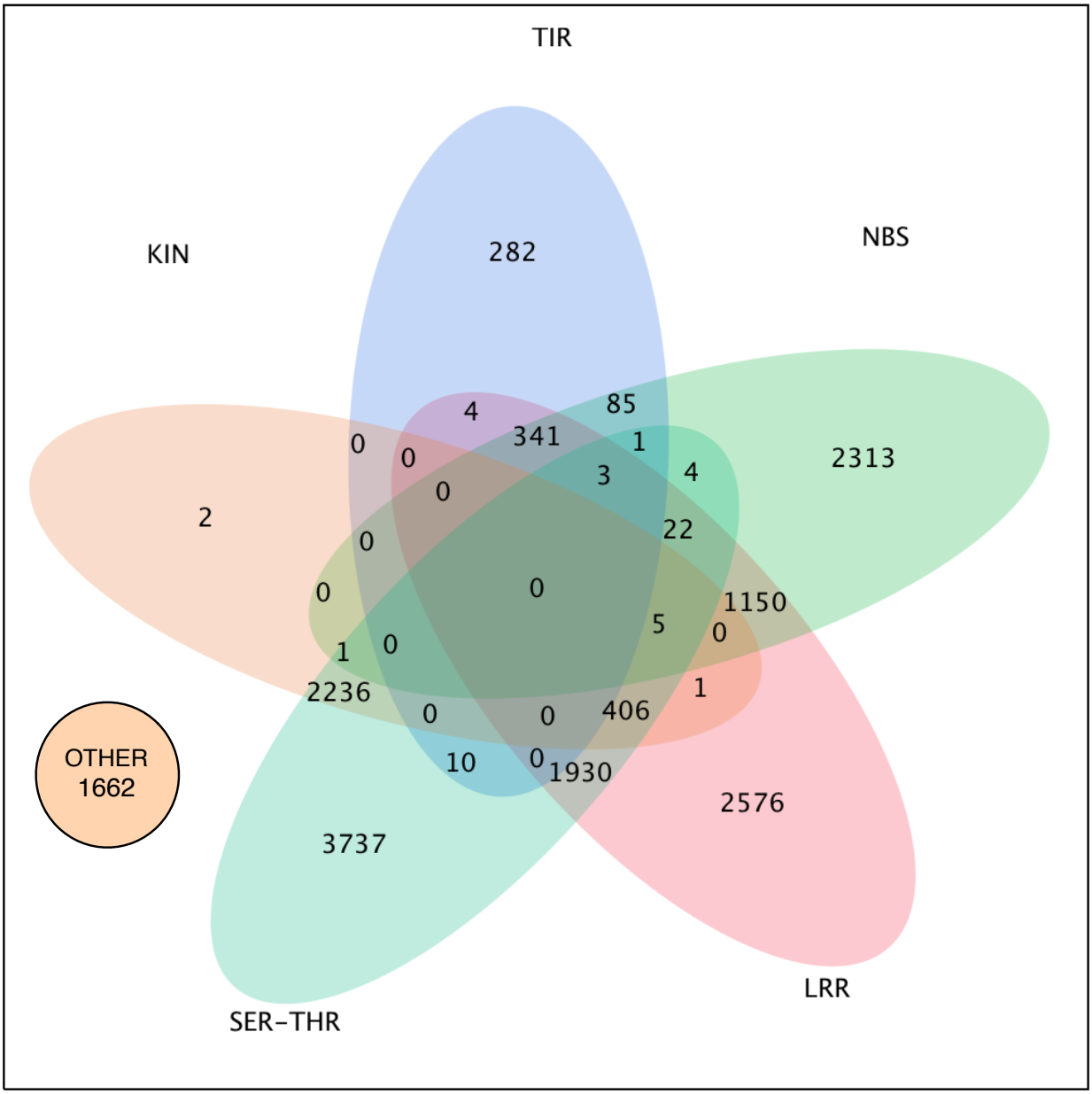


Figura 13. Venn rappresentativo di tutte le associazioni di domini proteici prodotti tramite l'analisi di DRAGO

3.5.1 La classe "Other"

Tramite DRAGO è stato possibile effettuare delle predizioni anche sulla classe dei geni di resistenza definita "other". A causa delle caratteristiche peculiari delle sue sequenze, ad oggi non esistono studi che descrivano questi geni e che li caratterizzino. Partendo dai 13 geni R di referenza sopra descritti, il sistema di predizione messo a punto è stato in grado di raccogliere nei diversi set di dati 1662 sequenze correlate ad essi. È ovviamente errato parlare di queste sequenze come facenti parte di un'unica classe con specifiche caratteristiche, ma è pur vero che tali sequenze mostrano un'elevata omologia per uno o più geni, i cui domini non sono tipici, ma la cui funzione è quella di resistenza e per

questo motivo è stato ritenuto opportuno la loro caratterizzazione ed il loro inserimento nel PRGdb. L'inserimento delle sequenze della classe "other" all'interno del PRGdb permette non solo di creare una risorsa sui geni di resistenza completa, ma anche di permettere, a coloro che studiano geni di resistenza con caratteristiche peculiari, di avere un punto di riferimento in cui trovare valide informazioni.

3.5.2 Nuove associazioni proteiche

Grazie alla predizione effettuata da DRAGO, è stato possibile evidenziare che, oltre alle classi conosciute appartenenti alla famiglia dei geni R, esistono proteine strutturate in modo da contenere molte altre associazioni dei domini tipici dei geni R o singoli domini. In tabella 3 è possibile osservare le sequenze predette da DRAGO divise per associazioni di domini ed il loro numero.

Tipologia di domini	N° Seq	Classe
Ser/thr	3737	Sconosciuta
LRR	2576	Sconosciuta
NBS	2313	Sconosciuta
KIN-Ser/thr	2236	Chinasi citoplasmatica (gene *pto*)
LRR-Ser/thr	1930	RLP
LRR-NBS	1150	CNL
LRR-KIN-Ser/thr	406	RLKb
TIR-NBS-LRR	341	TNL
TIR	282	Sconosciuta
TIR-NBS	85	Sconosciuta
LRR-NBS-Ser/thr	22	Sconosciuta
TIR-Ser/thr	10	Sconosciuta
NBS-LRR-KIN-Ser/thr	5	Sconosciuta
TIR-LRR	4	Sconosciuta
NBS-Ser/thr	4	Sconosciuta
TIR-NBS-LRR-Ser/thr	3	Sconosciuta
KIN	2	Sconosciuta
KIN-LRR	1	Sconosciuta
TIR-NBS-Ser/thr	1	Sconosciuta
NBS-KIN-Ser/thr	1	Sconosciuta

Tabella 4. Numero di proteine predette tramite DRAGO divise per associazioni di domini conservati

Le nuove associazioni trovate nei vari set di dati analizzati fanno supporre un più complesso panorama del sistema dei geni di resistenza, che non si ferma alle sole 4 classi trattate fino ad oggi. Nonostante siano molto chiare le caratteristiche delle proteine che fanno parte delle nuove classi, le nuove associazioni sono state definite, qui come nel PRG, come "classe sconosciuta" al fine di poter validare tramite esperimenti molecolari i dati ottenuti *in silico.*

Complessivamente le associazioni di domini nuove sono 15 ed il numero di sequenze ad esse associate è di gran lunga superiore a quello delle sequenze afferenti alle 4 classi conosciute. Il numero di sequenze varia a seconda della complessità della struttura proteica, costituendo classi molto numerose per proteine contenenti un singolo dominio o classi povere per strutture più complesse (figura 14).

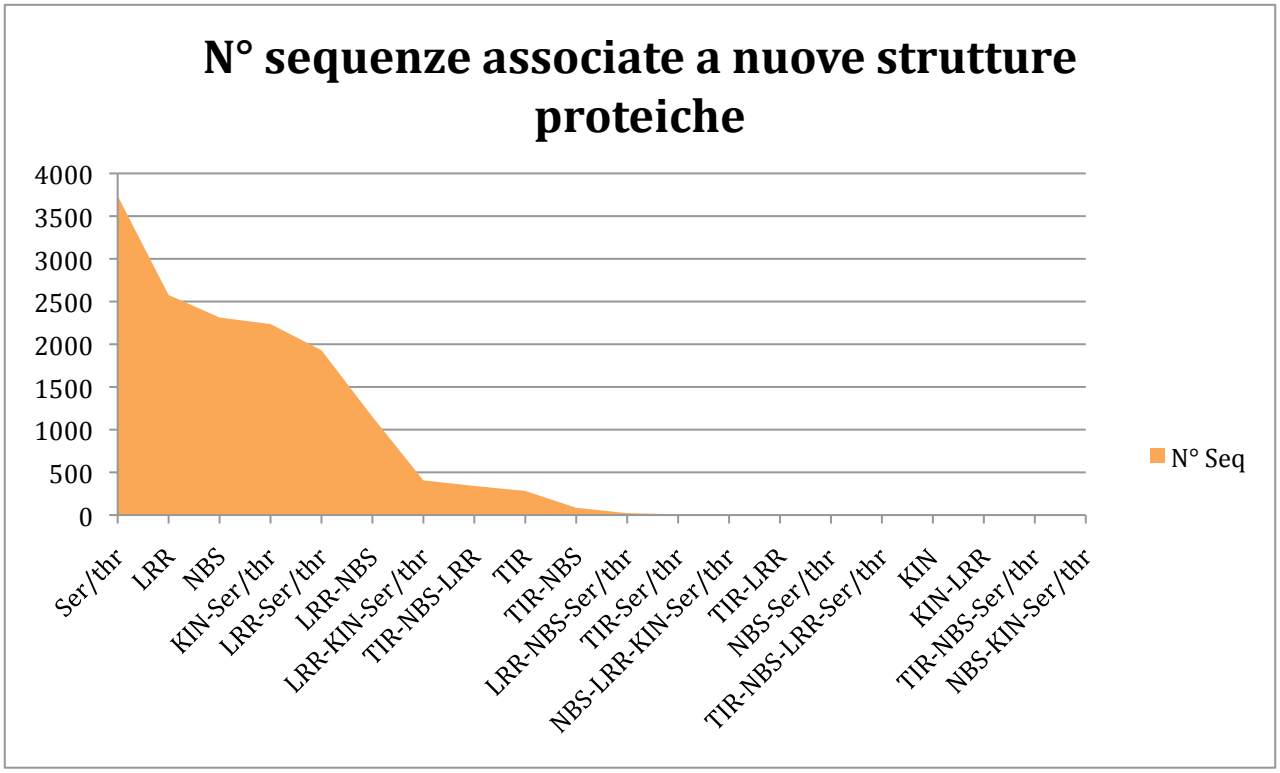


Figura 14. N° sequenze associate a nuove strutture proteiche

3.6 MATRIX: un sistema di predizione genica ad alta efficienza

Un'altra importante fonte, ancora poco sfruttata per la raccolta di informazioni e per la ricerca di nuovi geni R candidati, è la collezione di genomi vegetali parzialmente o totalmente sequenziati. Utilizzando le informazioni di sequenziamento è possibile avere una visione completa della genomica di un organismo e, costruendo strumenti adeguati, è possibile ricavare importanti informazioni sulle famiglie geniche presenti, sulla loro localizzazione e sulle loro caratteristiche.

Per sfruttare i genomi vegetali al fine di ottenere informazioni sulla famiglia dei geni R, è stato costruito uno strumento di predizione ad alta efficienza in grado di predire i putativi geni R da uno o più genomi contemporaneamente. Questo strumento è stato denominato MATRIX ed è in grado di lavorare su set di proteine annotate su genomi sequenziati. Il riconoscimento, all'interno di quelli che possono essere definiti "proteomi virtuali", delle proteine omologhe alle proteine R avviene tramite la produzione di una matrice di similarità tra le 73 proteine R di referenza e le proteine a funzione sconosciuta. Tramite la matrice è possibile discriminare e dividere dal set analizzato tutte le sequenze omologhe alle già funzionali proteine R.

3.6.1 Sviluppo di MATRIX

3.6.1.1 Dati di partenza e creazione dei moduli HMM

I 32 geni R di referenza, provenienti dalla famiglia delle *Solanaceae* già collezionati ed analizzati, sono stati divisi in classi filogenetiche di appartenenza. Tale divisione rispecchia a livello proteico un'omologia data dalla presenza di domini conservati. Ogni singolo ramo dell'albero filogenetico specifica una determinata classe di appartenenza delle proteine R ed in questo caso l'analisi ha prodotto un albero composto da tre rami ben distinti, corrispondenti alle tre classi di resistenza CNL, TNL e RLP. Poiché il set dei geni di referenza delle *Solanaceae* non contiene geni della classe RLK, per lo sviluppo del sistema di predizione di questa classe genica sono state utilizzate le 200 proteine RLK presenti nel genoma di *Arabidopsis thaliana*. Le proteine, per ogni singola classe, sono state allineate tra di esse e per ogni set è stata identificata una sequenza numerica consenso tipica di quella classe proteica. La sequenza numerica consenso è stata determina tramite la creazione di uno specifico *script* in grado di leggere l'allineamento, e calcolare per ogni singola posizione amminoacidica un valore, basato sulla matrice BLOSUM62 con gap -11, che ne rispecchi l'omologia. Tramite i valori ottenuti per ogni

singola posizione è stato possibile calcolare il valore medio di omologia di tutto l'allineamento e, grazie a questo valore, creare un processo attraverso il quale estrapolare dall'allineamento stesso tutte le regioni con valori più alti delle media e con lunghezza maggiore di 8 amminoacidi. Questo processo ha permesso di ottenere per ogni classe proteica della famiglia R regioni lunghe più di 8 AA molto conservate, che rispecchiano la presenza o meno di domini all'interno delle sequenze. Le regioni estratte tramite questa procedura sono state utilizzate per la produzione di profili HMM, rispecchiando fedelmente le caratteristiche delle regioni isolate (figura 15). In tutto sono stati creati 308 profili HMM divisi come segue:

- 146 moduli per la classe CNL
- 56 moduli per la classe RLK
- 48 moduli per la classe RLP
- 50 moduli per la classe TNL

Se per le classi TNL e RLP, le cui proteine sono risultate molto conservate, la specificità dei moduli ha prodotto durante i test preliminari una buona complementazione con le sequenze, per le classi CNL e RLK, le cui sequenze hanno mostrato forte variabilità, sono stati effettuati dei passaggi successivi al fine di creare più moduli che riuscissero a superare questa problematica. Studiando dettagliatamente la filogenesi della classe CNL delle *Solanaceae* è stato visto che la stessa è composta da due gruppi distinti e che all'interno del secondo gruppo esistono due sottodivisioni con peculiari caratteristiche. Inoltre per classe RLK la filogenesi ha mostrato la presenza di 3 gruppi distinti. Grazie a questa analisi i profili delle due classi in oggetto sono stati ridisegnati in modo da essere più specifici non solo a livello delle classi, ma anche delle sottodivisioni. In tutto sono stati prodotti:

- 42 moduli HMM per la classe CNL sottoclasse 1
- 38 moduli HMM per la classe CNL sottoclasse 2 divisione 1
- 66 moduli HMM per la classe CNL sottoclasse 2 divisione 2
- 22 moduli HMM per la classe RLK sottoclasse 1
- 14 moduli HMM per la classe RLK sottoclasse 2
- 20 moduli HMM per la classe RLK sottoclasse 3

I 308 moduli così divisi rappresentano in modo esaustivo le caratteristiche proteiche delle 4 classi dei geni R e possono essere utilizzati per la creazione di un sistema di analisi ad alta efficienza.

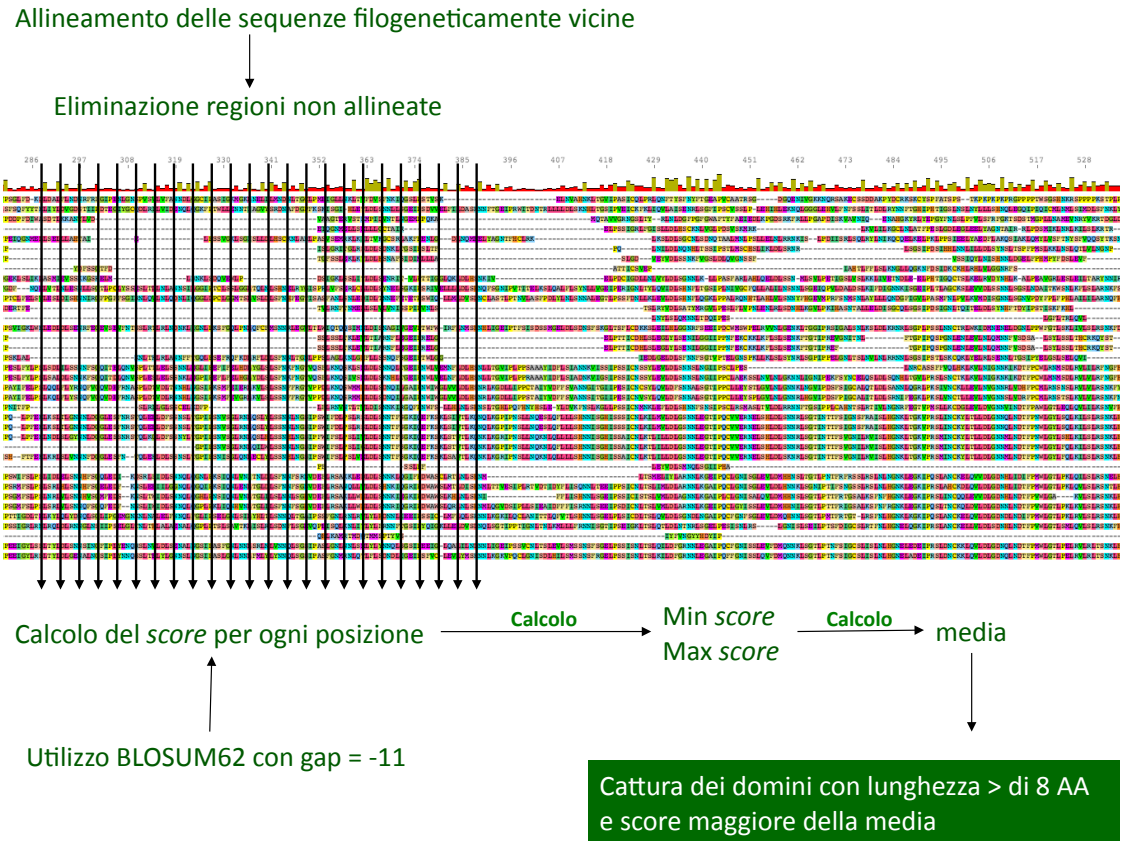


Figura 15. Rappresentazione schematica del sistema di costruzione dei profili HMM specifici per i geni R

3.6.1.2 Sviluppo del sistema MATRIX

I moduli HMM sono stati utilizzati per sviluppare la parte finale del sistema MATRIX. Un apposito *script* è stato creato per utilizzare i moduli HMM, per il vaglio delle proteine, alla ricerca di sequenze contenenti i domini conservati delle proteine R di referenza. Entrando nello specifico, questa complessa operazione viene effettuata lanciando ogni singolo profilo HMM su una proteina a funzione sconosciuta, se il modulo non incontrerà nessun omologia con la proteina analizzata si otterrà un valore pari a zero, nel caso in cui invece il modulo incontri una regione più o meno simile, lo *script* produrrà un valore tanto alto quanto alta è la specificità con la proteina stessa. Questo processo viene ripetuto su ogni singola proteina con tutti i 308 moduli, creando così una matrice di similarità che indica quali moduli sono contenuti nella proteina e con quale specificità. Con un set di dati contenente più di una proteina, questo processo può essere ripetuto in modo da ottenere una matrice che indica i dati non più di una sola proteina, ma dell'intero set di dati. Questo processo, nonostante complesso, è molto veloce ed è in grado di processare migliaia di proteine contemporaneamente. Tale caratteristica gli permette di analizzare interi genomi vegetali alla ricerca di proteine con caratteristiche

compatibili a quelle di resistenza. Per ridurre la grandezza delle matrici ed il numero di falsi positivi ottenuti, sono stati messi appunto tre utili accorgimenti che hanno reso i risultati più affidabili. In primo luogo la normalizzazione del dato ottenuto in relazione alla grandezza del set ed alla sua specificità con i moduli HMM, il secondo tramite l'eliminazione dopo l'ottenimento della matrice di tutte le proteine che non presentavano nessuna omologia, ed infine l'eliminazione di tutte le proteine con un'omologia inferiore ad 8 moduli HMM. Grazie a queste tre migliorie, attualmente MATRIX può essere non solo utilizzato indifferentemente su genomi di mono e di-cotiledoni, ma presenta anche un'alta percentuale di specificità per tutte le classi delle proteine R (figura 16).

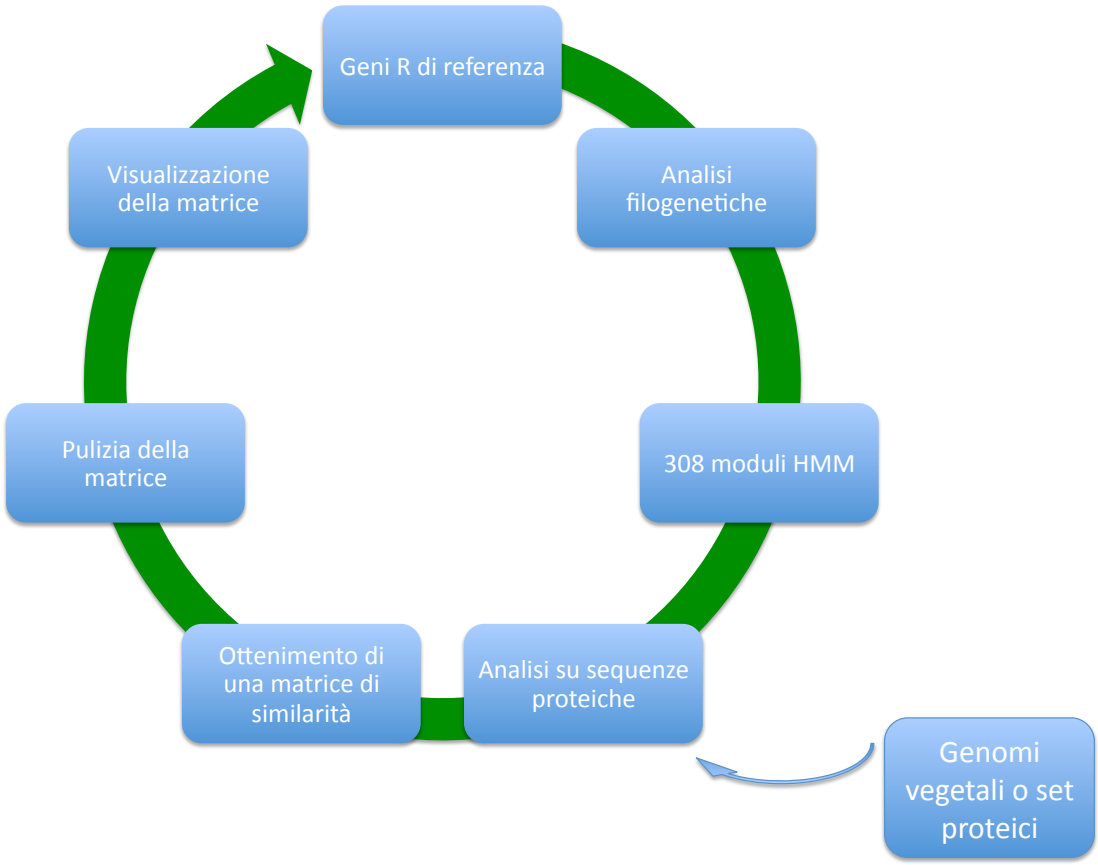


Figura 16. Schematizzazione del processo di predizione offerto da MATRIX

3.6.1.3 Utilizzo di un'interfaccia grafica per la visualizzazione dei dati

Analizzando set proteici numerosi, risulta molto complesso leggere le matrici numeriche prodotte da MATRIX. Per questo motivo i dati prodotti sono stati formattati in modo tale da essere letti da semplici software di raggruppamento e visualizzazione dei dati d'espressione (anche loro sono costituiti da matrici numeriche), così da poter essere

visualizzati in modo semplice e facile da interpretare. La figura 17 mostra la matrice prodotta dalla predizione dei geni R in pomodoro. Al fine di rendere più maneggevole la matrice, sono visualizzate solo le proteine che hanno alta omologia con i geni R di referenza, mentre tutte le altre sono state eliminate. Sull'asse delle ascisse sono posizionati tutti i profili HMM costruiti sui geni R divisi per classe R di appartenenza, mentre sulle ordinate sono posizionate le singole proteine analizzate, che rispecchiano omologia con i profili delle differenti classi. Le diverse colorazioni degli incroci tra le ascisse e le ordinate rappresentano la presenza o meno di quel profilo HMM all'interno della proteina, variando da nero (assenza di omologia) a rosso intenso (alta omologia). Tutte le proteine sono state raggruppate secondo l'algoritmo gerarchico (hierarchical clustering), creando gruppi ben distinti in base al profilo ottenuto. Questa tipologia di matrice permette anche di osservare le differenza tra le singole proteine dei raggruppamenti e mette in risalto la presenza di sottogruppi all'interno delle classi R.

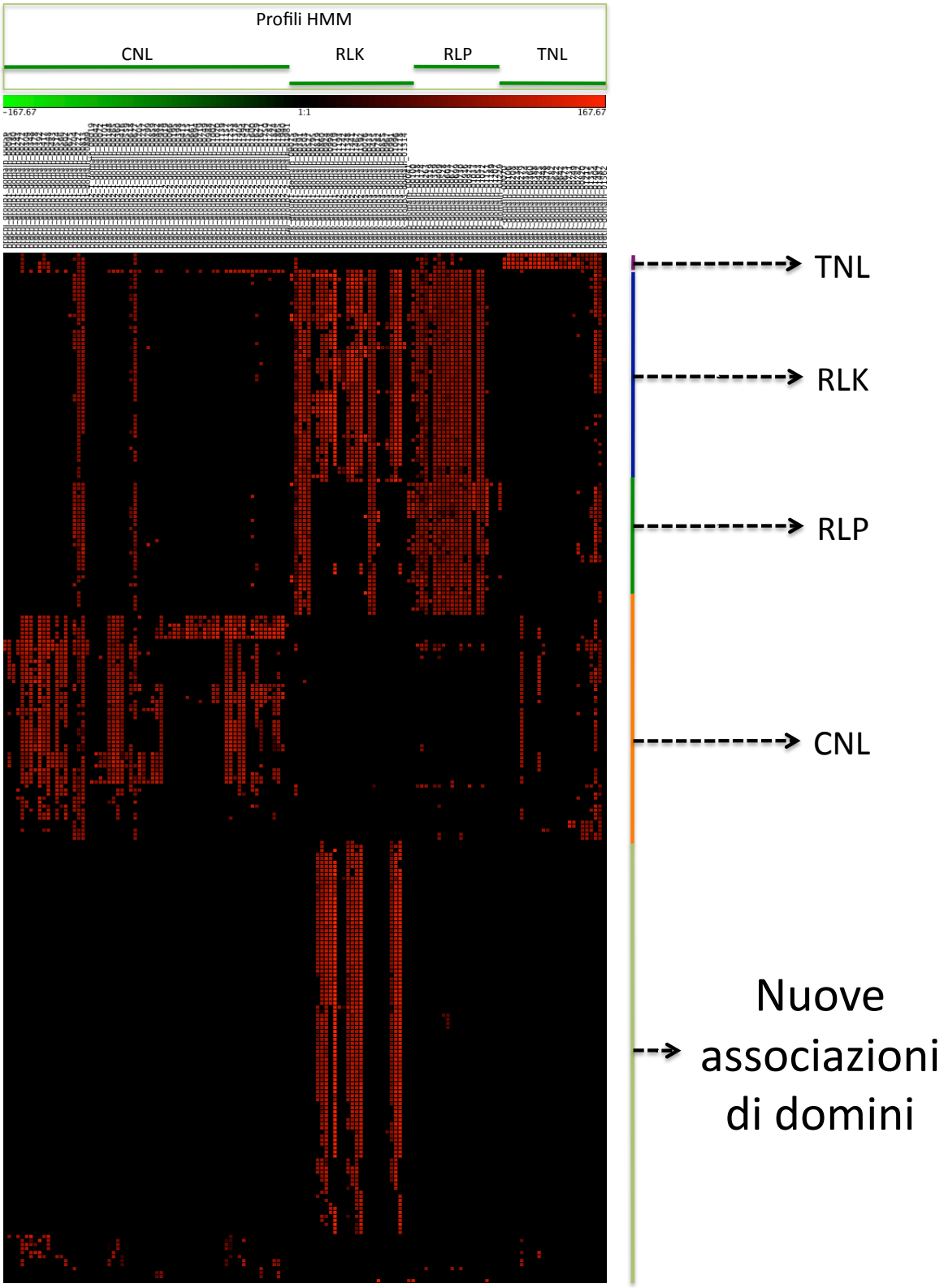


Figura 17. Risultato di MATRIX per l'analisi del genoma del pomodoro

3.6.2 Analisi tramite MATRIX di diversi set proteici

3.6.2.1 Il set delle Viridiplantae

Per testare l'efficienza e la potenza dello strumento di predizione sviluppato, è stato analizzato il set proteico composto da tutte le proteine vegetali del regno *Viridiplantae*, raccolte in NCBI. Il predittore non ha avuto problemi ad analizzare il set composto da 827753 proteine ed ha prodotto una matrice contenente 28524 proteine con regioni omologhe ai domini funzionali delle proteine R. A causa della grandezza del set ottenuto, non esistono ad oggi programmi di visualizzazione che permettano di rendere la matrice leggibile e quindi fruibile a tutti gli utilizzatori. In ogni caso, le informazioni sulle sequenze predette sono state conservate e nuovi strumenti sono in sviluppo per l'utilizzazione dei dati ottenuti.

3.6.2.2 Il set delle Solanaceae

In quanto famiglia modello per lo studio delle resistenze, le proteine derivanti dalle *Solanaceae* rappresentano un'importante fonte per la scoperta di nuovi candidati per la funzione di resistenza. Tramite MATRIX sono state analizzate 20580 proteine di questa famiglia, di cui 1181 sono risultate positive alla selezione *in silico.* Grazie alla plasticità del sistema di predizione, il set non solo è stato diviso in 13 sottoclassi, ognuna omologa ad un pool di proteine di referenza, ma è stato diviso anche in base alle 40 specie da cui le proteine derivano. Questa divisione permette di effettuare studi comparativi tra le diverse specie afferenti alle *Solanaceae* e tra i geni predetti e quelli già conosciuti (tabella 5, figure 18 e 19). Attraverso i dati riportati è possibile osservare i geni afferenti alle singole specie vegetali, confrontarli tra di essi, capire quali sono i genotipi su cui si è svolto un maggior lavoro d'analisi di sequenze, analizzare i geni R delle specie selvatiche e di nicchia. La tabella 5 illustra anche le omologie delle proteine predette con quelle di referenza, in modo da rendere facilmente comprensibile quali sono gli omologhi presenti nelle diverse specie afferenti alla famiglia delle *Solanaceae*. Per le specie con informazioni di sequenze (*S. tuberosum*, *S. lycopersicum* , *S. demissum*) è possibile anche tracciare e confrontare i profili delle resistenze in base alle percentuali di geni afferenti alle diverse classi R.

Nomi proteine R di referenza		rpi-blb1, rps1-k-1, rps1-k-2, R3a, I2, rps5, rps2, Rpm1, rpg1b, rpg8, rcy1, hrt, pi-ta, mla10	Bs2, tm2a, tm2-2, Gpa2, rx1,rx2	Rf, Prf, Sw5a, Hero, Mi1, rpi-blb2,	Ve1, Ve2, Cf9b, Cf4a, Cf2, Le6IX1, Le EIX2, Cf9, Cf4	Pvgip	P2, N, Bs4, Ry1, L6, M, rps4, rpp5, gro1.4	Pepr1, fls2, efr, Xa21, Er,	-	Pto	Rpg2, Ryfo	-	-	-	
		Cnl gr 1	Cnl gr. 2	Cnl gr. 3	Rlp gr. 1	Rlp gr. 2	Tnl	Rlk	Kin gr. 1	Kin gr. 2	Kin gr. 3	Kin gr. 4	Indef. gr.1	Indef. gr.2	Totale
Capsicum annuum				2		31			1	3	6				43
Capsicum chacoense			2												2
Capsicum chinense		1				1				6	2				10
Capsicum frutescens								1							1
Nicotiana benthamiana								1			1	2	1		5
Nicotiana glutinosa							1				1				2
Nicotiana plumbaginifolia						1									1
Nicotiana tabacum												1			1
Petunia x hybrida					2										2
	Cnl gr 1	Cnl gr. 2	Cnl gr. 3	Rlp gr. 1	Rlp gr. 2	Tnl	Rlk	Kin gr. 1	Kin gr. 2	Kin gr. 3	Kin gr. 4	Indef. gr.1	Indef. gr.2	Totale	
Solanum acaule		1			2									1	2
Solanum aethiopicum						2								2	2
Solanum arnezii_x_Solanum_hondelmannii						2			1	1		4		6	6
Solanum bresidens					3									3	3
Solanum bulbocastanum	14		2					1			1			18	18
Solanum chacoense							2							2	2
Solanum chilense				1	27									39	39
Solanum chmielewskii				1	1									4	4
Solanum circaeifolium								4						4	4
Solanum demissum	26		31		5		2	3	1	2		15	3	88	88
Solanum habrochaites				8	6				5	3				22	22
Solanum lycopersicoides				2	2				2					4	4
Solanum lycopersicum	3	4	20	20	5		5		1	10	10	1		79	79
Solanum melongena								1						1	1
Solanum neorickii				1					1					2	2
Solanum nigrum			1		3					1		2	4	8	8
Solanum peruvianum		1	2	6	39		2		10	8		1	3	69	69
Solanum phureja_x_Solanum_stenotomum												3		3	3
Solanum pimpinellifolium	2		3	25	19		1		3	2		1		56	56
Solanum pinnatisectum												3		3	3
Solanum sp._ VFNT			2											2	2
Solanum stoloniferum	2												3	5	5
Solanum sucrense								3		3		7		13	13
Solanum tarijense	1													1	1
Solanum torvum				2	1									3	3
Solanum tuberosum	14	9	9		19		11	4	1	2	6	3	34	27	139
Solanum tuberosum_subsp._andigena		1					2							3	3
Solanum vernei										2		1		3	3
Solanum verrucosum	2										3			2	2
Solanum virginianum														3	3
TOTALE															660
Somma delle colonne	65	18	72	70	165		16	19	12	48	49	28	64	34	660
Numero di proteine R di referenza	14	6	6	9	1		9	5	0	1	2	0	0	0	53

Tabella 5 Risultato della predizione, ad opera di MATRIX, su tutte le proteine provenienti da NCBI afferenti alla famiglia delle *Solanaceae* divise per specie e per omologia con i geni R di referenza

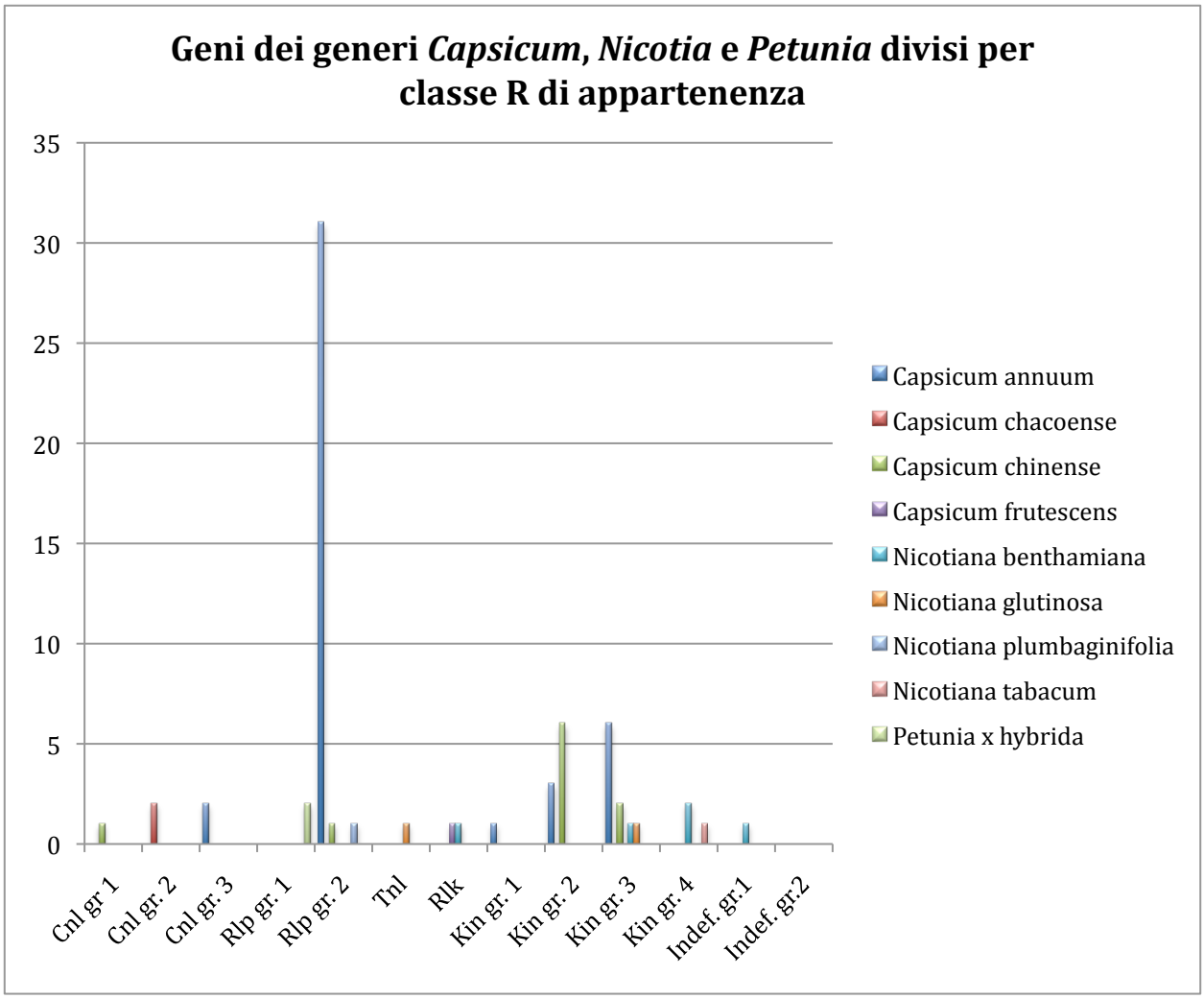


Figura 18. Distribuzione delle proteine predette tramite MATRIX dei generi *Capsicum*, *Nicotiana* e *Petunia* divisi per classe R di appartenenza

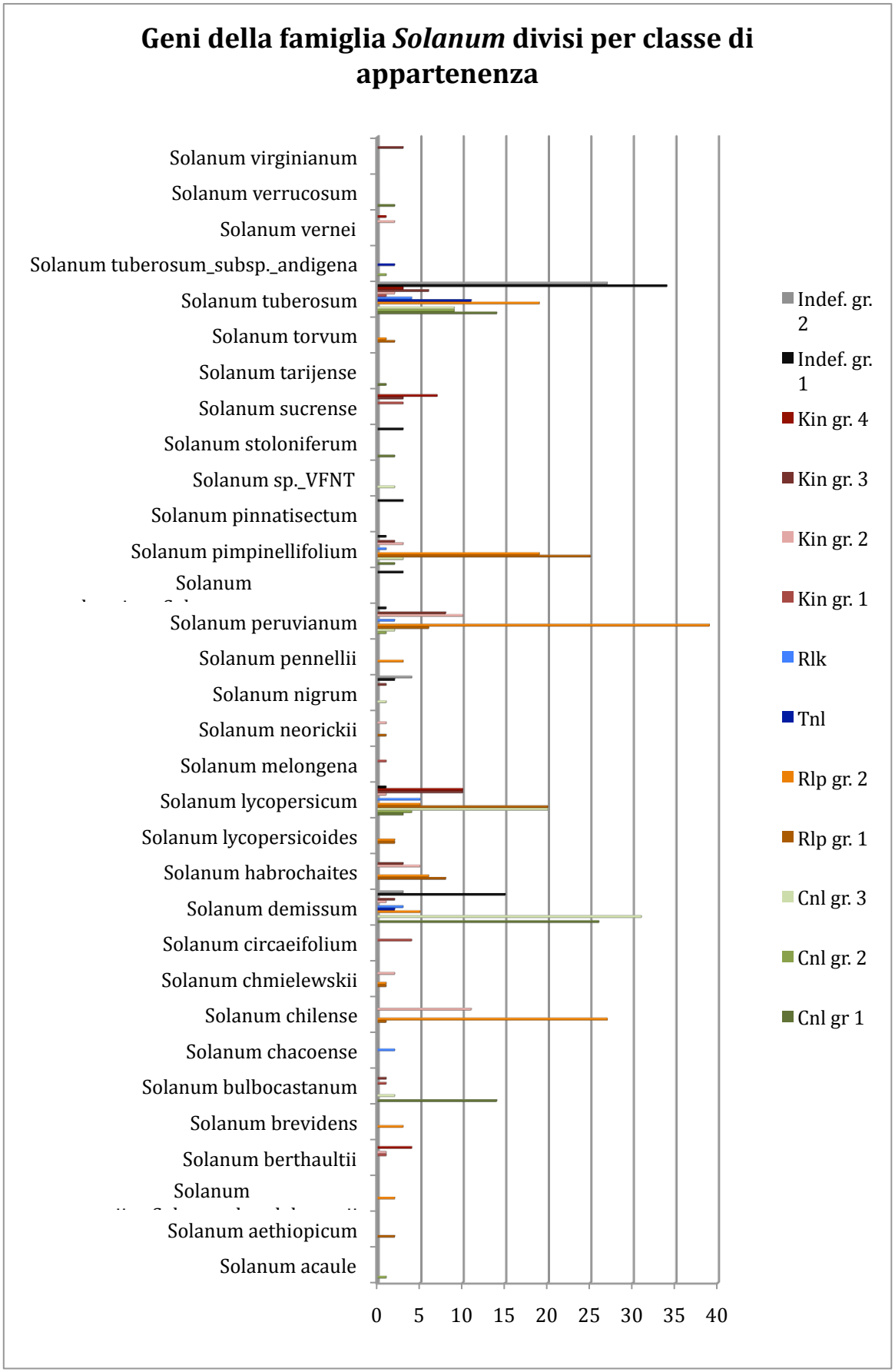


Figura 19. Distribuzione delle proteine predette tramite MATRIX del genere *Solanum* divisi per classe R di appartenenza

3.6.3 Analisi tramite MATRIX dei genomi vegetali sequenziati

Otto genomi vegetali, cinque già sequenziati e tre parzialmente sequenziati, sono stati analizzati al fine di creare un esaustivo panorama delle proteine di resistenza all'interno dei propri genomi. Per rendere i dati comparabili tra loro, i genomi sono stati analizzati contemporaneamente in un'unica analisi, andando poi a dividere le sequenze in base alla provenienza. La tabella 6 mostra i dati utilizzati con input e quelli ottenuti come output. Di fondamentale importanza è conoscere la grandezza del genoma di partenza ed il numero di proteine predette al suo interno, in modo da poter effettuare efficaci studi comparativi. Dai 5 genomi completamente sequenziati *Arabidopsis thaliana*, *Vitis vinifera*, *Popolus trichocarpa*, *Sorghum bicolor* ed *Oryza sativa* rispettivamente sono state ottenute 1102, 2044, 2361, 3089 e 1975 proteine correlate ai geni R, mentre per i tre genomi parzialmente sequenziati *Solanum lycopersicum* , *Lotus japonica* e *Zea mais* sono state ottenute rispettivamente 256, 1127, 114 proteine.

3.6.3.1 Analisi comparativa dei genomi sequenziati

Sui 5 genomi completi è stato possibile effettuare studi comparativi al fine di capire meglio i meccanismi che regolano la risposta agli stress biotici negli organismi vegetali. I genomi presi in esame si differenziano tra loro per molti aspetti come, per esempio, la famiglia tassonomica dalla quale provengono, la grandezza del genoma e l'appartenenza alla divisione delle mono o di-cotiledoni. Al fine di poter effettuare un'analisi comparativa sui genomi in questione, tutti i proteomi sono stati raccolti in un unico file. In questo modo la normalizzazione dei dati ed i calcoli dell'errore sono stati uniformati, al fine di poter eseguire un confronto anche tra sequenze provenienti da genomi diversi. Ad ogni proteina analizzata è stato poi assegnato un numero con le informazioni di provenienza, così da dividere automaticamente i dati ottenuti. I genomi comparati sono stati *Arabidopsis thaliana*, *Vitis vinifera*, *Popolus trichocarpa*, *Sorghum bicolor* ed *Oryza sativa* ed in tabella 6 è possibile osservare i dati ottenuti, il numero di proteine putative per la funzione R rispetto alla grandezza dei proteomi analizzati, il numero di proteine che rispecchiano i canoni delle 4 classi R e la percentuale di sequenze che contengono nuove associazioni di domini. Tramite quest'analisi è stato possibile tracciare un profilo delle resistenza di ogni specie, che possono anche graficamente essere comparate tra di loro (figura 21). Tramite la comparazione dei genomi è possibile anche osservare il

numero di proteine che ricadono nelle singole classi, che confermano i dati ottenuti anche su altri pool genici e tramite il predittore DRAGO.

Nome Specie		*Arabidopsis thaliana*	*Vitis vinifera*	*Popolus trichocarpa*	*Oryza sativa*	*Sorgum bicolor*
Monocotiledoni					•	•
Dicotiledoni		•	•	•		
Dimensione genoma (Mb)		115	475	480	390	697
Dimensione proteoma (N° seq.)		32825	30434	50221	66710	34496
Classi Proteiche	CNL	61	326	248	-	-
	TNL	88	96	52	-	-
	RLP	118	278	383	-	-
	RLK	248	277	324	-	-
Proteine con singoli domini o nuove associazioni		587	1067	1354	-	-
Proteine con strutture tipiche delle classi R		515	977	1007		
Totale (significatività elaborata con chi-quadro)		1102[a]	2044[b]	2361[c]	3089[d]	1975[e]
% di proteine putative R rispetto all'intero proteoma		3,36	6,71	4,7	4,63	5,72
% di proteine con nuove associazioni rispetto al set predetto		53,2	52,2	57,3	-	-

Tabella 6. Informazioni generali sui genomi completamente sequenziati e risultati dell'analisi di predizione di MATRIX

La figura 20 mostra la distribuzione dei geni delle singole classi rispetto alla grandezza del proteoma di ognuno di essi. Da questa distribuzione risulta evidente che la classe TNL in tutte le specie è la meno presente, mentre il genoma di *Vitis* è l'unico che ha una forte incidenza di geni CNL. I geni della classe RLK ed RLP sono presenti in numero

maggiore rispetto alle altre classi, con una particolare propensione di *Arabidopsis* ad avere un gran numero di geni RLK.

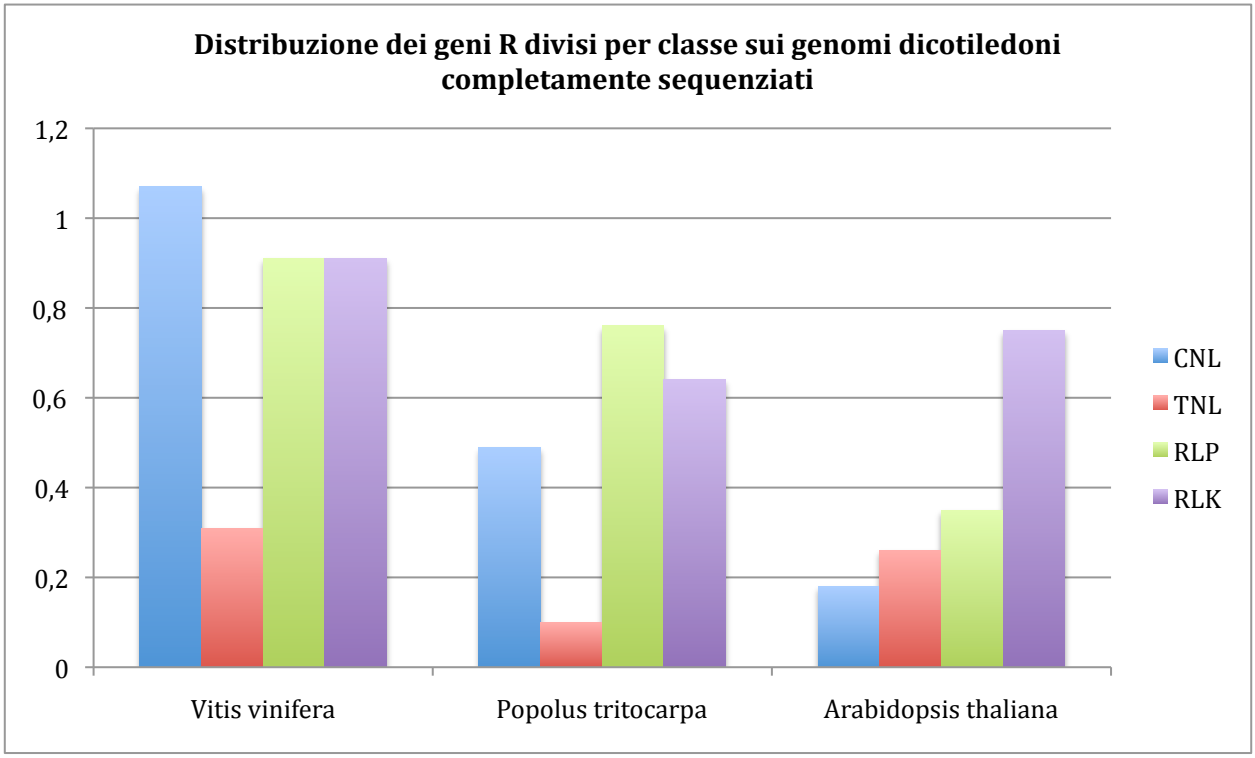


Figura 20. Distribuzione dei geni R divisi per classe sui genomi dicotiledoni completamente sequenziati

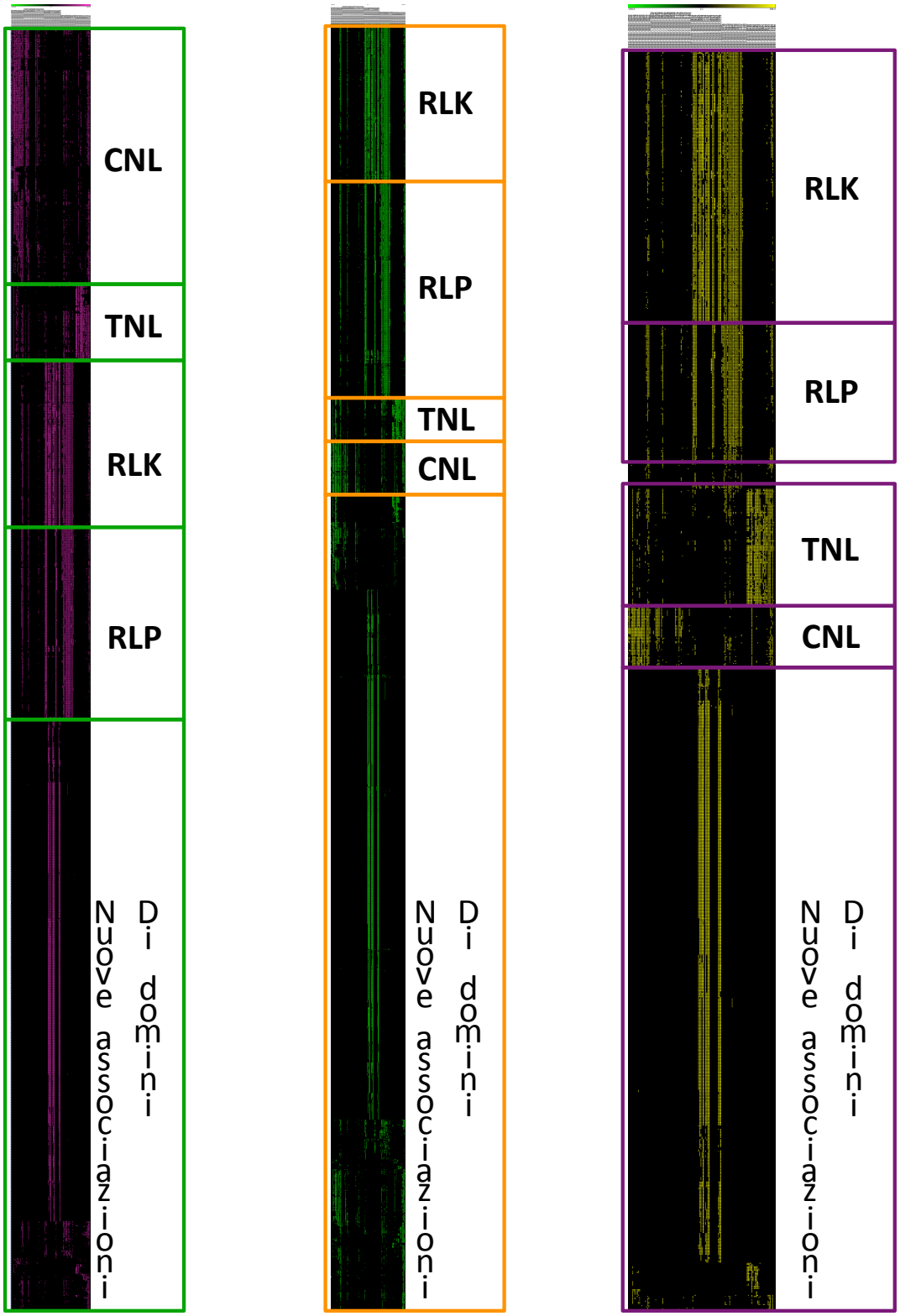


Figura 21. Profili di resistenza per i tre genomi dicotiledoni completamente sequenziati ottenuti tramite l'utilizzo di MATRIX

3.7 Utilizzo dei dati prodotti tramite MATRIX per uno studio approfondito del genoma del pomodoro

Tra i genomi analizzati tramite il sistema predittivo MATRIX, il pomodoro è stato scelto per effettuare studi di genetica e di genomica strutturale. Come è noto, il pomodoro è una delle specie da cui sono stati isolati e caratterizzati molti geni R. Della sua specie e della sua famiglia si conosco molte informazioni e la ricchezza dei dati presenti offre spunto per molti lavori a stampo evoluzionistico e molecolare. Approfondire le caratteristiche delle sequenze correlate alle resistenze presenti sul suo genoma può portare ad un ampliamento delle conoscenze ed aprire la strada a molte studi in diversi campi della biologia.

3.7.1 Classificazione delle sequenze predette

Dal genoma del pomodoro sono state predette 7126 proteine, di cui 256, analizzate tramite MATRIX, hanno mostrato caratteristiche interessanti. Questo set di sequenze, raggruppato in base alla presenza di domini conservati tipici delle proteine R, è stato diviso in 11 sotto-cluster ognuno con caratteristiche ed omologie specifiche con le proteine di referenza. In base all'omologia con i geni R funzionali isolati nelle Solanaceae, le 4 classi note sono state divise in sotto-classi al fine di produrre dati più chiari e specifici. Le 256 proteine predette sono state così divise:

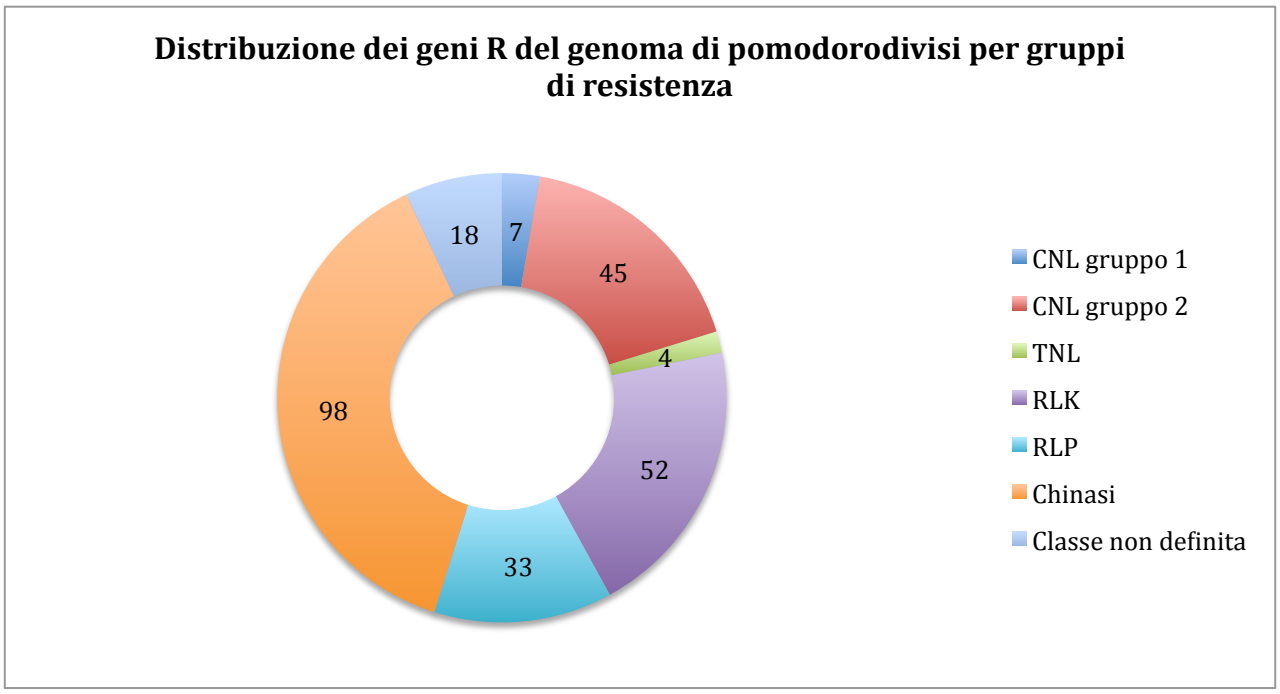


Figura 22 Distribuzione dei geni R del genoma di pomodoro divisi per gruppi di resistenza

Per ogni classe è stata effettuata un'analisi proteica tramite InterproScan per confermare la presenza dei domini predetti. Oltre alle 4 classi tipiche della famiglia delle proteine R, il genoma di pomodoro presenta anche interessanti sequenze composte da nuove associazioni di domini o da singoli domini la cui classe è illustrata come "non definita" (figura 22).

3.7.1.1 Le sequenze della classe non definita

Grazie ai dati ottenuti tramite DRAGO, è ormai chiaro che il panorama delle sequenze contenenti i domini, che danno funzionalità come proteine di resistenza, non si ferma alle sole 4 classi descritte in letteratura, ma presenta associazioni nuove ancora non descritte. In tutti i genomi analizzati sono presenti sequenze che non sono state raggruppate per via della loro struttura. Nel genoma di pomodoro sono state ritrovate 18 sequenze, il 7% dell'intero set predetto, con caratteristiche non comuni alle 4 classi dei geni di resistenza. Di seguito è possibile osservare la composizione delle singole proteine di questo set di dati composto da 7 proteine con il singolo dominio LRR, 9 proteine con il singolo dominio NBS, 1 proteina RLP dal dominio troncato ed un'interessantissima proteina con i tipici domini della classe CNL recante all'estremità N-terminale il dominio specifico RPW8, ritrovato nell'omonimo gene di *Arabidopsis thaliana* recante resistenza alla malattia "powdery mildew", portata da *Golovinomyces cichoracearum* (figura 23).

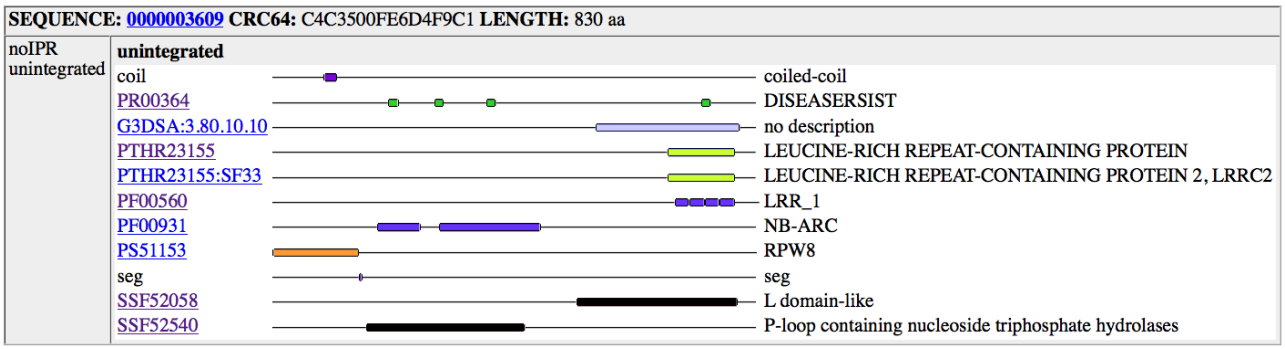


Figura 23. Esempio di proteina con una struttura proteica non definita

3.7.2 Creazione della mappa delle resistenza di pomodoro e localizzazione dei geni R

Le proteine predette così classificate sono state inserite in un database ad hoc ed i dati di sequenza sono stati arricchiti di tutte le informazioni provenienti dal progetto di sequenziamento del genoma del pomodoro. Per ogni sequenza sono state inserite informazioni su:

- Classe predetta
- Sequenza genica
- Contig di provenienza
- Coordinate genetiche del gene
- Cromosoma di provenienza

A queste informazioni, già di per sè interessanti, ne è stata aggiunta un'altra: sfruttando i dati già prodotti dall'ente SGN, SOL Genomic Network, sono stati raccolti tutti i marcatori COSII provenienti dal pomodoro e le loro sequenze sono state ancorate sui contig sequenziati. Con questo panorama informativo è stata creata una mappa fisica del genoma del pomodoro dove sono state visualizzate tutte le sequenze, correlate alla funzione di resistenza, e tutti marcatori COSII presenti sul genoma. Questa mappa, che integra i dati fisici del sequenziamento con i dati genetici dei marcatori, permette di visualizzare non solo la posizione delle sequenze relative ai geni R, ma anche la loro classe ed i marcatori COSII associati. Una mappa così, denominata "mappa R", unica nel suo genere, è stata resa disponibile (figura 24). Per le figure dei singoli cromosomi è possibile consultare l'appendice A.

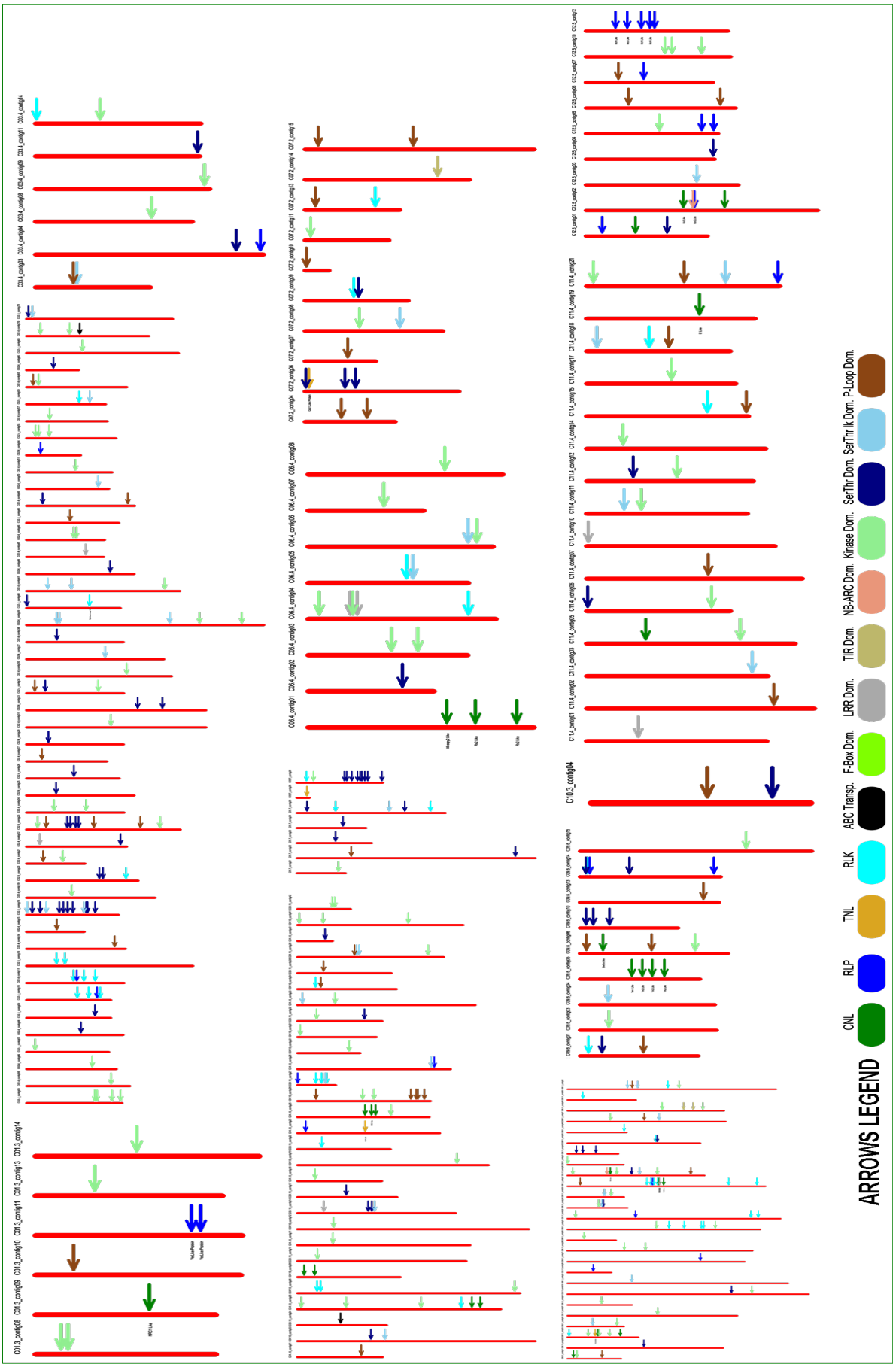


Figura 24. La mappa R del genoma del pomodoro

3.7.3 Analisi degli “Hot Spot” e della disposizione dei geni R

Tra le grandi potenzialità della “mappa R”, vi è quella di poter avere una visione globale del panorama genico di resistenza. Grazie a questa caratteristica è stato possibile portare avanti uno studio strutturale a livello dei cluster genici presenti sul genoma e capire se la distribuzione dei geni R tra i cromosomi è uniforme o meno. La figura 25 rappresenta il numero di geni divisi per cromosomi, tenendo presente il numero di contig sequenziati.

Com’è possibile notare, la distribuzione delle sequenze sui cromosomi non è uniforme ed i cromosomi ricchi di geni putativi per la funzione di resistenza sono il 2, 5, 6, 8, 9, 12. Inoltre, tramite la figura 26, è possibile approfondire la localizzazione delle singole sequenze rispetto ai contig ed affermare che le sequenze correlate alle resistenze sono in maggior numero strutturate in cluster e non in singole sequenze.

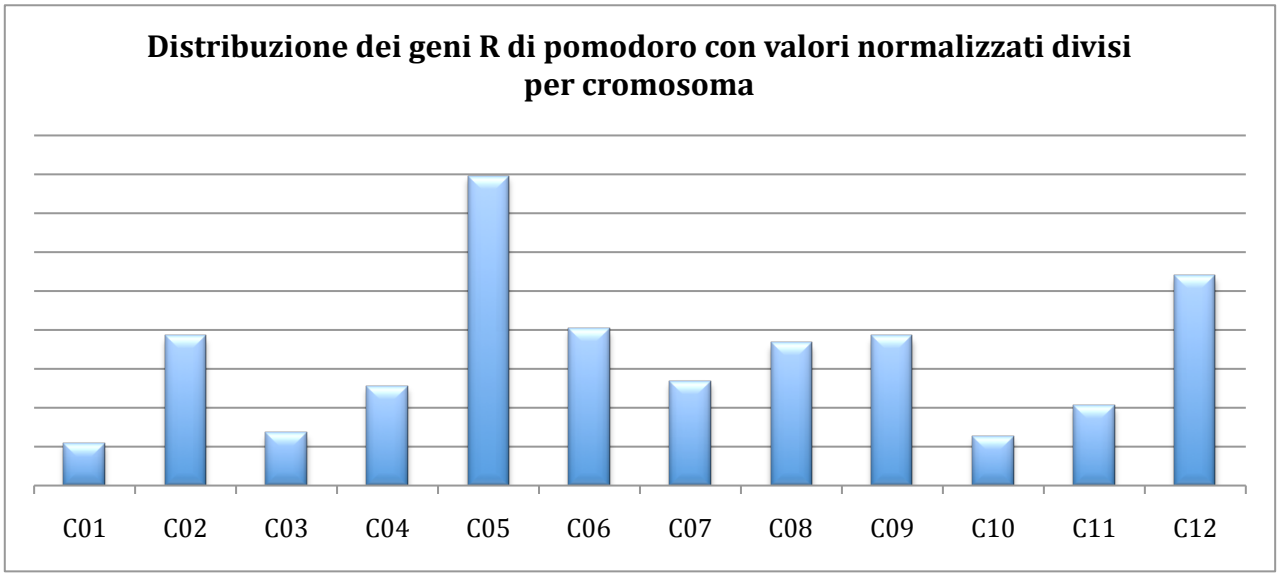


Figura 25. Distribuzione dei geni R di pomodoro con valori normalizzati divisi per cromosoma

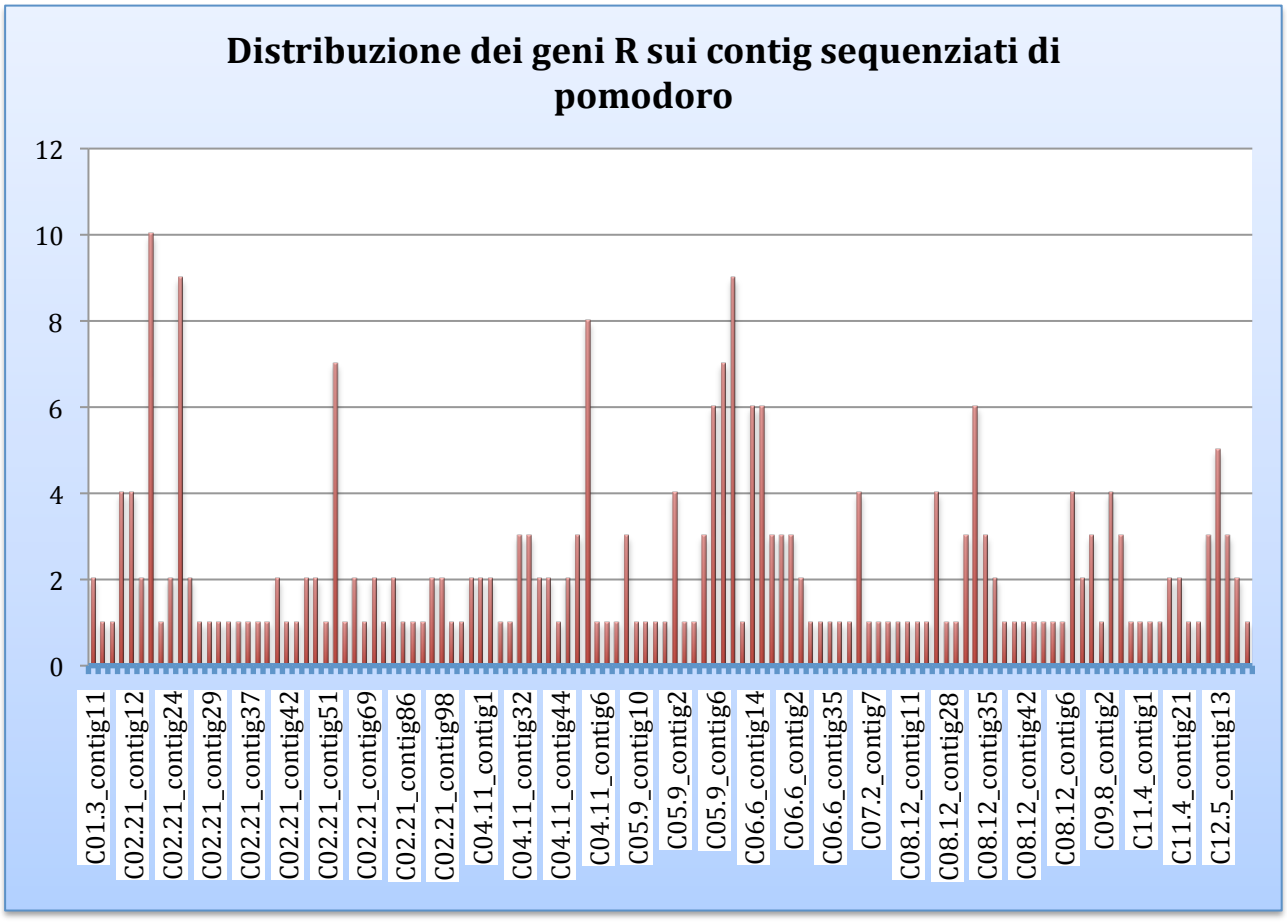


Figura 26. Distribuzione dei geni R sui contig sequenziati di pomodoro

Su 256 sequenze, i geni organizzati in cluster sono risultati il 72,7% mentre la percentuale di geni non clusterizzati è del 27,2%. La figura 26 mostra il numero di geni presenti nei diversi contig, evidenziando che i geni posizionati in modo singolo sono 69, contro i 187 posizionati in modo ravvicinato gli uni agli altri, e che la grandezza dei cluster, misurata per numero di geni posizionati in regioni adiacenti, può variare da un minimo di 2 geni ad un massimo di 10.

Le diverse classi in cui sono state divise le sequenze seguono andamenti diversi, sia per quanto riguarda la loro posizione sul genoma, sia per quanto riguarda il loro raggruppamento. I 4 geni della classe TNL si trovano su 4 cromosomi diversi e sono quindi rari e singoli. Invece, per i geni appartenenti alla classe CNL, RLP, RLK è stato notato non solo che sono distribuiti in modo non uniforme sul genoma, ma che possono disporsi in cluster costituiti da 2 a 6 geni (figura 27).

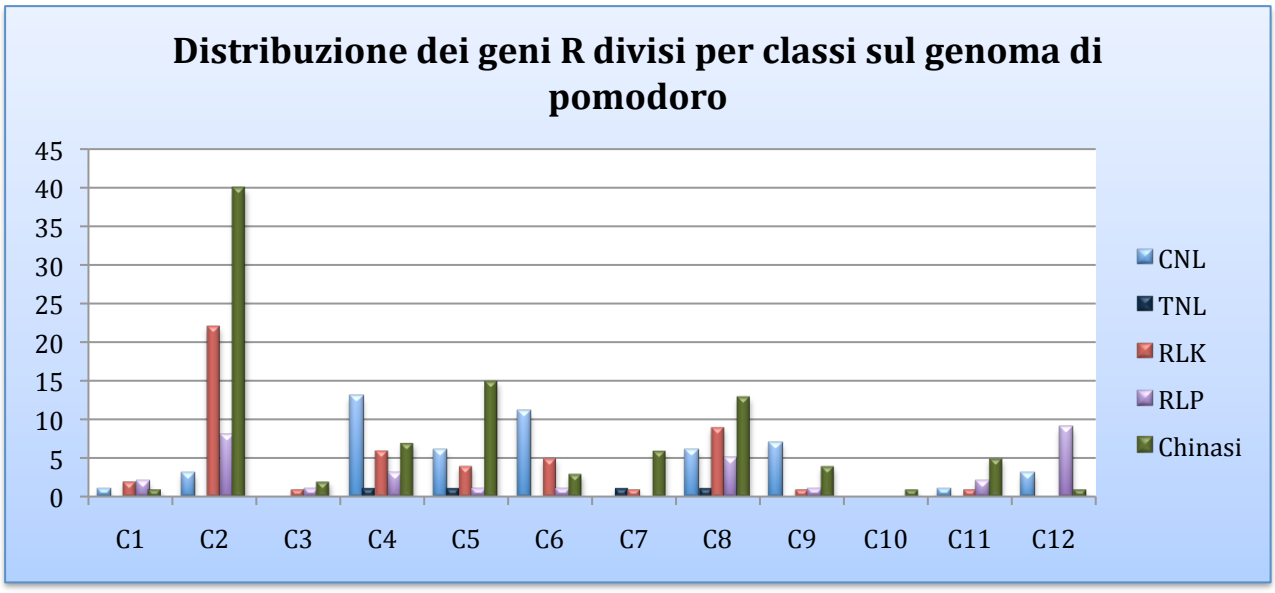
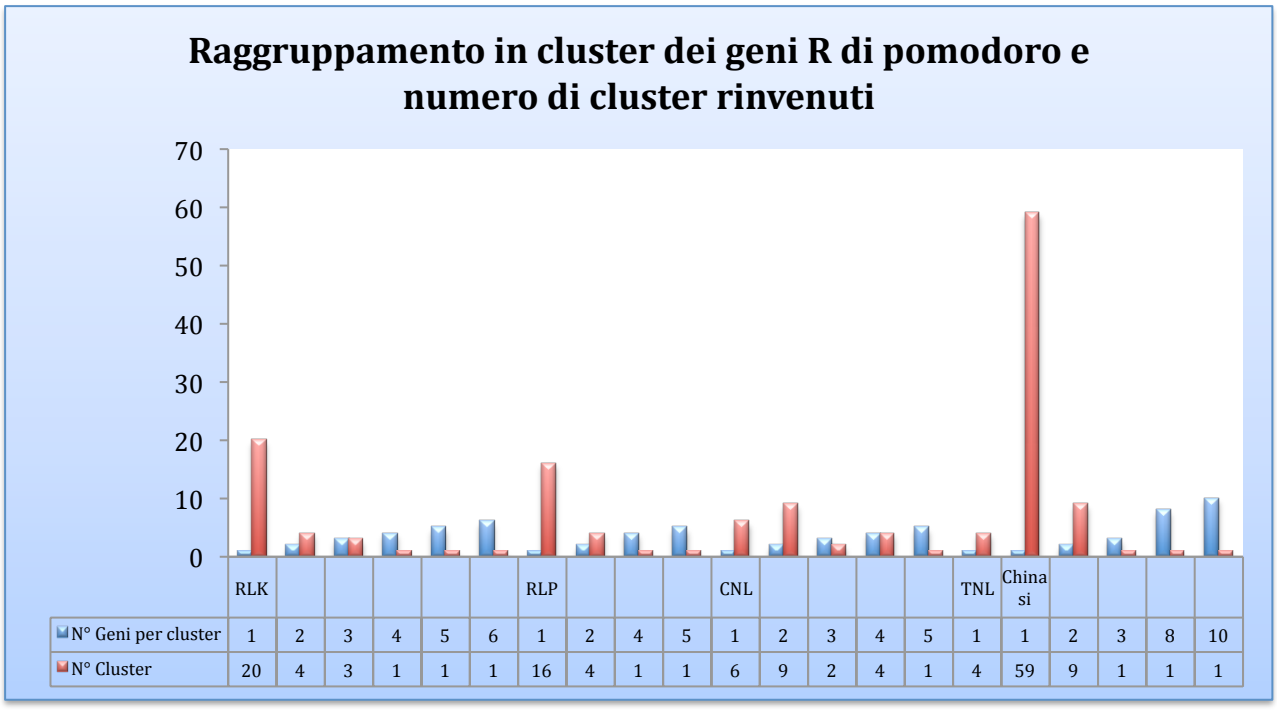


Figura 27. Distribuzione dei geni R divisi per classi sul genoma di pomodoro

Dalla figura 27 è facile notare come alcuni cromosomi sono deputati a classi R specifiche ed altri sono totalmente privi di geni R o di alcune classi. Il cromosoma 12 è ricco di geni della classe RLP, che risulta invece rara in tutti gli altri cromosomi, se non per alcuni geni sui cromosomi 1, 2, 4 e 8. I geni della classe CNL si trovano nella regione centrale del genoma, in particolare sui cromosomi 4, 5, 6, 8, 9 e 12, mentre i geni delle classi RLK e chinasi presenti in numero maggiore, raggiungono i loro picchi di presenza sui cromosomi 2,4,5,6 ed 8. Anche per quanto riguarda la concentrazione dei geni appartenenti alle diverse classi, è doveroso far notare come essi hanno caratteristiche diverse. I geni della classe TNL sono disposti in modo singolo, mentre quelli della classe delle chinasi sono strutturati in cluster anche di 10 geni. I geni della classe RLK possono arrivare ad essere organizzati in cluster di 6 sequenze, mentre quelli della classe RLP si organizzano in cluster composti al massimo di 5 geni (figura 28).



	RLK						RLP				CNL					TNL	Chinasi				
N° Geni per cluster	1	2	3	4	5	6	1	2	4	5	1	2	3	4	5	1	1	2	3	8	10
N° Cluster	20	4	3	1	1	1	16	4	1	1	6	9	2	4	1	4	59	9	1	1	1

Figura 28. Raggruppamento in cluster dei geni R di pomodoro e numero di cluster rinvenuti

3.7.4 Analisi delle duplicazioni delle sequenze

Il set di dati predetti mostra 16 coppie di geni duplicati. Le regioni interessate da questo fenomeno sono presenti sui cromosomi 2,4,5,6 e 11. Sul cromosoma 2 sono stati trovati 11 geni, di cui 10 identici sullo stesso cromosoma, ed uno identico ad un gene posizionato sul cromosoma 4. I cromosomi 5 e 11 presentano rispettivamente una sola duplicazione in regioni molto vicine tra di esse. Molto interessante è il caso del cromosoma 6, dove le coppie duplicate sono 8. Nella regione compresa tra il contig 14 e 15, si trovano 5 duplicazioni e nella regione tra il contig 16 e 17 altre 3. Un'analisi approfondita ha rivelato che la duplicazione è avvenuta su due interi cluster genici, composti rispettivamente da 5 e 3 geni identici ed altri paraloghi. Poiché il cluster è stato duplicato invertendo il suo orientamento, è possibile parlare di duplicazione tandem invertita (figura 29).

Tutti i geni duplicati sono mostrati in tabella 7, insieme ai cromosomi interessati dalla duplicazione ed alla tipologia di sequenza.

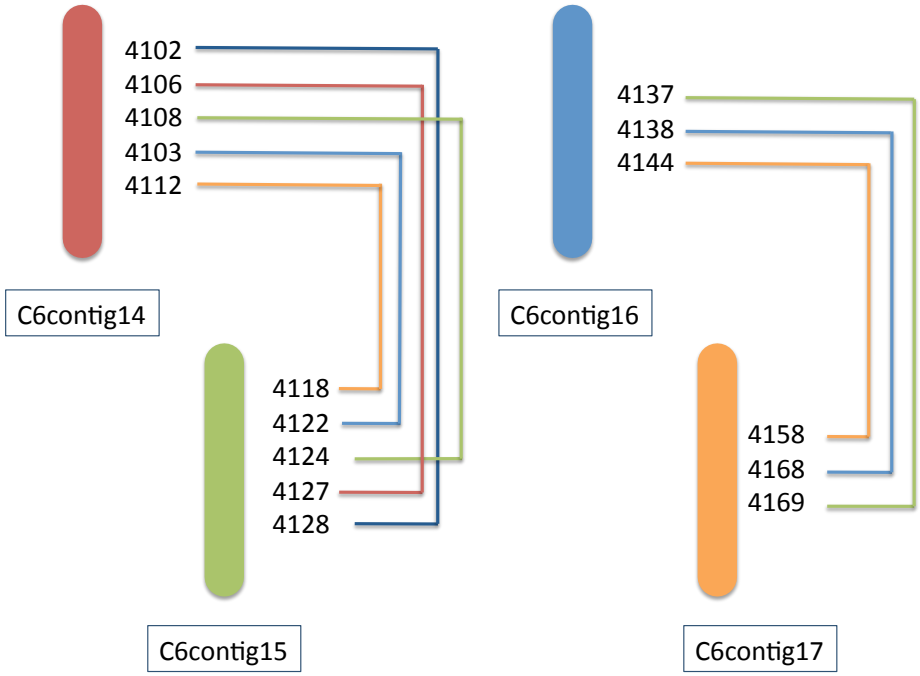


Figura 29. Rappresentazione grafica di due eventi di duplicazione sul cromosoma 6 di pomodoro

Nome Contig	Nome Gene		Nome Contig	Nome Gene	Classe genica
C11conting18	6662	→	C11conting21	6732	Chinasi
C2conting29	806	→	C2conting88	1878	Chinasi
C2conting32	859	→	C2conting34	920	Chinasi
C2conting40	1073	→	C2conting41	1103	Chinasi
C2conting63	1731	→	C2conting79	1513	Chinasi
C2conting9	433	→	C2conting10	2140	Chinasi
C4conting18	751	→	C2conting21	2583	CNL
C5conting5	3917	→	C5conting5	3930	RLK
C6conting14	4102	→	C6conting15	4128	CNL
C6conting14	4106	→	C6conting15	4124	CNL
C6conting14	4108	→	C6conting15	4122	CNL
C6conting14	4103	→	C6conting15	4127	CNL
C6conting14	4112	→	C6conting15	4118	CNL
C6conting16	4137	→	C6conting17	4169	CNL
C6conting16	4138	→	C6conting17	4168	CNL
C6conting16	4144	→	C6conting17	4158	RLK

Tabella 7. Elenco di tutti i geni R di pomodoro che hanno subito processi di duplicazione

3.7.5 Analisi dell'omologia delle sequenze appartenenti ai cluster genici

È noto che i geni R clusterizzati hanno spesso una forte omologia di sequenza. Grazie ai dati prodotti sul genoma di pomodoro, è possibile effettuare studi d'omologia di sequenza su tutti i geni predetti divisi in classi. Attraverso quest'approccio è possibile esaminare l'omologia delle diverse sequenze in relazione alla loro localizzazione. Con analisi filogenetiche, è stato possibile portare al luce lo stretto legame che sussiste tra l'omologia delle sequenze appartenenti al genoma di pomodoro e la loro localizzazione. Le analisi sono state effettuate per le classi CNL, RLP e RLK ed in tutte e tre è stato possibile osservare la disposizione ed il raggruppamento delle proteine afferenti ai diversi cluster genici (figura 30, 31, 32). L'attenta osservazione degli alberi prodotti mostra che esiste una correlazione tra la posizione genomica e l'omologia delle proteine, che molto spesso le sequenze adiacenti sono omologhe e che per alcuni cluster sono avvenuti fenomeni di transizione e duplicazione. Esistono anche casi in cui proteine distanti tra loro mostrano significativi livelli di omologia e casi, come l'analisi effettuata sulla classe RLP, dove l'omologia delle proteine risulta di molto inferiore rispetto alle altre classi R.

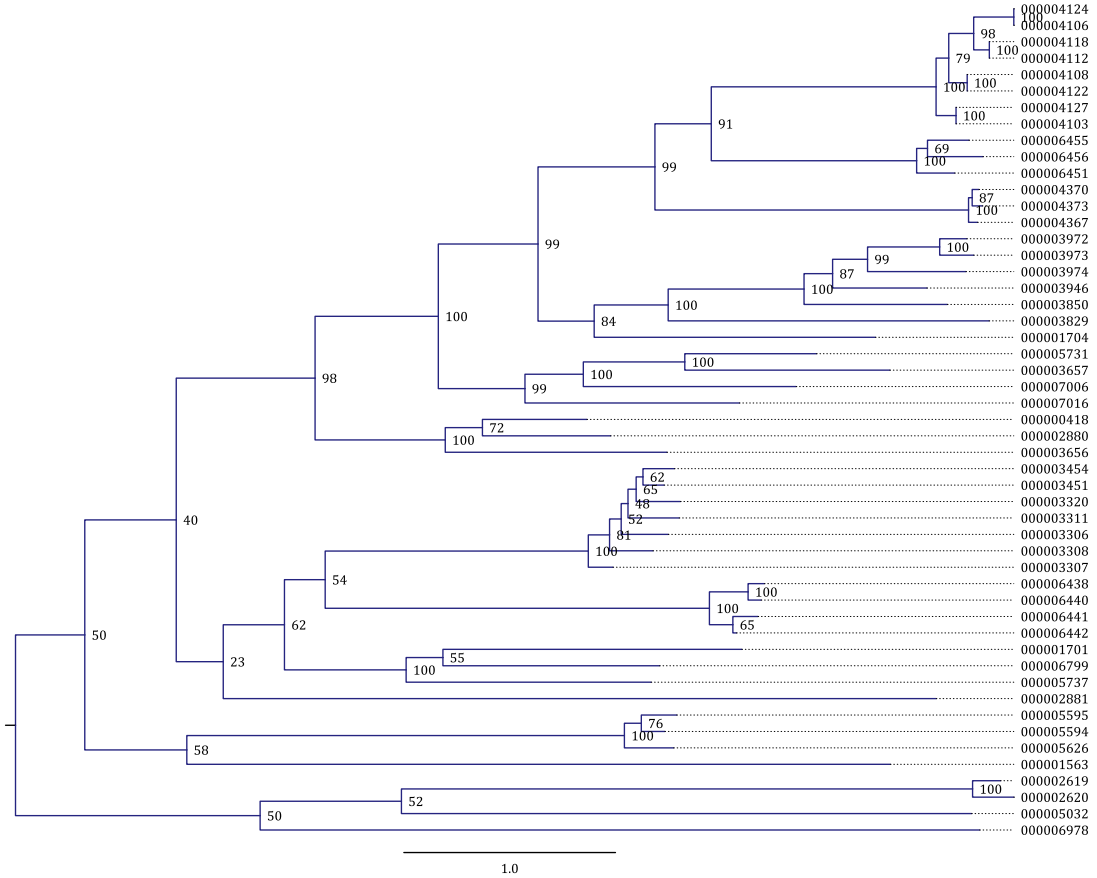


Figura 30. Albero filogenetico dei geni della classe CNL di pomodoro

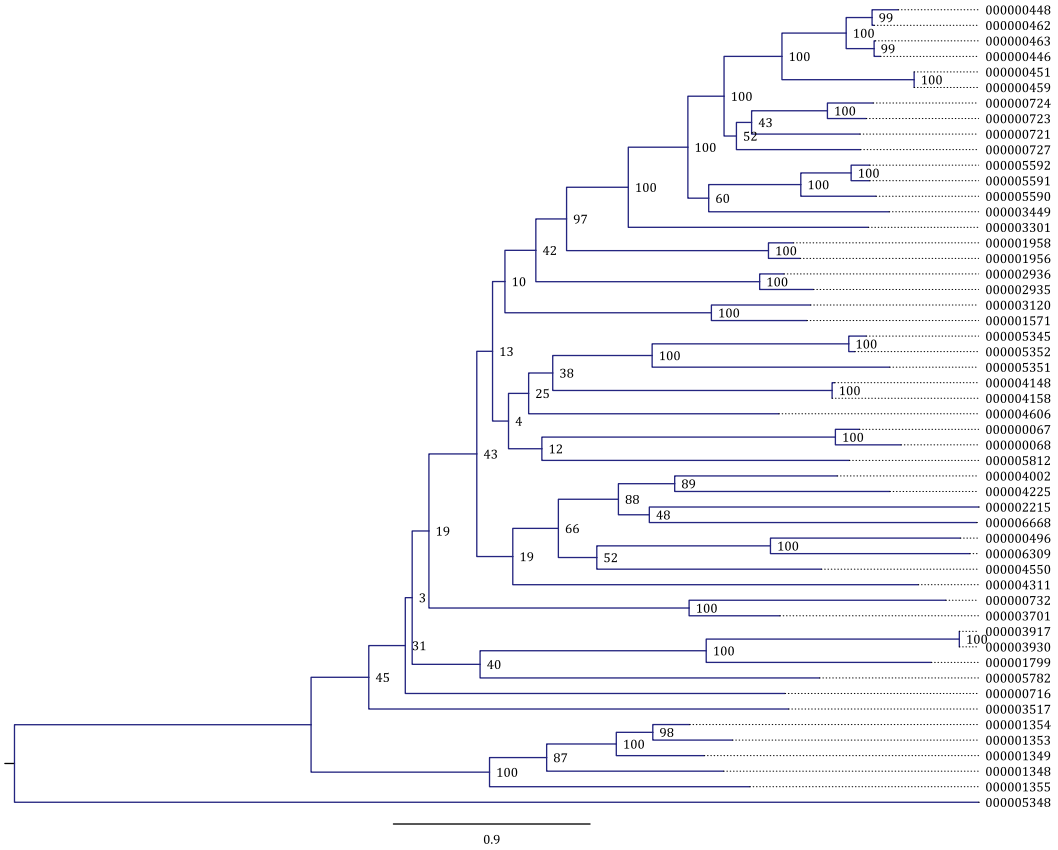


Figura 31. Albero filogenetico dei geni della classe RLK di pomodoro

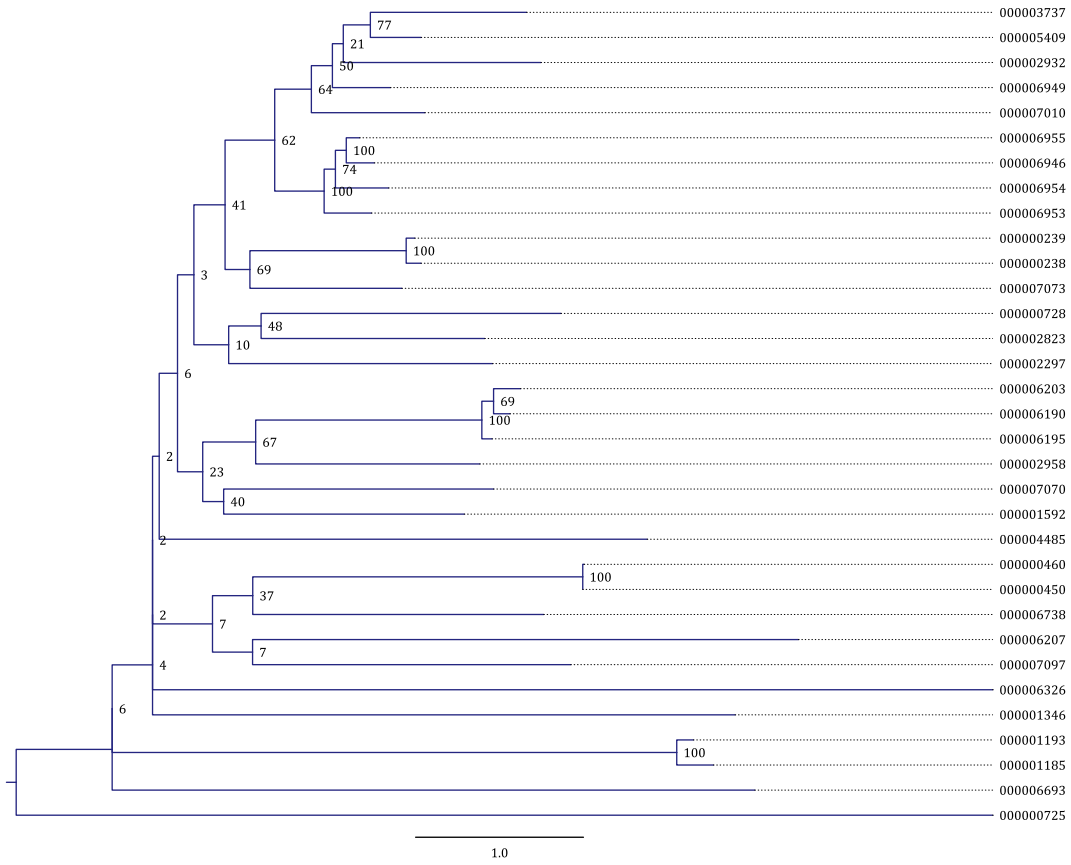
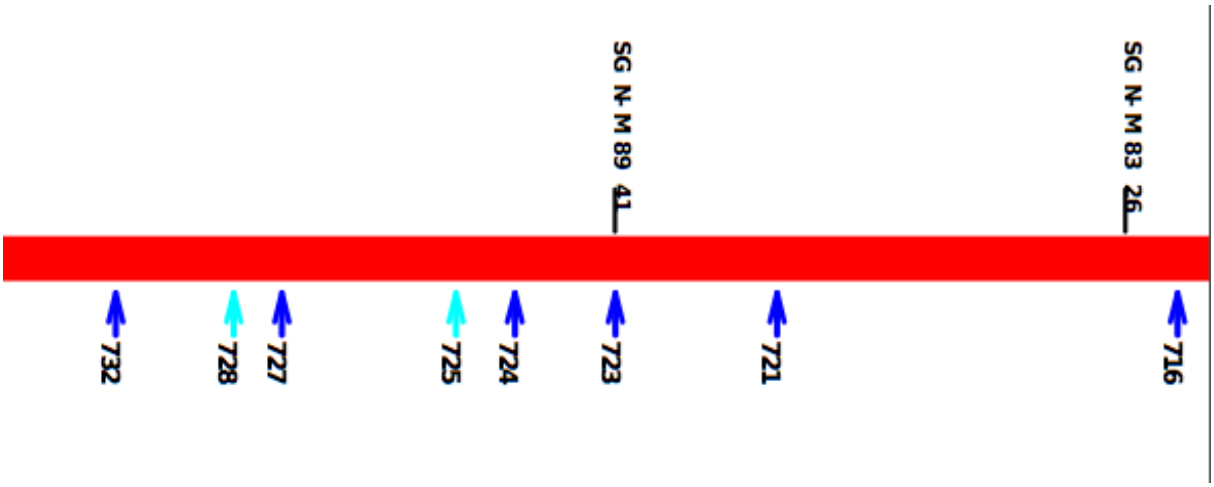


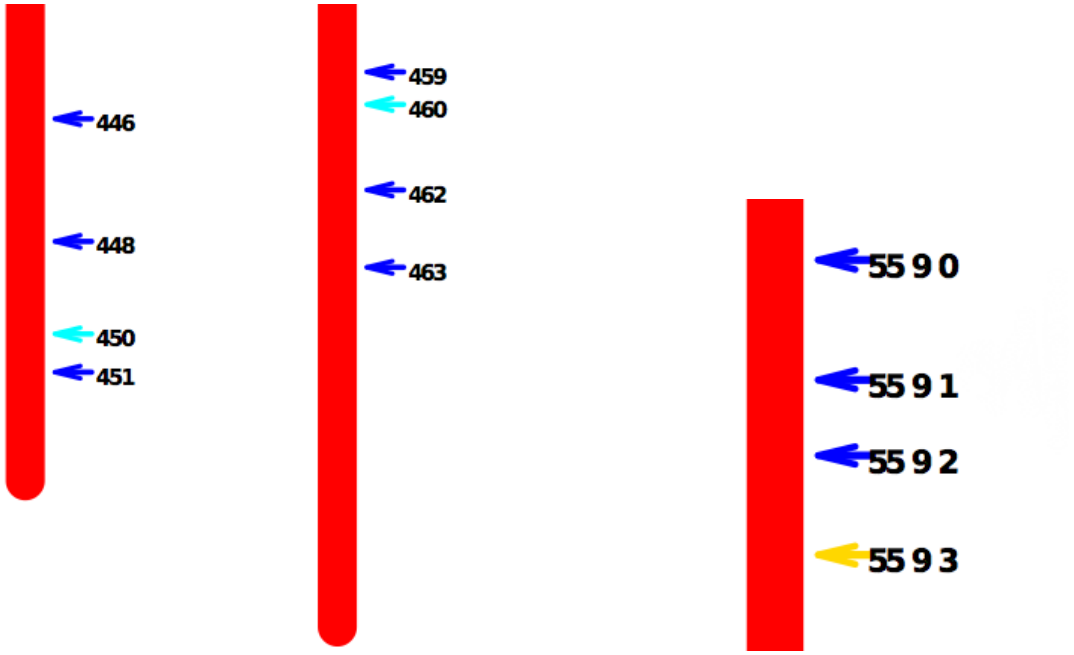
Figura 32. Albero filogenetico dei geni della classe RLP di pomodoro

3.7.6 Analisi dei cluster genici composti da geni afferenti a classi diverse

Un'altra interessante caratteristica, riscontrata analizzando il genoma del pomodoro, tramite la mappa R, è che ci sono alcuni cluster composti da geni provenienti da classi diverse. Sono un chiaro esempio i due cluster presenti sul cromosoma 2, sui contig 11 e 12, composti da tre geni della classe RLK ed uno della classe RLP; un cluster del cromosoma due composto da 6 geni RLK e 2 RLP sul contig 25; un cluster presente sul cromosoma 8 contig 33 composto da 3 geni RLK ed una chinasi. Come è possibile osservare dalla figura 33, anche la disposizione dei cluster e dei geni ad essi afferenti non sembra casuale.



Cromosoma 2 contig 25



Cromosoma 2 contig 11-12

Cromosoma 8 contig 33

RLK

RLP

Chinasi

Figura 33. Cluster genici rinvenuti sul genoma di pomodoro composti da geni afferenti a classi R differenti

3.7.7 Analisi della classe RLK di pomodoro

I geni della famiglia RLK hanno la caratteristica di essere recettori ad ampio spettro, che riconoscono elicitori più generici e trasducono il segnale di difesa, attraverso un dominio chinasico. Grazie agli strumenti utilizzati in questo lavoro di tesi, nei set di sequenze analizzati, sono state trovate interessanti informazioni. Il genoma del pomodoro presenta 52 sequenze della classe RLK, presenti sul cromosoma 2 ed 8 rispettivamente ed a seguire sui cromosomi 4, 5 e 6. Dai paragrafi precedenti è possibile notare che questa classe genica spesso clusterizza sia tra gli stessi geni RLK, sia tra geni di classi diverse. Inoltre tre coppie geniche risultano duplicate e due geni sono risultati altamente omologhi ai geni di *Arabidopsis FLS2* e *EFR*, entrambi implicati nel riconoscimento batterico rispettivamente della flagellina e del fattore di elongazione. La filogenesi ha mostrato che esistono diversi raggruppamenti di sequenze afferenti alla classe RLK, ma con caratteristiche diverse, come ad esempio la presenza o meno del dominio LRR e la tipologia del dominio chinasico. Da questi risultati preliminari si evince che la classe RLK è presente anche nel genoma di pomodoro e delle *Solanaceae,* aprendo la strada ad interessanti ipotesi.

3.8. Analisi filogenetica delle sequenze appartenenti al genere *Solanum*

Attraverso l'utilizzo del sistema di predizione MATRIX, sono state analizzate molte sequenze e sono state estrapolate diverse informazioni utili dai dati grezzi provenienti da esperimenti di genomica su larga scala. Le sequenze provenienti dal genere *Solanum*, ovvero le circa 20000 proteine collezionate da NCBI e le circa 7000 ottenute tramite il sequenziamento del genoma del pomodoro, sono state oggetto di approfondimento attraverso l'uso di un'ampia varietà di sistemi informatici. Di particolare importanza, dopo l'assegnazione di una putativa funzione ad una proteina, è la sua caratterizzazione mediante il confronto con altre sequenze, al fine di ottenere interessanti dati sull'interconnessione delle proteine all'interno delle specie, famiglie e generi a cui appartengono. Per fare ciò è stato creato un set di dati, contenente tutte le proteine predette dalla famiglia delle *Solanaceae* di NCBI, le proteine predette sul genoma di pomodoro ed i già clonati geni di resistenza. Il pool proteico così composto è stato analizzato e raggruppato tramite MATRIX, al fine di associare le proteine provenienti da tre set diversi e separarli in base alle caratteristiche di sequenza, per poterne studiare la

struttura, la correlazione e l'omologia. L'esperimento ha raggruppato le sequenze in 12 gruppi che, con la guida delle proteine R di referenza, sono stati divisi come segue:

Nome gruppo	N° proteine	Omologia con proteine R
CNL 1	91	*rpi-blb1, rps1-k-1, rps1-k-2, R3a, I2, rps5, rps2, Rpm1, rpp13, rpp8, rcy1, hrt, pi-ta, mla10*
CNL 2.1	24	*Bs2, tm2a, tm2-2, Gpa2, rx1,rx2*
CNL 2.2	95	*Ri, Prf, Sw5a, Hero, Mi1, rpi-blb2,*
TNL	32	*P2, N, Bs4, Ry1, L6, M, rps4, rpp5, gro1.4*
RLP 1	98	*Ve1, Ve2, Cf9b, Cf4a, Cf2, LeEIX1, Le EIX2, Cf9, Cf4*
RLP 2	196	*Pgip*
RLK	78	*Pepr1, fls2, efr, Xa21, Er,*
KIN 1	25	-
KIN 2	48	*Pto*
KIN 3	135	*Rpg1, Rfo*
KIN 4	45	-
Non Definito	37	-

Tabella 8. Numero di sequenze della famiglia delle *Solanaceae* divise in gruppi di resistenza

Per ogni gruppo è stata effettuata un'accurata analisi filogenetica, e le sequenze presenti negli alberi ottenuti sono state, tramite lo *script* realizzato appositamente, etichettate in modo da rendere ottimale la lettura dell'albero.

3.8.1 Ottenimento di alberi filogenetici delle putative proteine R

Gli alberi sono stati sottoposti ad una serie di modifiche grafiche, che non hanno influito in alcun modo sulla qualità dei dati prodotti. Per migliorarne la leggibilità, la provenienza del set di dati e le caratteristiche delle sequenze sono state evidenziate con colori diversi. Le proteine R funzionali sono state evidenziate aggiungendo le descrizioni sul patogeno e la malattia che bloccano, ed una completa descrizione è stata affiancata al nome di ogni sequenza putativa, in modo da poter risalire facilmente alla stessa. A causa della voluta ridondanza del set delle proteine di referenza e del set di proteine proveniente da NCBI, ad ogni proteina R ne sarà associata una con bootstrap di 100 che rappresenta la stessa proteina proveniente dal set NCBI. La corretta associazione delle proteine identiche, provenienti da diversi set, rappresentano un ulteriore controllo della corretta analisi performata da MATRIX e dai software di filogenesi. Per tutte le classi analizzate, la filogenesi ha confermato i raggruppamenti ottenuti impiegando MATRIX per l'analisi di predizione. Le proteine analizzate si dispongono in base all'omologia di sequenza sull'albero a prescindere dal set di appartenenza. Grazie a quest'analisi, per ogni gene di resistenza conosciuto, è stato possibile associare i propri omologhi conservati in NCBI ed i propri omologhi presenti sul genoma di pomodoro, osservandoli in rapporto con gli altri geni R. I rami prodotti dalla filogenesi si distinguono per specie da cui proviene la sequenza, creando così un panorama nel quale le *Solanaceae* hanno un ruolo principale, e poi, a seguire, *Arabidopsis* ed altre specie. Da quest'analisi è possibile capire anche quali sono i geni più o meno studiati nel corso del tempo, trovandosi spesso con rami contenenti decine di sequenze omologhe provenienti da molte specie diverse, oppure con pochi omologhi per gene di referenza. Il numero di sequenze omologhe provenienti da NCBI influenza spesso la disposizione delle sequenze di pomodoro, che a volte vengono posizionate più distanti, ma solo per una relativa minore omologia con i geni di referenza. Sia nel set dei geni predetti da pomodoro, sia nel set di NCBI, per tutti i geni di referenza è stato trovato il proprio corrispettivo con distanza tra le due sequenze di 0. Molto interessanti risultano le sequenze che si posizionano più lontano rispetto ai geni di referenza, in quanto hanno la stessa struttura ed un'elevata omologia con il gene di referenza, ma presentano quel giusto grado di polimorfismo, interessante da analizzare per la scoperta di geni con nuove funzioni di riconoscimento. Inoltre, le sequenze del genoma di pomodoro, che risultano omologhe tra loro in quanto parte di un cluster genico, sono altamente simili a geni di referenza presenti in altre

specie. Grazie a questa tipologia di analisi è stato possibile discriminare dalle sequenze presenti in NCBI omologhe ai geni di referenza tutte le sequenze tagliate o parziali che spesso confondono gli utenti (figure 34-39).

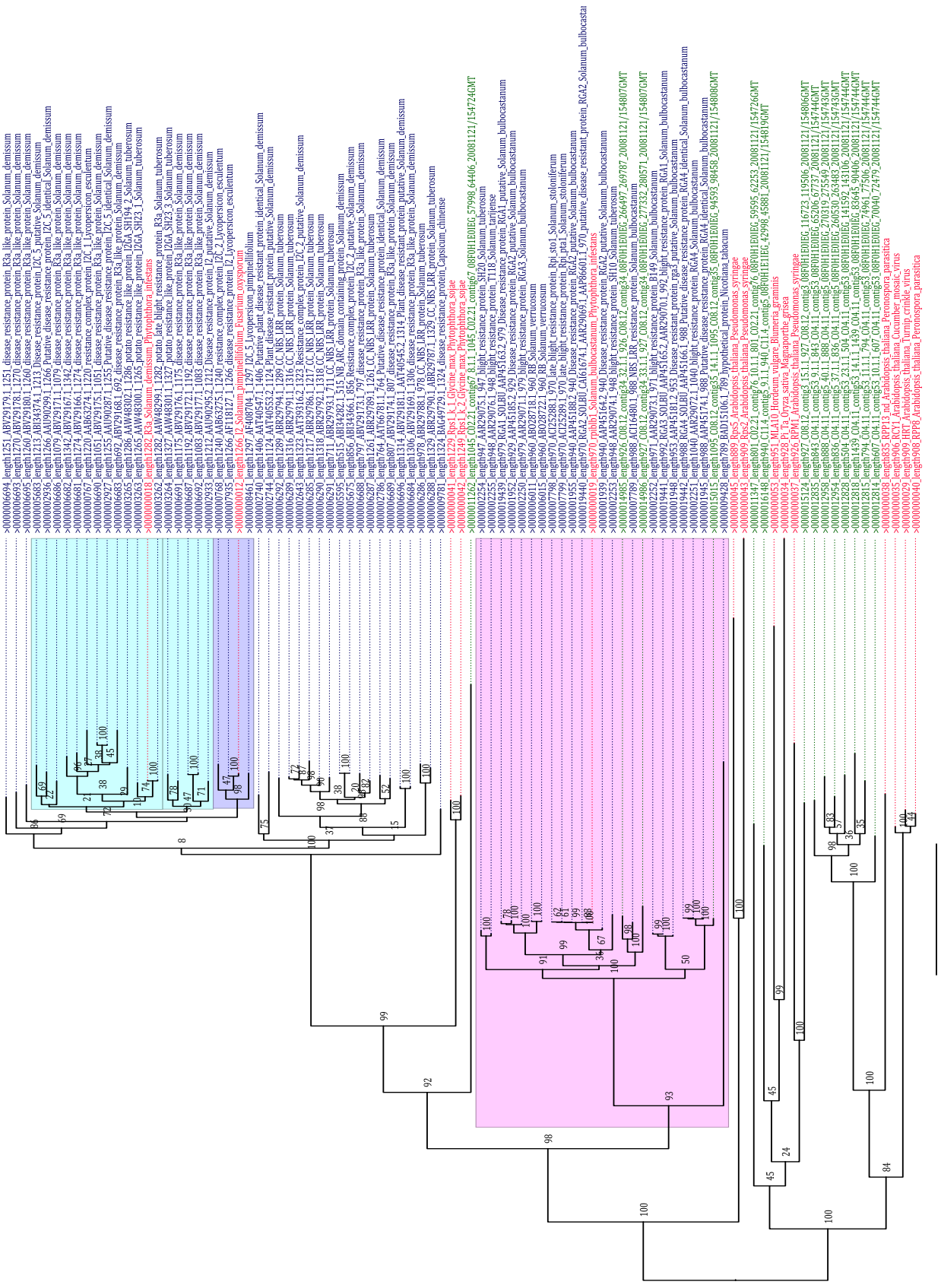


Figura 34. Filogenesi del gruppo di resistenza CNL1

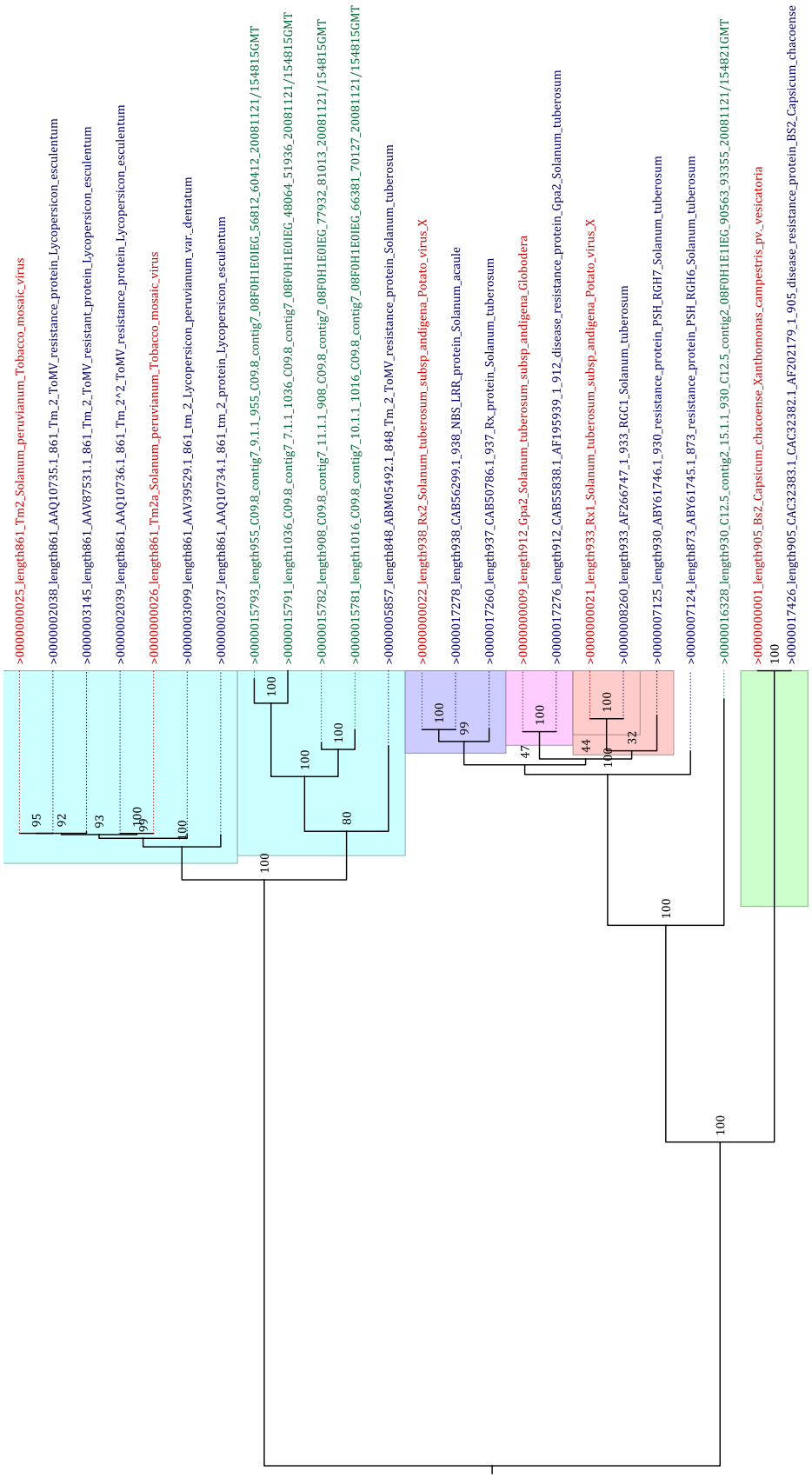


Figura 35. Filogenesi del gruppo di resistenza CNL2.1

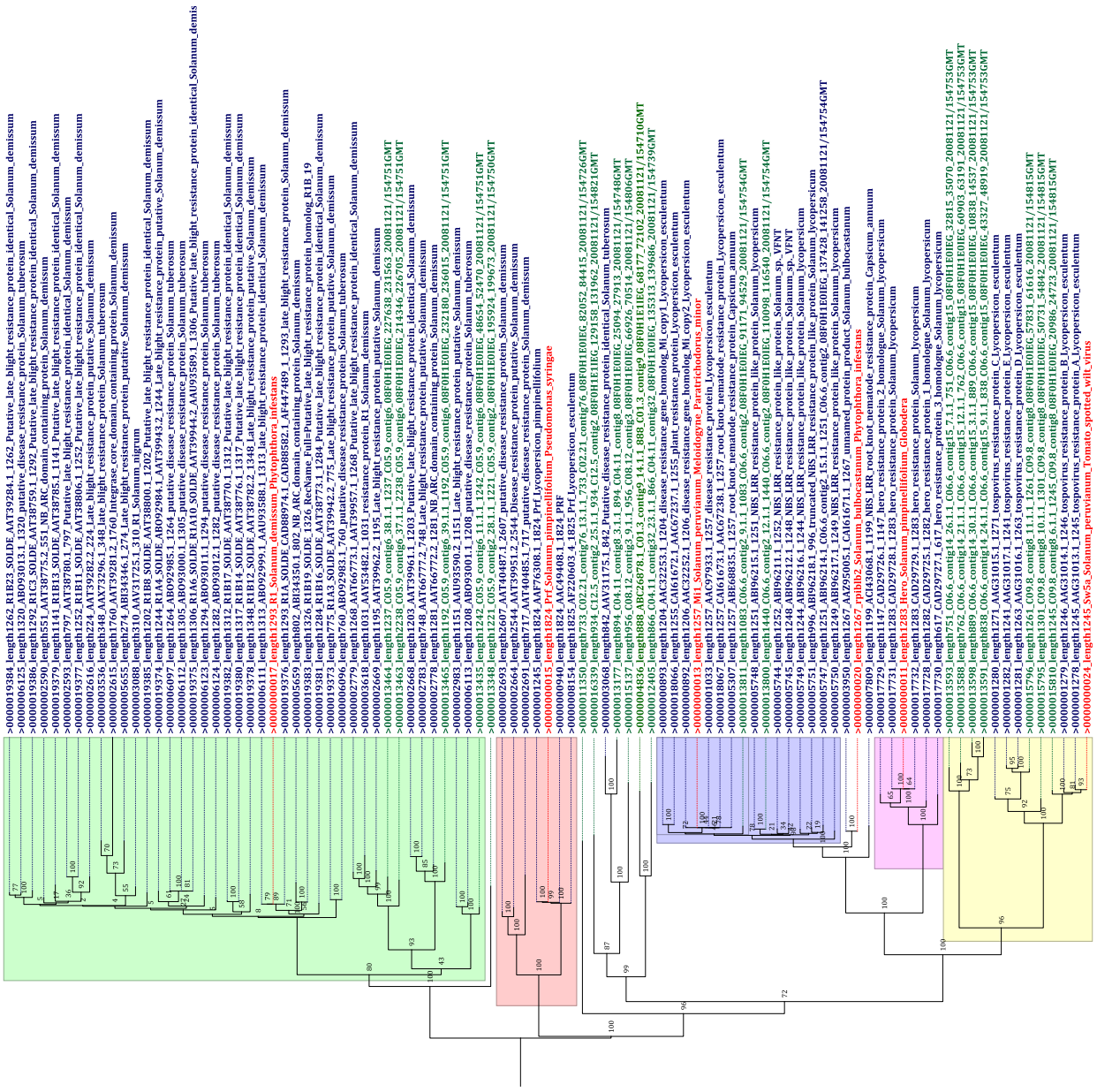


Figura 36. Filogenesi del gruppo di resistenza CNL2.2

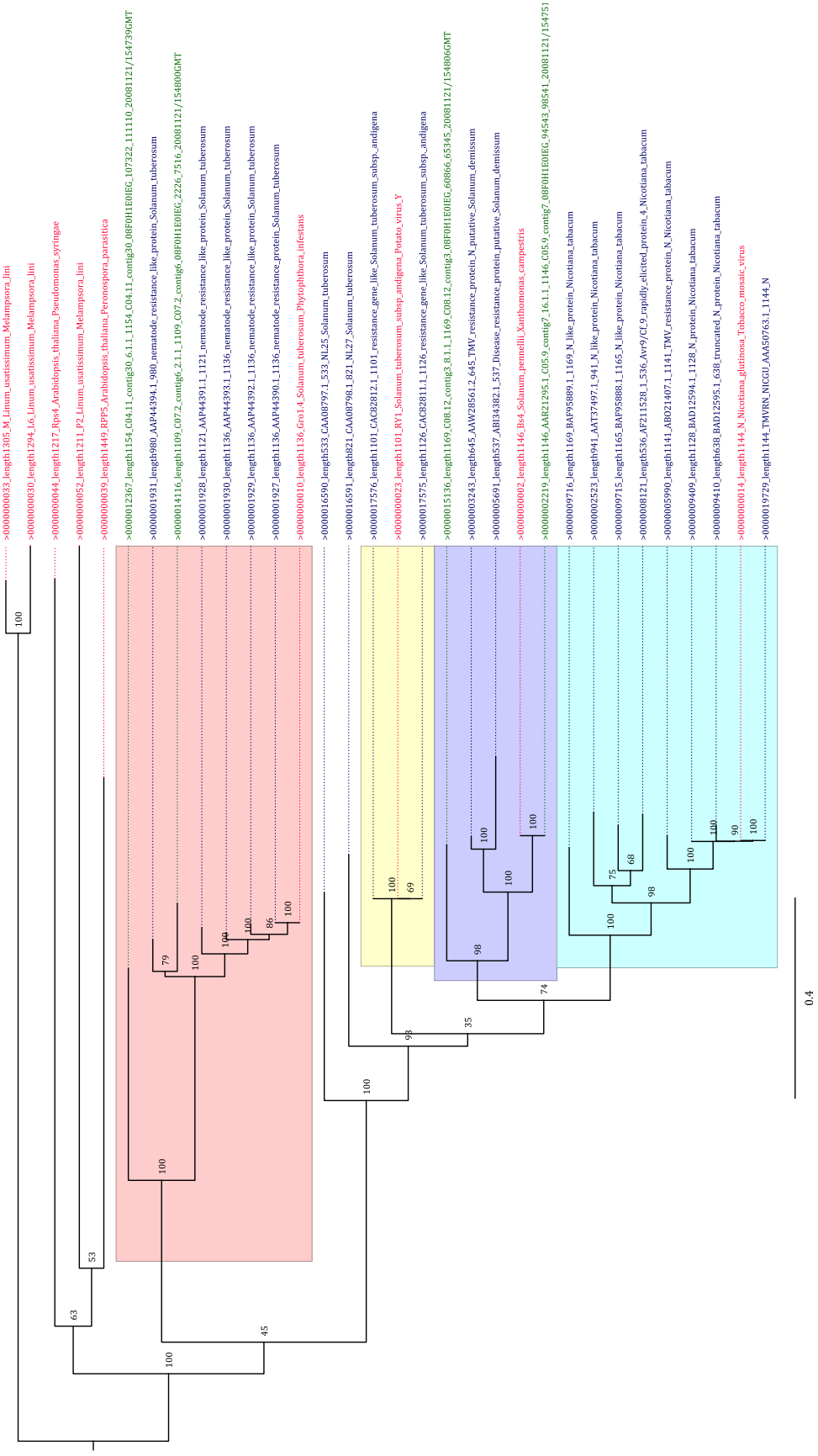


Figura 37. Filogenesi del gruppo di resistenza TNL

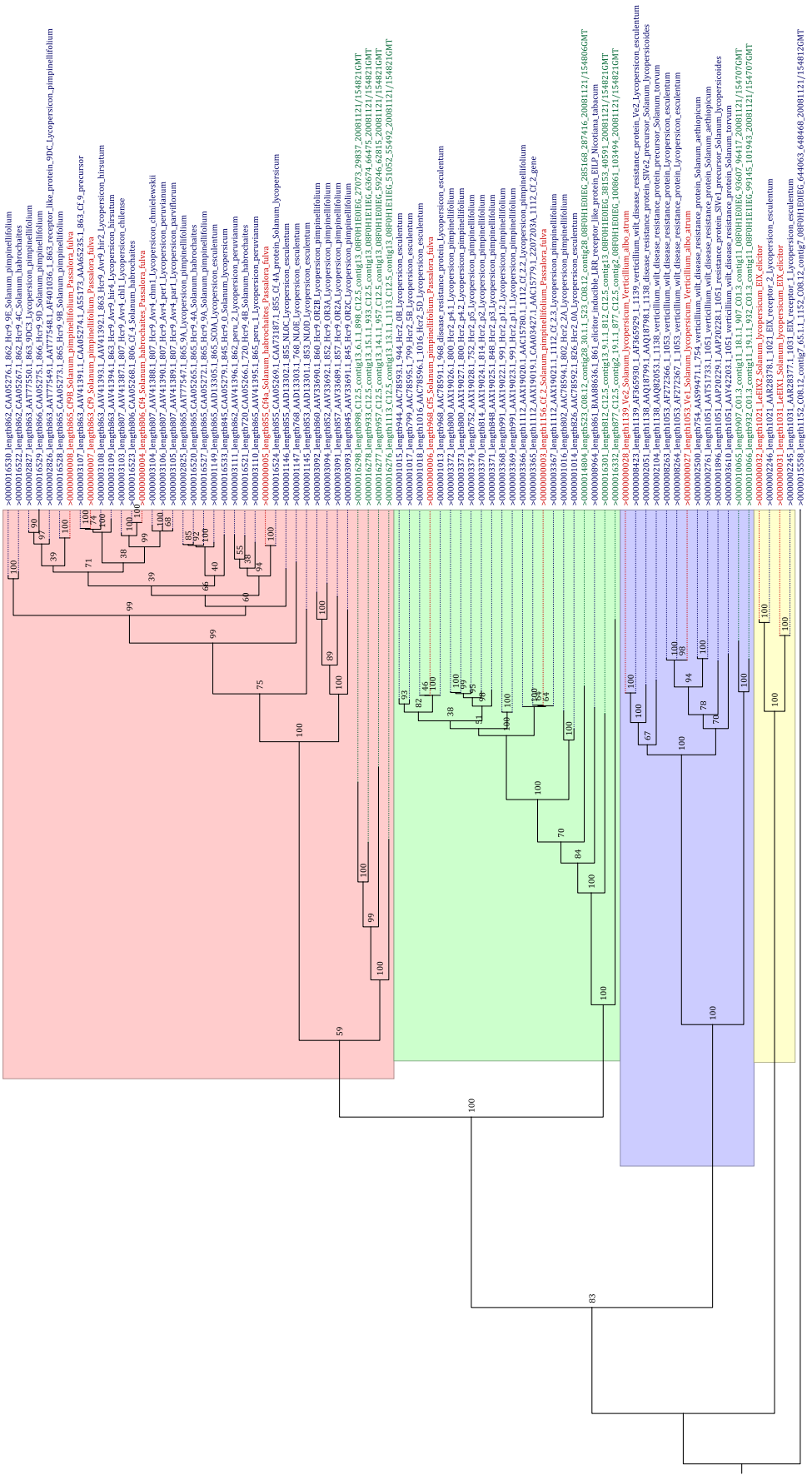


Figura 38. Filogenesi del gruppo di resistenza RLP

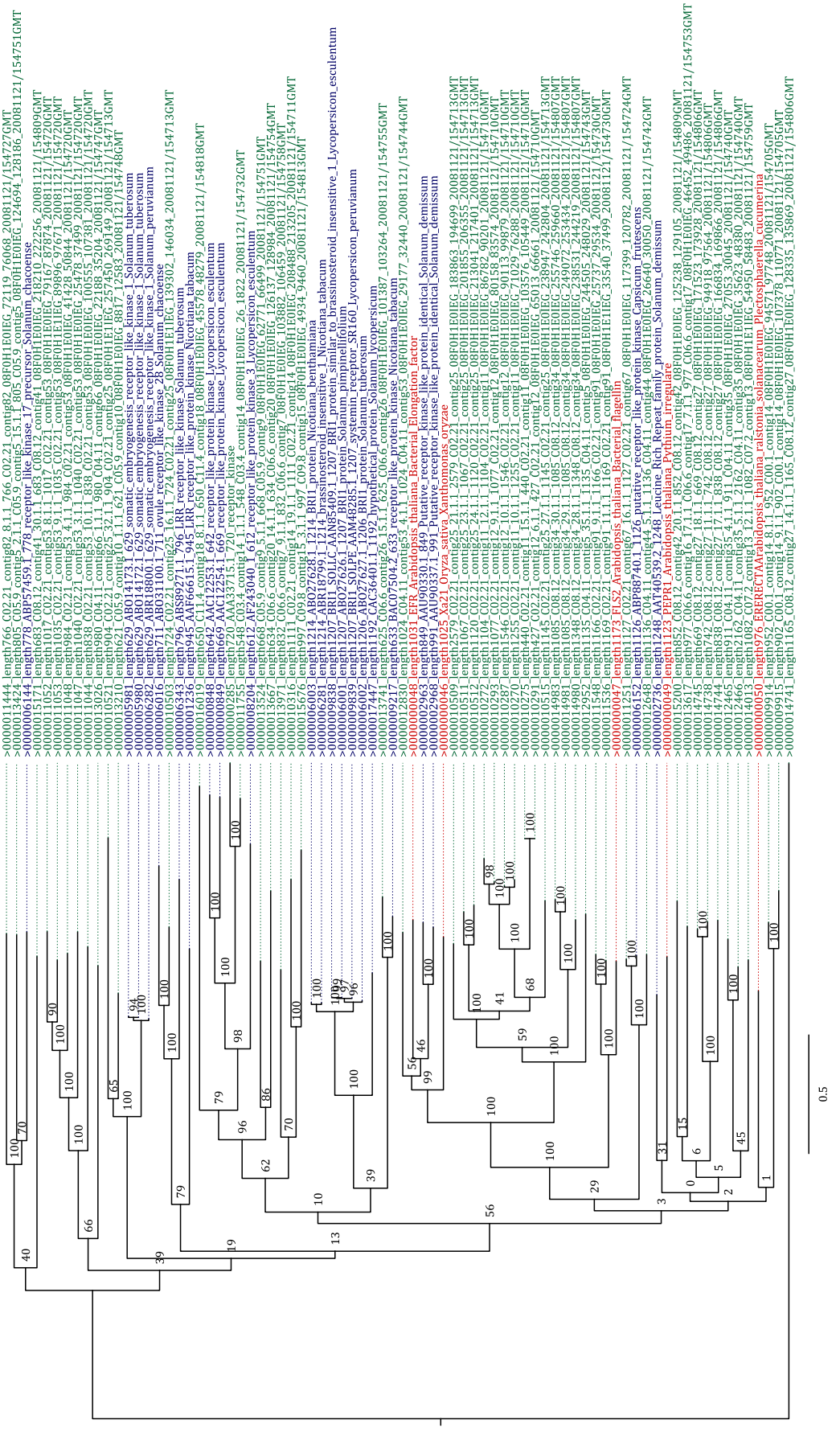


Figura 39. Filogenesi del gruppo di resistenza RLK

3.9 Profilo dei geni R nel genere *Solanum*

Tutte le analisi esposte fin'adesso hanno permesso di tracciare un particolareggiato profilo dei singoli geni R. Il PRG contiene 73 geni R di referenza, di cui 32 appartenenti alla famiglia delle *Solanaceae.* I sistemi di predizione specifici, messi a punto in diversi set di dati, hanno permesso di portare alla luce migliaia di sequenze putative per la funzione di resistenza. In quanto famiglia modello per lo studio delle resistenze, i dati provenienti dalle *Solanaceae* e dal genoma del pomodoro sono stati studiati approfonditamente e sono state create delle correlazioni tra i geni di referenza e le sequenze predette.

Dai dati filogenetici esposti nel capitolo precedente, è possibile notare che per la classe CNL sono state trovate 210 sequenze, di cui 26 dei geni R di referenza. Per i geni *I2* ed *R3a* (che tra essi hanno un omologia 76%) oltre alla sequenza identica di controllo, sono stati ritrovati rispettivamente 2 e 19 proteine altamente omologhe. A causa della loro similarità, sono state trovate altre 19 proteine appartenenti allo stesso gruppo ma con un'omologia non discriminabile per I2 o R3a (figura 34). Le proteine citate provengono in maggior parte da *Solanum demissum* e *tuberosum,* alcune da *Solanum lycopersicum* ed una da *Capsicum chinense*. Per il gene *rpi-blb1* sono state ritrovate 23 sequenze omologhe, di cui 3 provenienti dal genoma di pomodoro, le altre da diverse specie selvatiche di patata ed una da *Nicotiana tabacum.* Per i geni *Tm-2* e *Tm-2a* sono state trovate 3 sequenze provenienti da NCBI dalla specie *Solanum lycopersicum* e 4 dal genoma di pomodoro. Per il geni *Rx1*, *Rx2* e *Gpa2* sono state messe in evidenza alcune sequenze omologhe provenienti da NCBI ed una proveniente dal genoma di pomodoro (figura 35). Per *Bs2* non sono stati ritrovati omologhi se non il gene identico dal set NCBI. Infine i geni *Hero*, *rpi-blb2*, *Mi1.2*, *Sw-5*, *R1* e *Prf* hanno mostrato di avere molti omologhi provenienti sia dal set NCBI sia dal genoma del pomodoro (figura 36).

Per la classe TNL, i cui geni sono più conservati, meno concentrati in singole regioni cromosomiche e più rari, le sequenze omologhe ai 9 geni R, di cui 4 appartenenti alle *Solanaceae*, sono risultate 32. Per i geni *Ry* e *N* le omologie sono state riscontrate solo con le proteine di NCBI, mentre per i geni *Gro1.4* e *Bs4* sono stati trovati diversi omologhi anche nel genoma di pomodoro (figura 37).

Molto più numerose sono risultate le classi RLP e RLK: per la classe RLP sono state analizzate filogeneticamente 294 sequenze di cui 10 geni R di referenza. Com'è possibile vedere dalla figura 38, i geni sono raggruppati in 4 blocchi ben distinti contenenti

rispettivamente i geni *Cf9* e *Cf4* nel primo, i geni *Cf2* e *Cf5* nel secondo, i geni *Ve1* e *Ve2* nel terzo ed infine i geni *EIX1* ed *EIX2* nel quarto. Tutti i geni hanno sequenze altamente omologhe, provenienti sia dal set NCBI sia dal genoma di pomodoro. La classe RLK, composta da 74 sequenze di cui solo 5 appartenenti al set dei geni di referenza, presenta dati molto interessanti (figura 39). Infatti, nonostante sia la classe meno studiata in pomodoro, è quella che contiene il maggior numero di sequenze. Come per gli altri set di dati, anche quest'analisi ha prodotto un cospicuo numero di sequenze, la cui struttura non è tipica delle quattro classi R conosciute e le cui caratteristiche verranno approfondite nel capitolo della discussione.

3.10. Selezione di un pool di geni per studi di caratterizzazione molecolare

Lo studio accurato dei geni R, predetti attraverso l'utilizzo di MATRIX sul genoma di pomodoro, ha permesso di identificare geni con ottime caratteristiche per essere candidati R. L'unione delle informazioni sulla posizione di tali proteine, sulla loro struttura ed sui loro marcatori associati hanno evidenziato molte sequenze interessanti, tra cui il gene 4701 afferente alla classe TNL ed i cluster 3972-73-74 e 6438-40-41-42 appartenenti alla classe CNL. Il gene 4701, sintenico al gene R di patata *Gro1.4,* è un ottimo candidato per espletare una possibile funzione di resistenza contro *Fusarium oxysporum* razza 2, mentre i cluster risultano rispettivamente sintenici e paraloghi al gene *R1* di *Solanum demissum*. Inoltre sono stati indagati gli altri tre geni della classe TNL (candidati 2835, 3993 e 5730) per lo studio dei polimorfismi e della loro evoluzione nelle specie coltivate e selvatiche di pomodoro. Per il candidato 4701 è stato scelto di utilizzare una strategia mirata all'isolamento ed al clonaggio del gene, mentre per gli altri candidati delle classi CNL e TNL è stato scelto di verificarne la presenza in specie di pomodoro selvatico e coltivato al fine di studiarne la conservazione tra le diverse specie e le loro caratteristiche.

3.10.1 Analisi di amplificazione su pomodoro e specie affini

Diverse coppie di primer sono state costruite sui dieci geni appartenenti alle classi CNL e TNL di pomodoro. Le amplificazioni sono state effettuate su diversi genotipi di pomodoro: *S. lycopersicum cv Heinz*, *S. pennelli*, *S. pimpinnellifolium*, *S. peruvianum*, *S. hirsutum* e per tutti i geni le indagini preliminari tramite amplificazione PCR hanno riportato una presenza a livello di DNA su tutte le specie selvatiche tranne che per il gene 6441 assente in *S. peruvianum.* Questi risultati preliminari mostrano l'ubiquità dei

geni R e l'elevato livello di conservazione tra specie affini. Essi sono il punto di partenza per lo studio della conservazione di tali geni e per lo studio del loro polimorfismo ed aprono la strada alla complementazione di nuove sequenze per la funzione di resistenza (figura 40).

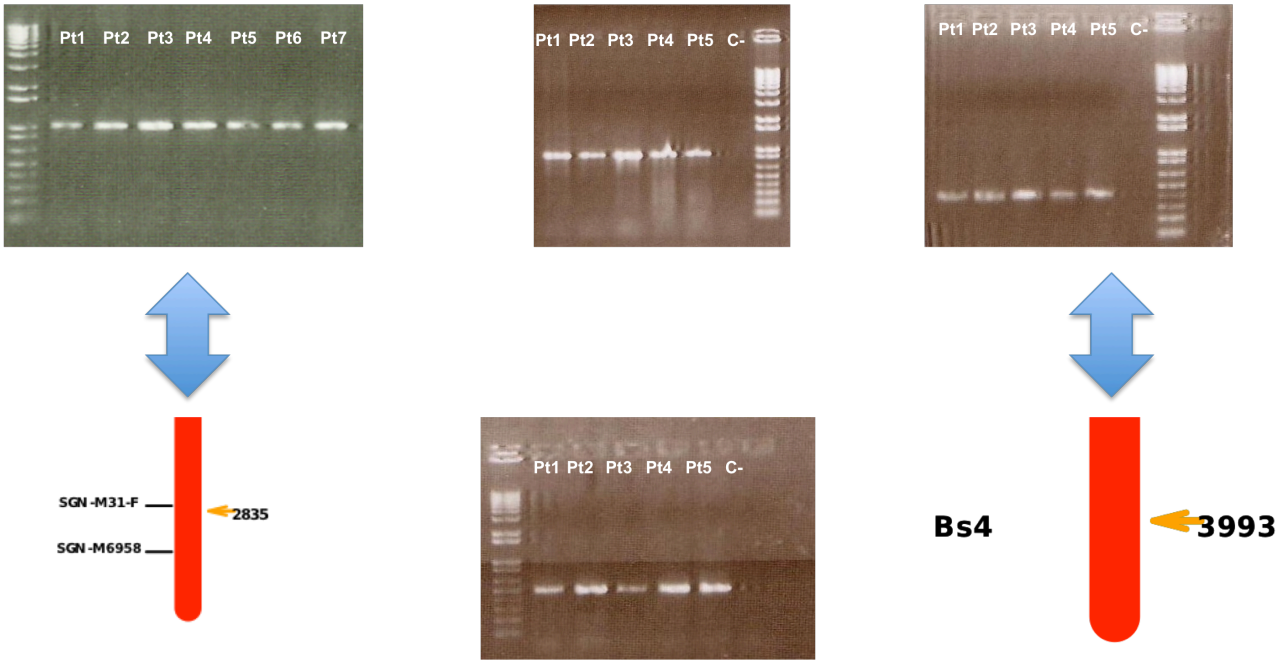


Figura 40. Amplificazione dei geni della classe TNL presenti sul genoma di pomodoro

3.10.2 Caratterizzazione e clonaggio del gene 4701

Il gene 4701 posizionato sul cromosoma 7 contig 6, appartenente alla classe TNL, ha una struttura composta da tre domini, ordinati come segue: TIR-NBS-LRR ed ha un'alta omologia e sintenia con il gene Gro1.4 di patata. La presenza del gene 4701 sul genoma di pomodoro è stata verificata tramite analisi PCR, attraverso l'utilizzo di diverse coppie di primer costruiti all'interno ed all'esterno della regione genomica (figura 41). La presenza di tale gene è stata verificata anche su due genomi che portano resistenza al patogeno *Fusarium oxysporum* razza 2: sulla specie *S. pennelli* e sulla linea di introgressione *IL7-3.* Su questi genotipi il candidato 4701 è presente e sono state effettuate analisi di retro-trascrizione ed amplificazione al fine di studiarne l'espressione. L'analisi ha mostrato un'espressione differenziale tra i diversi genotipi, rivelando la presenza del trascritto solo nei due genotipi resistenti ed una sua assenza nella varietà coltivata Heinz (figura 42). Il gene 4701, differenzialmente espresso, è stato

quindi isolato nella linea di introgressione *IL7-3* ed è stato clonato in un vettore di entrata gateway.

Il gene così clonato può attualmente essere utilizzato per esperimenti di clonaggio funzionale attraverso tecnologia gateway e per la sua complementazione in pianta.

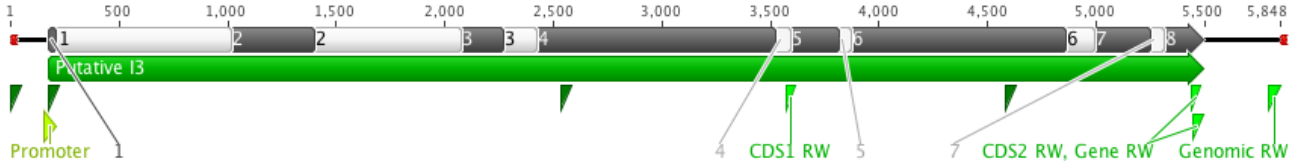


Figura 41. Caratteristiche del gene 4701

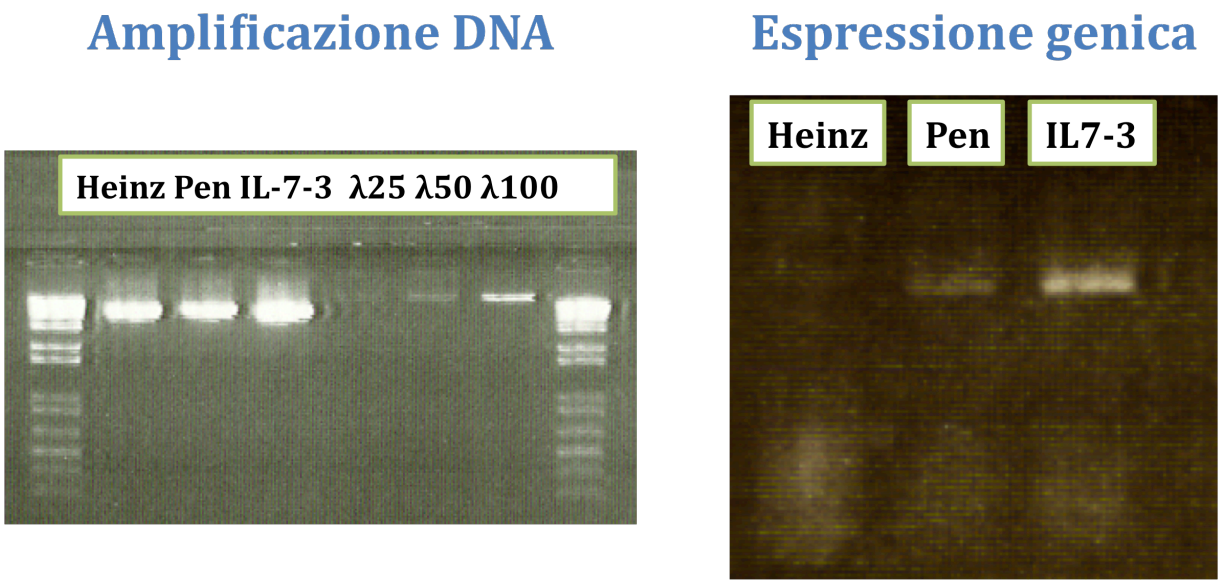


Figura 42. Analisi di amplificazione su DNA e cDNA del gene 4701 della classe TNL

3.11 Profilo delle classi geniche nella famiglia dei geni R

Questo studio ha portato alla luce importanti informazioni non solo a livello dei singoli geni di resistenza, ma anche a livello delle classi che compongono la famiglia dei geni R. L'analisi di sequenze già presenti e di sequenze predette tramite due sistemi distinti di predizione ha portato a poter associare specifiche caratteristiche alle classi R.

La famiglia dei geni R di referenza può essere divisa in 5 classi ben distinte, 4 contenenti domini che combinati in specifiche associazioni espletano la funzione di resistenza ed una, comunemente chiamata "other", nella quale sono racchiusi tutti quei geni che hanno si funzionalità conclamata, ma all'interno dei quali non è possibile trovare delle caratteristiche comuni. Anche se questa classe non è stata ampiamente trattata, oltre

alle 11 sequenze di referenza, sono state predette circa 1600 sequenze nuove, tramite DRAGO. Inoltre tramite i sistemi di predizione utilizzati è stato evidenziato che:

- non esistono solo sequenze con associazioni di domini tipici delle 4 classi descritte in letteratura
- esistono molte altre associazioni di singoli o più domini non ancora approfonditi a livello molecolare.
- il numero di proteine afferenti alle diverse classi è inversamente proporzionale alla loro complessità
- i genomi vegetali presentano profili di resistenza che variano per il numero e la tipologia delle proteine R
- la grandezza del genoma e del proteoma delle specie sequenziate non è correlata al numero di geni R presenti nel loro patrimonio genico
- il genoma di pomodoro contiene proteine afferenti alla classe RLK
- l'organizzazione, la disposizione e le caratteristiche dei geni R in pomodoro non è casuale, ma segue regole di posizionamento e raggruppamento ben precise.

Infine, l'integrazione delle conoscenze prodotte (figura 43) ha permesso di creare una risorsa unica per lo studio delle resistenze vegetali e di individuare un pool di candidati da analizzare e caratterizzare a livello molecolare.

Figura 43. Schema riassuntivo delle risorse sviluppate e dei dati prodotti per studiare i geni R in pianta

4. Discussioni

Le piante, organismi fondamentali per la nostra sussistenza e per quella del nostro pianeta, utilizzano accurati e sofisticati sistemi di difesa per interagire con la realtà circostante, soppravvivere ed evolversi. Alla base dell'interazione pianta-patogeno, della resistenza o suscettibilità delle piante, c'è una complessa macchina molecolare in grado di riconoscere i patogeni, attivarsi e difendere a più livelli (fisico, chimico, molecolare) l'organismo attaccato (16). Questa breve descrizione del sistema immunitario vegetale fa capire quanto esso sia profondamente interconnesso con gli altri apparati della pianta e quanto sia complicato approfondire quest'argomento a causa della presenza di più attori e della complessità delle interazioni. Tra le diverse modalità attraverso cui le piante esplicano questa funzione esiste quella in cui i principali attori sono i geni sentinella R, ovvero di resistenza (32). Questi geni producono proteine in grado di riconoscere specifiche proteine del patogeno e trasdurre un segnale cellulare che attiva le risposte di difesa (102). Lo scopo di questo lavoro è stato caratterizzare la famiglia genica R, per approfondire i meccanismi attraverso i quali espletano la loro funzione, analizzare i geni già conosciuti e predirne ed isolarne di nuovi.

Approfondendo questa tematica è stato subito chiaro che, nonostante l'importanza di tale famiglia, mancava un dato fondamentale: il numero e le caratteristiche dei geni R isolati fino ad oggi. Per questo motivo, il primo passo di questo lavoro è stato una minuziosa ricerca bibliografica che ha portato a creare la più grande e precisa catalogazione della famiglia dei geni di resistenza (9).

L'ultimo manoscritto riguardo la catalogazione dei geni R, risalente al 2007, ne raccoglieva 55 (25), mentre la catalogazione offerta in questo lavoro ne raccoglie 73. A causa del continuo aggiornamento delle conoscenze, sono stati creati dei sistemi automatici per aggiornare continuamente il catalogo, in modo da avere a disposizione dati sempre nuovi e precisi. Oltre a migliorare la conoscenza dei singoli membri afferenti a questa famiglia, questa catalogazione ha permesso di effettuare un'approfondita analisi filogenetica.

La letteratura riporta che la famiglia dei geni R è divisa in 4 classi, ognuna con caratteristiche specifiche che rendono le proteine discriminabili le une dalle altre. L'analisi filogenetica e gli studi sui domini proteici effettuati in questo lavoro rispecchiano questa classificazione e producono informazioni aggiuntive interessanti (24, 16).

Le strutture proteiche delle 4 classi, per quanto composte da domini diversi, hanno tutte la stessa funzione: riconoscere una molecola e trasdurre un segnale. La funzione di riconoscimento è data in tutte le classi da un dominio specifico: quello LRR. Questo dominio, con funzioni di legame e riconoscimento di altre molecole (103), ubiquitario in tutti i geni di resistenza, può essere collegato a diversi trasduttori che ne completano la funzione. I trasduttori delle proteine R fino ad oggi isolati possono essere: molto complessi, come il caso della classe TNL dove sono presenti i domini TIR (102) ed NBS (104) che trasducono il segnale e legano ATP; mediamente complessi, come la classe CNL che contiene il legante ATP (dominio NBS) e al posto del dominio TIR una regione superavvolta, o come la classe RLK che ha un dominio chinasico, legato al dominio LRR (105), trasduttore di segnale; poco complessi, come la classe RLP dove la trasduzione del segnale avviene direttamente dal dominio LRR intimamente correlato ad un dominio serin-treonin chinasico (106).

Osservando l'albero filogenetico prodotto, è chiaro che il ramo più vicino alla radice è quello dei geni CNL. In seguito l'albero si divide in due rami: il primo, che porta alla formazione della classe TNL, e il secondo che, dividendosi a sua volta, porta alla formazione delle classi RLP ed RLK. Le caratterisitche molecolari dei geni delle due famiglie, così come è anche confermato dall'analisi filogenetica, lasciano pensare ad una comune evoluzione che poi ha lasciato il posto ad un evento di differenziazione riguardante solo uno dei domini proteici e lasciando invece la restante parte della sequenza invariata. Questa suddivisione potrebbe supportare la teoria di un'origine comune di tutti i geni R, considerando che tutte le classi hanno in comune il dominio LRR e che tutte le classi svolgono la stessa funzione. L'evoluzione potrebbe essere avvenuta da un modello più complesso ad uno più semplice per garantire una maggiore efficienza (minore consumo energetico per la pianta, capacità di reagire più velocemente in situazione di stress). La costituzione delle quattro classi potrebbe anche essere avvenuta attraverso quattro processi evolutivi paralleli, che oggi non siamo in grado di mettere in evidenza per la quantità insufficiente di informazioni. In ogni caso, l'albero

prodotto ben evidenzia le diversità delle quattro classi, mettendo in risalto le caratteristiche di ognuna. Nonostante i geni R possano facilmente essere classificati nelle rispettive classi, la loro struttura è tale che un singolo cambiamento amminoacidico può cambiare completamente il riconoscimento e la funzione di una proteina (81). Infatti, ogni gene R funziona in modo diverso anche dal suo parente filogeneticamente più stretto, rendendo quindi impossibile associare ad una determinata struttura un particolare tipo di riconoscimento per un patogeno o per una classe biotica. Questo tipo di analisi è stata condotta sia in questo lavoro sia in altri studi, ma non è stato possibile stabilire nessun legame tra la classe della proteina ed il riconoscimento per un particolare tipo di patogeno; anzi diversi lavori dimostrano che tutte le classi contengono proteine che danno resistenza a tutte le tipologie di organismi patogeni (16, 19, 27).

A prescindere dalla loro origine, le differenze e le analogie tra le diverse classi della famiglia R hanno permesso di effettuare interessanti studi volti alla caratterizzazione dei geni R ed alla comprensione dei meccanismi di difesa nei vegetali. Inoltre, poiché l'insorgere della malattia in un organismo vegetale è un evento raro rispetto all'insorgere della resistenza, è giusto pensare che i geni R siano molti di più dei soli 73 isolati, e che ogni organismo ne possa avere diverse centinaia (48). Proprio per questo motivo è stato pensato, a partire dai geni R già conosciuti, di sviluppare dei sistemi di predizioni *in silico* che esplorino il patrimonio di sequenze nucleotidiche e proteiche, prodotte nel corso degli ultimi anni, alla ricerca di nuovi putativi geni R.
L'idea di sviluppare sistemi *in silico* per la ricerca dei geni R, invece dell'utilizzo di un classico approccio di caratterizzazione genica attraverso la biologia molecolare, è nata osservando gli sviluppi dei sistemi di analisi genomica degli ultimi anni. Tali analisi, in grado di produrre migliaia di sequenze in un unico esperimento, hanno "inondato" le banche dati mondiali con tantissime nuove informazioni ancora non analizzate (51). In questo panorama scientifico è plausibile pensare che attualmente non ci sia la necessità di produrre nuove informazioni di sequenza, ma che sia necessario, invece, trovare sistemi in grado di caratterizzare e rendere fruibili i dati già prodotti. Ad oggi è essenziale capire il modo migliore d'utilizzo dei dati di genomica, trascrittomica e proteomica e creare strumenti in grado non solo di correggere gli inevitabili errori computazionali dati dall'analisi massiva dei dati, ma anche di creare sistemi in grado di

filtrare i dati per poter essere sfruttati nuovamente in esperimenti di biologia molecolare, per l'avanzamento delle conoscenze e per il chiarimento dei meccanismi molecolari (107, 108).

Lo sviluppo dei sistemi di predizione ha lo scopo di creare un ponte tra le informazioni racchiuse nei db pubblici e il mondo della genetica vegetale (109). Inoltre, per le caratteristiche della famiglia dei geni R, è plausibile pensare che all'interno dei dati pubblici siano presenti moltissime sequenze, la cui funzione ad oggi è sconosciuta e che invece potrebbero nascondere importanti funzioni molecolari. Lo scopo di questo lavoro è creare un sistema attraverso il quale da grandi quantitativi di dati si possa arrivare alla scelta di geni candidati da caratterizzare a livello molecolare. Grazie alla velocità con cui i nuovi sistemi di sperimentazione sono prodotti, sarà possibile, in futuro, creare strumenti informatici specifici per ogni famiglia genica, che permetteranno di ottenere dati altrettanto interessanti quanto quelli ottenuti, in questo lavoro, per la famiglia dei geni R (9).

Sono stati sviluppati tre specifici strumenti bioinformatici per riconoscere e caratterizzare sequenze altamente simili ai geni di resistenza, per estrapolarle dal loro set di appartenenza e per raccoglierle in un database dedicato all'organizzazione dei dati in semplici schede consultabili attraverso diversi tipi di interrogazioni. Il primo strumento è costituito da un database per la raccolta dei dati di predizione e di letteratura, mentre gli altri due possono essere definiti predittori, in quanto predicono ed associano una possibile funzione a sequenze proteiche la cui funzione è sconosciuta. Il database sviluppato è il PRGdb ei predittori sviluppati in questo lavoro sono DRAGO (9) e MATRIX.

Il primo strumento è stato creato affinché i dati di predizione prodotti possano essere di dominio pubblico ed utilizzati dalla comunità scientifica. Questa è una risorsa specifica per i geni di resistenza dove è possibile trovare tutte le informazioni e le caratteristiche dei 73 geni R funzionali e tutte le sequenze predette tramite il sistema DRAGO. Tale risorsa, costituita da un database ed un sito web, www.prgdb.org, è la prima ed unica nel suo genere e ha caratteristiche all'avanguardia per quanto riguarda l'organizzazione dei dati e la loro visualizzazione. Il numero totale di sequenze presenti è di 16844 divise come segue: 73 geni R di referenza, 6308 proteine provenienti da NCBI, 10463 proteine predette tramite DRAGO. Per i geni di referenza è stata creata una base dati imponente,

ricercando le informazioni da molte fonti e collegandole tra loro. Questo lavoro ha creato una base solida per lo studio della famiglia dei geni R a cui tutti possono attingere per ottenere una visione chiara dei singoli processi di resistenza. Grazie all'utilizzo di schemi di database relazionali, è possibile associare informazioni multiple ad un singolo dato, rendendo possibile per esempio l'associazione di un gene a più di una resistenza o di un gene a più di una specie (85). Un'altra caratteristica importante, oltre alla plasticità della struttura razionale del db, è la dinamicità delle ricerche create al suo interno, grazie alle quali, oltre a raggiungere velocemente le informazioni desiderate, un utente può costruire attraverso una semplice interfaccia grafica una *"query"* su misura per le sue esigenze. Com'è possibile notare nella parte bassa dell'home page, la ricerca che nei risultati è definita come "avanzata", non è altro che una interfaccia per la strutturazione di *query* dinamiche. Infatti un utente può selezionare uno o più domini tipici dei geni R, selezionarne l'associazione ed esplorare le sequenze ottenendo solo quelle con le caratteristiche desiderate.

Successivamente alla nascita del PRGdb sono stati sviluppati DRAGO e MATRIX attraverso cui sono state analizzate rispettivamente 600mila sequenze UniGene ed 800mila sequenze proteiche provenienti da diversi set di dati. Un numero così cospicuo di dati analizzati ha portato a raccogliere nel primo caso circa 10mila UniGene putativi per la funzione di resistenza, nel secondo caso circa 40mila proteine. Un risultato così diverso, per due strumenti sviluppati al fine di produrre predizioni con egual capacità discriminativa, non sorprende a causa della ridondanza dei set analizzati con MATRIX e della caratteristica dei geni R che, poiché strutturati in cluster genici, hanno per ogni gene trascritto diversi omologhi non trascritti (21). Un numero così di alto di sequenze raccolte conferma la presenza di molte sequenze interessanti in set già esistenti e disponibili pubblicamente.

Analizzando gli strumenti dal punto di vista tecnico è possibile notare che DRAGO è un sistema attraverso il quale un gene espresso viene tradotto in proteina, selezionato e catalogato tramite l'analisi dei suoi domini proteici. MATRIX, invece, è un sistema sviluppato *ex novo*, dove, tramite profili HMM creati appositamente sui geni R, è possibile andare ad effettuare predizioni su larga scala.

Prima di comparare i due sistemi è meglio approfondire alcune tematiche inerenti ai singoli sistemi di predizione.

DRAGO non è un software, ma è una "pipeline", ovvero un "percorso", composto da diverse tappe, attraverso cui le sequenze passano e subiscono modificazioni o producono informazioni. Tale percorso è stato strutturato su misura per la famiglia dei geni di resistenza e, essendo stato progettato accuratamente, ha prodotto ottimi risultati. La pipeline funziona in questo modo: una sequenza Unigene viene tradotta in proteina, la proteina viene comparata con le proteine R già conosciute e, se la comparazione ha prodotto alti livelli di omologia, la proteina sarà analizzata a livello dei domini proteici. Se questi domini sono associati in modo analogo ad una delle 4 classi R già conosciute, allora la proteina viene catalogata come putativa per quella classe; se invece le associazioni di domini sono sconosciute, viene catalogata come classe sconosciuta e ne vengono studiate le caratteristiche (9). Il percorso effettuato dalle sequenze non è altro che una concatenazione di diversi software, in cui i dati di uscita di uno diventano i dati di entrata del successivo, che permettono di tracciare un profilo per ogni nuova sequenza registrando tutte le informazioni prodotte nei diversi passaggi. Importanza rilevante hanno avuto gli strumenti di controllo posti tra i diversi passaggi effettuati: stringenti, per evitare di raccogliere sequenze non interessanti, ed elastici, per permettere di raccogliere sequenze simili ai geni R ma con caratteristiche diverse. I dati che hanno permesso di capire se le condizioni impostate potevano essere adatte alla scopo sono stati i risultati stessi. Infatti, testando set di dati contenenti un numero di geni R noti, DRAGO restituiva con una precisione del 100% solo i geni R. I dati evidenziati hanno portato alla luce proteine con associazioni di domini totalmente nuove rispetto alle classi conosciute, così come si desiderava avvenisse. Per validare il sistema, tutte le informazioni prodotte sono state analizzate manualmente e tutte le proteine controllate singolarmente.

MATRIX, invece, è un programma di predizione creato *ex novo* che nonostante si basi su algoritmi e software già conosciuti, propone un nuovo sistema per la predizione delle strutture proteiche. Come molti software di analisi proteiche (PFAM utilizza lo stesso sistema di funzionamento, ma in modo più generico), si basa sull'utilizzo dei profili HMM (100). Utilizzando i profili HMM di PFAM (per i domini LRR, NBS e TIR), per la ricerca di geni R su set di dati noti, si riusciva ad ottenere solo il 25% del numero reale di geni R presenti in quel determinato set. In questo progetto sono stati costruiti profili HMM

molto stringenti a partire dalle sequenze appartenenti alla famiglia genica R. I nuovi moduli HMM non solo rappresentano i domini conservati di una specifica classe R, ma tracciano un profilo delle singole proteine R. Per la prima volta, quindi, lo studio dei geni R non avviene analizzando i domini conservati delle proteine ed osservando le loro posizioni o le loro associazioni, ma basandosi sui profili creati per le classi geniche nella loro interezza. Le proteine analizzate attraverso MATRIX non sono più contenitori di domini proteici, le cui regioni variabili possono essere trascurate, ma sequenze amminoacidiche in grado di assumere specifiche conformazioni e svolgere specifiche funzioni. È facile capire quali sono gli sviluppi di un sistema del genere: attraverso MATRIX non vengono predette proteine con i domini uguali o simili ai geni di resistenza già conosciuti, ma vengono predette proteine con profili uguali o simili ai geni R. Con una sola ricerca è possibile osservare tutte le sequenze con profili comparabili ai geni R e ordinarle in base alla loro omologia con le proteine R.
Analizzando grossi set di dati con questo sistema non solo è possibile predire tutte le proteine simili ai geni R, ma è facilissimo vedere anche quali proteine nuove sono associate a proteine la cui funzione è già nota. MATRIX non è stato costruito come un asettico strumento bioinformatico, ma come uno strumento genetico che sfrutta l'informatica, dando un senso biologico ai risultati che produce.

Avendo brevemente approfondito le caratteristiche dei due strumenti sviluppati, è possibile adesso compararne le caratteristiche e capirne il loro giusto utilizzo, iniziando dal dire che: entrambi sono progettati per lavorare attraverso algoritmi per lo studio delle omologie di sequenza; utilizzano come base di partenza le caratteristiche dei geni di resistenza già clonati; lavorano, anche se in modo diverso, su sequenze proteiche e sui loro domini conservati.

I due sistemi sono differenti per la scelta dei set dei geni di resistenza di partenza da utilizzare come base informativa: per il sistema DRAGO sono stati utilizzati tutti i 73 geni R, mentre per il sistema MATRIX solamente i geni appartenenti al genere *Solanum,* per le classi R CNL, TNL ed RLP, e tutti i geni appartenenti ad *Arabidopsis thaliana* per la classe RLK (non presente nel genere *Solanum*). Questa scelta è dovuta al fatto che, per DRAGO, maggiore è il numero di geni di referenza maggiore sarà la possibilità di predire sequenze putative per la funzione di resistenza; per MATRIX, più caratteristiche ha in

comune il set di referenza meglio sarà costruito il profilo delle classi R. Oltre ai dati di base per lo sviluppo dei sistemi, un'altra differenza si riscontra nei set di dati che i due predittori possono analizzare. Le proteine analizzate da DRAGO derivano da geni trascritti dagli organismi vegetali, mentre le proteine analizzate da MATRIX sono annotazioni sui genomi vegetali sequenziati. Nel primo caso i vantaggi delle sequenze provenienti da ESTs sono rappresentati dalla loro alta probabilità di essere proteine funzionali, o comunque presenti nel patrimonio di un organismo, e dall'assenza di sequenze mal predette, troncate o non funzionali; mentre lo svantaggio di un set di dati del genere è la perdita di informazioni importanti sui geni ad espressione indotta e su tutte le sequenze non trascritte ma che contengono importanti caratteristiche strutturali ed evolutive (133). Nel secondo caso, invece, i vantaggi di set di dati come i genomi annotati sono rappresentati da una visione globale del proteoma di un organismo e dalla possibilità di utilizzare i dati prodotti non solo per la caratterizzazione di nuovi geni, ma anche per altre tipologie di esperimenti; mentre gli svantaggi sono la presenza di errori dovuti all'errato assemblaggio di sequenze, e quindi alla mal annotazione dei geni sul genoma, e l'impossibilità di capire, se non con esperimenti molecolari, la reale funzionalità delle proteine analizzate (134, 135, 136). È chiaro quindi che dati differenti producono risultati simili ma con significati molto diversi e che un possibile modo di eliminare gli svantaggi dei differenti set è quello dell'utilizzo combinato o parallelo dei due sistemi sviluppati. Un possibile scenario futuro potrebbe essere lo studio del trascrittoma di un organismo attraverso DRAGO, lo studio del suo proteoma attraverso MATRIX, la comparazione dei dati e l'ottenimento di un profilo di resistenza dove, oltre alle informazioni sui geni predetti, si riesca a discernere le sequenze trascritte da quelle silenziate.

Oltre a processare dati diversi, i due sistemi possono essere utilizzati per scopi diversi: DRAGO è più utile per raccogliere putativi geni R funzionali adatti per essere caratterizzati ed utilizzati, mentre le possibilità offerte da MATRIX sono di ottenere il profilo delle resistenze di un intero organismo ed associare le sequenze ottenute ai già funzionali geni R, permettendo di effettuare interessanti studi di genomica strutturale, comparativa e di biologia evolutiva. Importante è anche il sistema di classificazione finale delle proteine prodotte: per DRAGO sono state utilizzate le informazioni di InterProScan e le sequenze classificate secondo i loro domini conservati, mentre per MATRIX la classificazione avviene grazie a software di raggruppamento che operano

sulla matrice prodotta, permettendo di organizzare le proteine in base al loro profilo ed associarle con i già isolati geni R. In quest'ultimo caso il risultato non è fine a se stesso in quanto non solo si caratterizza la nuova sequenza come appartenente ad una classe R, ma si effettua anche un lavoro di omologia con i geni R, collocando la nuova proteina più o meno distante da una proteina R di referenza. Nonostante i due sistemi di classificazione utilizzati siano diversi, il risultato ottenuto è molto simile: i nuovi putativi geni di resistenza vengono classificati in modo molto chiaro tramite i loro domini o i loro profili e le sequenze con nuove associazioni di domini o con singoli domini, in entrambi i casi, vengono classificate come simili ai geni R ma con struttura sconosciuta.

I risultati ottenuti dai due strumenti sono il frutto di un'attenta analisi dei processi computazionali e da una profonda conoscenza delle strutture proteiche associate ai geni R. I parametri di stringenza utilizzati sono stati calibrati dopo numerose sperimentazioni, permettendo di trovare un preciso equilibrio tra affidabilità e plasticità degli strumenti stessi. Il parametro più importante per DRAGO è il "cut-off" del BLAST scelto per decidere se un gene possa essere definito come un omologo, mentre per MATRIX i parametri sui quali si può "giocare" sono svariati, a partire dalla specificità degli allineamenti per la costruzione dei profili, finendo alla specificità dei profili stessi o ai parametri di pulizia della matrice prodotta. Proprio per quest'ultimo parametro bisogna spendere due parole sul perché si è scelto di eliminare dalla matrice tutte le sequenze con un'omologia minore di 8 profili HMM. È stata effettuata questa scelta perché alcuni domini delle proteine R, come quello chinasico o quello serin-treonin chinasico, sono molto comuni e sono presenti in altre famiglie proteiche oltre a quella dei geni R (137); inoltre la presenza di un singolo dominio non è sinonimo di putatività per la funzione di resistenza. Per queste due motivazioni, poiché il dominio più piccolo è composto da 9 profili HMM, eliminare tutte le proteine predette con meno di 8 profili equivale ad eliminare tutte le sequenze composte o da un unico dominio non specifico per i geni R o da sequenze con un'omologia molto bassa. Infine è bene chiarire quali sono i principi su cui si basano le visualizzazioni delle matrici prodotte da MATRIX. Per rendere fruibile a tutti i risultati di predizione, le matrici sono state visualizzate secondo uno schema in cui le gradazioni di diversi colori rappresentano una maggior o minor omologia. Tali programmi permettono anche di raggruppare i geni utilizzando diversi algoritmi in grado di dividere i geni predetti in gruppi di appartenenza (92).

Tutti i dati prodotti tramite gli strumenti sviluppati sono stati analizzati approfonditamente e sono state raccolte informazioni molto interessanti sui vari aspetti genetici della famiglia R. Sicuramente il dato più interessante riguarda le nuove associazioni di domini ritrovate tanto nei dati analizzati da DRAGO quanto in quelli analizzati da MATRIX. Infatti, in tutti i set di dati, oltre alle strutture tipiche delle 4 classi ben conosciute, sono state ritrovate strutture con singoli domini o con nuove associazioni di questi. Questo risultato diventa ancor più interessante quando i dati analizzati provengono sia da sequenze sicuramente espresse sia da sequenze genomiche, come in questo caso, poiché fanno ipotizzare non solo che nei genomi esista una gran variabilità di sequenze simili ai geni R, ma anche che tali strutture possano avere una funzione biologica in quanto trascritte. Il predittore DRAGO tra le sequenze Unigene ha portato alla luce ben 16 nuove associazioni di domini proteici e molte altre strutture interessanti, dove i domini vengono spesso ripetuti più volte, come il caso del gene di pioppo, Entrez DQ513219, strutturato con domini TIR-NBS-TIR (116).

Da subito è stato chiaro che i domini tipici della famiglia R sono associati in strutture non ancora conosciute per tipologia e per funzionalità. Dai dati ottenuti risulta che le nuove associazioni di domini possono essere più complesse o meno complesse rispetto alle strutture afferenti alle classi conosciute. Infatti, il numero di domini può variare da 1 a 4 ed il numero di sequenze correlate a tali strutture è inversamente proporzionale alla complessità delle proteine. Prima di approfondire il significato di queste nuove informazioni è bene domandarsi se le strutture isolate producono proteine funzionali e se tali strutture possono essere associate ad una putativa funzione di resistenza. Non esistono dati sperimentali che confermano la funzionalità o meno delle proteine predette. È pur vero però, che poiché molte di esse sono trascritte, è altrettanto probabile che molte di esse siano anche tradotte. Ad oggi non esistono geni R di referenza che hanno strutture diverse da quelle delle quattro tipiche classi. C'è da notare che le nuove associazioni predette contengo prevalentemente domini tipici dei geni R. Associazioni che non contengono domini LRR o che contengono domini non correlati alla resistenza potrebbero avere funzione diversa o essere semplici trasduttori di segnali cellulari implicati in processi diversi (111).

Tuttavia osservando i dati di predizione ci si pone una serie di domande a cui rispondere: perché non esistono tutte le associazioni di domini? Perché esistono

strutture con molte sequenze ed altre rare? Perché alcuni domini sono molto più presenti di altri? Perché il numero di proteine con domini singoli è decine di volte superiore a quello di proteine con strutture più complesse?

Si potrebbe trovare risposta in un'unica ipotesi che potrebbe essere definita di "efficienza e plasticità". Tale teoria si basa sulle caratteristiche che i geni R devono avere, per espletare al meglio la loro funzione: devono essere efficienti nel riconoscere specificamente un patogeno, ma al tempo stesso plastici per far fronte al processo di co-evoluzione pianta-patogeno (112). Per avere queste due caratteristiche le strutture dei geni R devono essere efficienti, ovvero aver la miglior associazione di domini che gli permetta di riconoscere al meglio i patogeni, e plastici, ovvero avere un serbatoio genomico, probabilmente composto da singoli domini, che grazie a diversi processi di ricombinazione possano essere ricombinati in geni R con nuove funzioni di riconoscimento, così come avviene in altri organismi e nei mammiferi (in modo molto più veloce) per la produzione delle immunoglobuline (113, 114, 115). Osservando le predizioni ottenute in funzione del concetto di efficienza e plasticità, i dati acquisiscono un significato molto più profondo. Infatti, i singoli domini molto numerosi potrebbero essere il serbatoio della plasticità e le associazioni delle 4 classi note rappresentano le strutture più efficienti. Infatti le 4 strutture conosciute sono le classi più numerose, subito dopo le sequenze con i singoli domini.

I dati prodotti da DRAGO trovano conferma anche in quelli prodotti da MATRIX, anche se in quest'ultimo caso le sequenze contenenti singoli domini sono in numero minore a causa dei parametri di stringenza più selettivi che eliminano le sequenze, a bassa omologia, con un solo dominio.

Dai dati di predizione è possibile delineare un panorama dove probabilmente le quattro classi conosciute sono quelle che hanno la struttura più adatta ad esplicare la funzione di resistenza, dove le proteine con i singoli domini creano una variabilità strutturale a cui è possibile accedere per produrre nuovi sistemi di riconoscimento, e dove le nuove classi predette sono nuove associazioni strutturali funzionali in grado di lavorare o singolarmente o associate in complessi proteici con più attori. Infine, dai dati esaminati è possibile capire perché gli esperimenti di genetica classica hanno sempre isolato geni delle quattro classi principali (molto più numerose rispetto alle altre classi) ed inoltre è possibile ipotizzare un'evoluzione della famiglia R dove, attraverso sistemi di ricombinazione ancora non chiariti (118), si vadano a creare strutture proteiche diverse

e con diverse associazioni di domini, contenenti il dominio LRR nel caso di un'interazione a singolo gene o in grado di formare strutture più complesse in assenza di esso.

Come attraverso i dati ottenuti da DRAGO è stato possibile approfondire le nuove associazioni proteiche, così tramite MATRIX è stato possibile tracciare un profilo delle resistenze vegetali dei singoli organismi sequenziati. Nonostante anche MATRIX abbia portato alla luce la presenza in tutti i genomi di proteine con struttura diversa da quella conosciuta per le 4 classi, tali dati non sono stati approfonditi non solo perché MATRIX è stato calibrato solo per lo studio della genomica delle proteine appartenenti alle 4 classi conosciute, ma anche perché i parametri di stringenza utilizzati non sono impostati per la predizione di proteine con singoli domini tipici della famiglia delle resistenze. Nonostante tale parametro possa essere modificato e le analisi rielaborate per la ricerca di sequenze con singoli domini, questo risulterebbe sconveniente per diversi motivi: le annotazioni dei genomi sequenziati non sono affidabili al 100%, ma anzi l'errore commesso più comunemente è proprio quello di unire due geni o troncarne uno in due, stravolgendo in questo caso tutte le percentuali di predizione (119); non utilizzare MATRIX ad alta stringenza determina la raccolta di sequenze non correlate direttamente ai processi di resistenza e di tutte le proteine contenenti i domini più conservati, per esempio le chinasi, e non quelle con i domini più variabili ma anche più interessanti, per esempio le LRR. Per tutte queste considerazioni MATRIX, costruito appunto per scopi diversi rispetto a DRAGO, non ha funzione di esplorare i genomi alla ricerca di nuove associazioni, ma ha lo scopo di tracciare i profili di resistenza di genomi interi a partire da profili conosciuti (45, 58, 59, 60, 61).

Tramite la predizione di DRAGO è stato possibile focalizzarsi sull'analisi della classe "other" che racchiude tutte le proteine che danno resistenza con meccanismi diversi da quelli delle classiche classi dei geni R. Per meccanismi diversi si intende strutture proteiche completamente diverse, spesso non contenenti domini conservati e che in ogni caso non posso essere correlati a nessuna delle quattro classi con proteine con domini conservati. Appartenenti a questa classe ci sono per esempio il gene *Asc-1* che funziona detossificando le cellule dalle fumotossine del fungo *Alternaria alternata* (120) o il gene *mlo* che conferisce resistenza al patogeno *Blumeria graminis* (68). La scelta di

inserire questi 11 geni all'interno del sistema di predizione di DRAGO ha permesso di isolare ben 1662 sequenze con alta omologia. Non contenendo domini conservati tali sequenze non possono essere né analizzate come appartenenti ad un'unica famiglia né associate a processi di resistenza conosciuti così com'è avvenuto per le altre quattro classi. È chiaro però che anche questi geni hanno un'importanza fondamentale per lo studio delle resistenze ed in parte sono correlati alle altre classi, non solo perché potrebbero portare alle luce diversi meccanismi di resistenza meno studiati, ma anche perché è possibile trovare in futuro proteine composte dai tipici domini conservati e da domini di proteine appartenenti alla classe "other". Questa ipotesi è stata in parte già validata quando sono stati analizzate due proteine particolari: la prima *RFO1*, che nella filogenesi dei geni R ha un braccio molto lungo che non gli permette di ottenere un posizionamento con alto livello di affidabilità, poichè oltre ai domini LRR e Ser-thr tipici della sua classe di appartenenza, RLP, contiene anche il dominio HMA tipico invece di tutta la famiglia dei trasportatori di metalli (121); la seconda, il gene di pomodoro, 3609 (vedi risultati paragrafo 3.7.1.1), che isolato tramite MATRIX mostra la tipica struttura della classe CNL con associato il tipico dominio del gene di *Arabidopsis* RPW8 appartenente alla classe "other". Inoltre, è stato ritenuto opportuno inserire i geni "other" tra quelli di referenza ed effettuare analisi predittive su di essi in quanto la risorsa PRGdb è nata per dare una visione d'insieme sui diversi processi di resistenza di cui anche questa categoria genica fa parte.

Se DRAGO ha prodotto importanti risultati per quanto riguarda l'organizzazione strutturale delle proteine R, MATRIX ha prodotto importanti risultati per quanto riguarda i profili di resistenza nelle specie già sequenziate, la genomica strutturale dei geni R e la filogenesi delle diverse classi della famiglia R. Diverse volte durante la descrizione dei risultati è stato sottolineato che i dati di predizione di MATRIX per essere analizzati sono stati divisi in sotto classi ognuna di esse omologa ad un determinato set di proteine R. Questa divisione, che non rispecchia più le quattro classi tipiche, ma è molto più specifica, è data proprio perché MATRIX basa la sua predizione sul profilo dei geni R e non sui suoi domini conservati.
I genomi vegetali analizzati sono stati in tutto 8 di cui 5 completamente sequenziati, e 3 parzialmente; 5 appartenenti alla divisione delle dicotiledoni mentre 3 a quella delle monocotiledoni. Le analisi effettuate sui genomi completamente sequenziati hanno

tracciato un profilo globale del panorama delle resistenze delle specie analizzate ed i dati estrapolati hanno permesso la comparazione dei profili genici delle diverse specie. La percentuale delle proteine R presenti in ogni specie varia dal 3 al 6 % e le comparazioni tra il numero di proteine R predette e la grandezza dei proteomi dei diversi genomi, effettuate tramite il test del qui-quadro, evidenziano che le differenze sono significative, dimostrando che il numero di geni R in ogni genoma è diverso l'uno dall'altro (122, 123). Del resto non sorprende questo dato in quanto i 5 genomi appartengono a specie totalmente diverse tra loro, con storie evolutive e pressioni selettive diverse (124, 125). L'interpretazione dei profili di resistenza delle diverse specie mette in risalto la caratteristica di molte sequenze ad avere domini LRR più o meno conservati determinando la presenza in tutte le classi di profili HMM relativi al dominio LRR. Tale dato può essere interpretato come una conservazione più o meno spinta di tale dominio tra le diverse classi R e può essere utilizzato per tracciare l'ipotesi che vede un origine comune dei geni R a partire da questo dominio.

Utilizzando la classificazione filogenetica proposta da Ziemann (126), basata su un gruppo di enzimi molto conservati presenti in tutte le specie di piante è possibile datare la comparsa di *Vitis vinifera* a circa 125 milioni di anni, ed a seguire *Popolus trichocarpa* ed *Arabidopsis thaliana*. Questa filogenesi pone quindi *Arabidopsis* come la specie più giovane tra le tre e vite come la più antica. Tali dati possono essere utilizzati per interpretare meglio la figura 17 dove si nota una distribuzione diversa di geni R appartenenti alle quattro classi. Infatti, premettendo che tutte le classi hanno la possibilità di instaurare resistenza a tutti i diversi stress biotici è possibile che la differenza di distribuzione possa essere dovuta a caratteristiche strutturali insite alle specie stesse o in processi evolutivi. Le considerazioni che possono essere fatte osservando la divisione dei geni predetti, la filogenesi dei geni R di referenza e tenendo presente l'ipotesi dell'efficienza e plasticità sono: che *Vitis* ha il proteoma più piccolo ma il numero di geni di resistenza maggiore e che ha un numero maggiore di geni R citoplasmatici (ovvero CNL e TNL) rispetto agli altri due genomi che sono spostati verso le classi con geni R di membrana. Tenendo conto dei cicli di replicazione, e quindi del tasso di evoluzione, che vede *Arabidopsis* in testa seguita da *Popolus* e quindi da *Vitis,* è possibile osservare che i profili di resistenza mostrano una presenza sempre maggiore di geni R di membrana a scapito di quelli citoplasmatici, come del resto avviene anche nei mammiferi dove l'evoluzione ha conservato prevalentemente i geni di membrana

(113). Questo dato messo in evidenza dai profili dei diversi genomi va verso un processo evolutivo che vede la presenza di un numero maggiore di proteine di membrana a scapito di quelle citoplasmatiche.

Oltre all'analisi dei genomi completamente sequenziati tramite MATRIX sono stati analizzati i genomi parzialmente sequenziati di altre 3 specie. In particolare l'attenzione del lavoro si è focalizzata sul profilo di resistenza del genoma del pomodoro. Essendo stato costruito sui geni del genere *Solanum*, MATRIX ha una particolare efficienza nel discriminare le differenze tra le proteine predette sul genoma del pomodoro.
In quanto non ancora completamente sequenziato, tale analisi non è servita per tracciare un profilo di resistenza definitivo di pomodoro, ma per effettuare uno studio di genomica strutturale sulla disposizione ed il raggruppamento dei geni R. Queste tipologie di analisi sono state effettuate unendo i dati di predizione del genoma, i dati di sequenziamento prodotti dalla comunità scientifica (38) ed i dati sui marcatori molecolari associati al pomodoro (127). L'unione dei diversi dati ha permesso di predire i putativi geni R presenti nel genoma, localizzarli sui cromosomi ed associarli ai marcatori genetici. Questo lavoro ha permesso, di creare una mappa, unica nel suo genere, poiché unisce informazioni genetiche ad una mappa fisica, definita "mappa R di pomodoro", dove è possibile osservare la localizzazione e la distribuzione dei putativi geni R ed inoltre ha permesso di effettuare studi approfonditi sulla disposizione, clusterizzazione e distribuzione dei geni R predetti. Attraverso la descrizione dei risultati ottenuti è risultato evidente che i geni R hanno disposizioni e distribuzioni particolari: disposizioni poiché si trovano spesso associati in cluster composti da geni appartenenti ad una o più famiglie, distribuzioni poiché sono distribuiti in modo non uniforme privilegiando specifici cromosomi e specifiche regioni cromosomiche (128, 129). Un analisi di genomica strutturale di questo tipo ha permesso di confermare molti dati sulla distribuzione dei geni R ed inoltre di iniziare a tracciare un quadro definitivo sulla duplicazione, la disposizione e il raggruppamento di questa famiglia genica in una specie modello per lo studio delle resistenza come quella del pomodoro. L'analisi del genoma del pomodoro attraverso il sistema MATRIX oltre a produrre le interessanti informazioni che ben vengono descritte nella sezione dei risultati, mostra una notizia singolare: la presenza di molti geni appartenenti alla classe RLK. Tale informazione è ritenuta di estrema importanza poiché la letteratura sui geni di resistenza, in pomodoro,

non ha prodotto alcun risultato su questa classe se non per l'isolamento di un gene omologo a FLS2 di *Arabidopsis* (130). Tramite le analisi effettuate in questo lavoro sono state ritrovate invece ben 52 sequenze strutturalmente riconducibili alla classe RLK il che conferma che il pomodoro, come tutte le altre specie vegetali, contiene nel suo patrimonio proteine di questa tipologia. Il pomodoro contiene due tipologie di proteine RLK, la prima composta da proteine contenenti i domini ser/KIN, la seconda composta proteine contenenti i domini ser/KIN ed LRR. Di questa prima tipologie esiste in pomodoro il gene *pto*, mentre per la seconda non esistevano ad oggi evidenze sperimentali sulla loro presenza. Approfondendo il meccanismo di funzionamento di *pto* (131) è possibile che la classe Ser-thr/KIN mancando del dominio LRR non ha funzioni di recettore ma solo di trasduttore di segnale essendo quindi parte di un sistema più complesso dove la resistenza necessita di più proteine con funzioni diverse (117, 132).

Oltre ad uno studio strutturale sull'organizzazione genomica dei geni R, il set predetto è stato analizzato per evidenziare il rapporto tra l'omologie delle sequenze delle singole classi in base alla loro distanza. Quest'analisi è stata effettuata intraprendendo studi filogenetici sulle sequenze delle classi R del set di pomodoro e studiando l'omologia delle sequenze tra di esse in base alla loro localizzazione cromosomica. Attraverso quest'analisi è stato possibile verificare la presenza di cluster genici all'interno delle diverse classi R e la presenza di cluster con alti livelli di omologia.
Il sistema di predizione MATRIX è stato utilizzato su un set di sequenze provenienti da tutta la famiglia delle *Solanaceae* in primo luogo perché questa famiglia è un modello per lo studio delle resistenze, secondo perché essendoci diversi dati prodotti sulle specie selvatiche è interessante esplorarli e cercare di costruire una relazione tra i geni provenienti dalle diverse specie.
Il raggruppamento di MATRIX ha diviso i geni predetti in diverse sotto classi ognuna con caratteristiche simili ad un gruppo ben definito di geni R. Da questi dati è stata prodotta la tabella 4 attraverso la quale facilmente è possibile osservare la distribuzione dei geni R tra le diverse specie di *Solanaceae*. Inoltre per collegare le sequenze provenienti dalle diverse specie è stata effettuata un'analisi filogenetica unendo i geni predetti a quelli di referenza e sviluppando una serie di associazioni tra i due set di dati. Grazie all'analisi filogenetica è stato possibile capire quali e con che specificità le sequenze sono correlate tra loro e, con le informazioni aggiuntive sulla specie e sul set provenienza, per ogni

gene di referenza è stato possibile creare un profilo di omologia. Questi profili, oltre a racchiudere tutte le informazioni già presenti in letteratura ma difficili da trovare e raccogliere, hanno anche ripulito i set di dati eliminando tutte le informazioni errate, come l'inserimento di sequenze non corrette o inserite in banca dati in modo non esatto. Tali analisi sono risultate ancor più utili proprio per le informazioni aggiuntive sulle proteine inserite durante la composizione degli alberi. Infatti, poiché il formato per le analisi filogenetiche deve utilizzare proteine le cui sigle discriminative possono essere composte solo da 10 caratteri sarebbe stato impossibile leggere i risultati in modo esaustivo. Per questo motivo è stato sviluppato uno *script* apposito per annotare le informazioni sugli alberi ottenuti indicando il set di provenienza, la descrizione delle proteina se annotata, la posizione genomica e per i geni di referenza le informazioni sui patogeni, ottenendo in questo modo un quadro completo per ogni singola proteina.

5. Conclusioni

I geni di resistenza hanno grossa importanza nel settore della genetica vegetale e del miglioramento genetico. Le loro caratteristiche genetiche e strutturali li rendono interessanti sia da un punto di vista scientifico sia nell'ottica di un loro utilizzo pratico. Sono una famiglia affascinante da studiare e capire i loro meccanismi evolutivi e di interazione con altre proteine rappresenta una sfida per il chiarimento di molte caratteristiche molecolari e di funzionamento. Questo studio propone un approccio innovativo per l'analisi dei geni R, utilizzando dati prodotti da esperimenti di genomica su larga scala, ed integra i risultati al fine di creare un *corpus* unico di conoscenze.

Un importante obbiettivo del miglioramento genetico è, da sempre, l'utilizzo dei geni R in programmi per la produzione di varietà con resistenze naturali nel rispetto dell'uomo e dell'ambiente. La creazione di una risorsa accessibile gratuitamente via web, specifica per i geni R; la loro catalogazione e caratterizzazione; lo sviluppo di strumenti di predizione genica specifici; la selezione di nuovi candidati per la lotta agli stress biotici, fa di questo lavoro una base salda per l'esplorazione dei meccanismi molecolari d'interazione pianta-patogeno e per l'isolamento di nuovi geni R da utilizzare, attraverso programmi di miglioramento genetico, nelle varietà del prossimo futuro.

6. Bibliografia

1. SIMS D. A,. (1997) Concepts of plant biotic stress. Some insights into the stress physiology of virus-infected plants, from the perspective of photosynthesis : Plant response to stress. Physiologia Plantarum, 100, 11.

2. Keen, N.T. (1990) Gene-for-gene complementarity in plant-pathogen interactions. Annu Rev Genet, 24, 447-463.

3. Hain, R., Reif, H.J., Krause, E., Langebartels, R., Kindl, H., Vornam, B., Wiese, W., Schmelzer, E., Schreier, P.H., Stocker, R.H. et al. (1993) Disease resistance results from foreign phytoalexin expression in a novel plant. Nature, 361, 153-156.

4. Hammerschmidt, R. and Dann, E.K. (1999) The role of phytoalexins in plant protection. Novartis Foundation symposium, 223, 175-187; discussion 188-190.

5. Bonas, U. and Lahaye, T. (2002) Plant disease resistance triggered by pathogen-derived molecules: refined models of specific recognition. Current opinion in microbiology, 5, 44-50.

6. Hammond-Kosack, K.E. and Jones, J.D. (1996) Resistance gene-dependent plant defense responses. The Plant cell, 8, 1773-1791.

7. Flor, H.H. (1971) Current Status of the Gene-For-Gene Concept. Annual Review of Phytopathology, 9.

8. Ryals, J.A., Neuenschwander, U.H., Willits, M.G., Molina, A., Steiner, H.Y. and Hunt, M.D. (1996) Systemic Acquired Resistance. The Plant cell, 8, 1809-1819.

9. Sanseverino, W., Roma, G., De Simone, M., Faino, L., Melito, S., Stupka, E., Frusciante, L. and Ercolano, M.R. (2009) PRGdb: a bioinformatics platform for plant resistance gene analysis. Nucleic acids research.

10. Heil, M. and Bueno, J.C. (2007) Herbivore-induced volatiles as rapid signals in systemic plant responses: how to quickly move the information? Plant signaling & behavior, 2, 191-193.

11. Levine, A., Tenhaken, R., Dixon, R. and Lamb, C. (1994) H2O2 from the oxidative burst orchestrates the plant hypersensitive disease resistance response. Cell, 79, 583-593.

12. Feys, B.J. and Parker, J.E. (2000) Interplay of signaling pathways in plant disease resistance. Trends Genet, 16, 449-455.

13. Anderson, J.P., Badruzsaufari, E., Schenk, P.M., Manners, J.M., Desmond, O.J., Ehlert, C., Maclean, D.J., Ebert, P.R. and Kazan, K. (2004) Antagonistic interaction between abscisic acid and jasmonate-ethylene signaling pathways modulates defense gene expression and disease resistance in Arabidopsis. The Plant cell, 16, 3460-3479.

14. Kauffmann, S., Legrand, M., Geoffroy, P. and Fritig, B. (1987) Biological function of ;pathogenesis-related' proteins: four PR proteins of tobacco have 1,3-beta-glucanase activity. The EMBO journal, 6, 3209-3212.

15. Crute, I.R. and Pink, D. (1996) Genetics and Utilization of Pathogen Resistance in Plants. The Plant cell, 8, 1747-1755.

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

17. Nurnberger, T., Brunner, F., Kemmerling, B. and Piater, L. (2004) Innate immunity in plants and animals: striking similarities and obvious differences. Immunological reviews, 198, 249-266.

18. Gebhardt, C. and Valkonen, J.P. (2001) Organization of genes controlling disease resistance in the potato genome. Annu Rev Phytopathol, 39, 79-102.

19. Friedman, A.R. and Baker, B.J. (2007) The evolution of resistance genes in multi-protein plant resistance systems. Current opinion in genetics & development, 17, 493-499.

20. Grube, R.C., E. R. Radwanski, and M. Jahn. (2000) Comparative genetics of disease resistance within the solanaceae. Genetics, 155, 873-887.

21. van der Vossen, E.A., van der Voort, J.N., Kanyuka, K., Bendahmane, A., Sandbrink, H., Baulcombe, D.C., Bakker, J., Stiekema, W.J. and Klein-Lankhorst, R.M. (2000) Homologues of a single resistance-gene cluster in potato confer resistance to distinct pathogens: a virus and a nematode. Plant J, 23, 567-576.

22. Martin, G.B., Bogdanove, A.J. and Sessa, G. (2003) Understanding the functions of plant disease resistance proteins. Annual review of plant biology, 54, 23-61.

23. Tai, T.H., Dahlbeck, D., Clark, E.T., Gajiwala, P., Pasion, R., Whalen, M.C., Stall, R.E. and Staskawicz, B.J. (1999) Expression of the Bs2 pepper gene confers resistance to bacterial spot disease in tomato. Proceedings of the National Academy of Sciences of the United States of America, 96, 14153-14158.

24. Bent, A.F. (1996) Plant Disease Resistance Genes: Function Meets Structure. The Plant cell, 8, 1757-1771.

25. van Ooijen, G., van den Burg, H.A., Cornelissen, B.J. and Takken, F.L. (2007) Structure and function of resistance proteins in solanaceous plants. Annu Rev Phytopathol, 45, 43-72.

26. Liu, J., Liu, X., Dai, L. and Wang, G. (2007) Recent progress in elucidating the structure, function and evolution of disease resistance genes in plants. Journal of genetics and genomics = Yi chuan xue bao, 34, 765-776.

27. Hulbert, S.H., Webb, C.A., Smith, S.M. and Sun, Q. (2001) Resistance gene complexes: evolution and utilization. Annu Rev Phytopathol, 39, 285-312.

28. Tiffin, P. and Moeller, D.A. (2006) Molecular evolution of plant immune system genes. Trends Genet, 22, 662-670.

29. Kuang, H., Woo, S.S., Meyers, B.C., Nevo, E. and Michelmore, R.W. (2004) Multiple genetic processes result in heterogeneous rates of evolution within the major cluster disease resistance genes in lettuce. The Plant cell, 16, 2870-2894.

30. McDowell, S.A.S. (2006) Recent insights into R gene evolution. Mol Plant Pathology, 7, 11.

31. Martin, G.B., Brommonschenkel, S.H., Chunwongse, J., Frary, A., Ganal, M.W., Spivey, R., Wu, T., Earle, E.D. and Tanksley, S.D. (1993) Map-based cloning of a protein kinase gene conferring disease resistance in tomato. Science (New York, N.Y, 262, 1432-1436.

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

33. Mackey, D., Holt, B.F., 3rd, Wiig, A. and Dangl, J.L. (2002) RIN4 interacts with Pseudomonas syringae type III effector molecules and is required for RPM1-mediated resistance in Arabidopsis. Cell, 108, 743-754.

34. Riely, B.K. and Martin, G.B. (2001) Ancient origin of pathogen recognition specificity conferred by the tomato disease resistance gene *Pto*. Proceedings of the National Academy of Sciences of the United States of America, 98, 2059-2064.

35. Couch, B.C., Spangler, R., Ramos, C. and May, G. (2006) Pervasive purifying selection characterizes the evolution of I2 homologs. Mol Plant Microbe Interact, 19, 288-303.

36. Hawkes, J. (1999) The economic importance of the family Solanaceae. Royal Botanic Gardens, 1, 1.

37. Knapp, S., Bohs, L., Nee, M. and Spooner, D.M. (2004) Solanaceae-a model for linking genomics with biodiversity. Comparative and functional genomics, 5, 285-291.

38. Mueller, L.A., Tanksley, S.D., Giovannoni, J.J., van Eck, J., Stack, S., Choi, D., Kim, B.D., Chen, M., Cheng, Z., Li, C. et al. (2005) The Tomato Sequencing Project, the First Cornerstone of the International Solanaceae Project (SOL). Comparative and functional genomics, 6, 153-158.

39. Cai, D., Kleine, M., Kifle, S., Harloff, H.J., Sandal, N.N., Marcker, K.A., Klein-Lankhorst, R.M., Salentijn, E.M., Lange, W., Stiekema, W.J. et al. (1997) Positional cloning of a gene for nematode resistance in sugar beet. Science (New York, N.Y, 275, 832-834.

40. Shen, K.A., Chin, D.B., Arroyo-Garcia, R., Ochoa, O.E., Lavelle, D.O., Wroblewski, T., Meyers, B.C. and Michelmore, R.W. (2002) Dm3 is one member of a large constitutively expressed family of nucleotide binding site-leucine-rich repeat encoding genes. Mol Plant Microbe Interact, 15, 251-261.

41. Tanksley, S.D., Ganal, M.W., Prince, J.P., de Vicente, M.C., Bonierbale, M.W., Broun, P., Fulton, T.M., Giovannoni, J.J., Grandillo, S., Martin, G.B. et al. (1992) High density molecular linkage maps of the tomato and potato genomes. Genetics, 132, 1141-1160.

42. Tanksley, S.D., Ganal, M.W. and Martin, G.B. (1995) Chromosome landing: a paradigm for map-based gene cloning in plants with large genomes. Trends Genet, 11, 63-68.

43. Giraudat, J., Hauge, B.M., Valon, C., Smalle, J., Parcy, F. and Goodman, H.M. (1992) Isolation of the Arabidopsis ABI3 gene by positional cloning. The Plant cell, 4, 1251-1261.

44. Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A. et al. (2001) The sequence of the human genome. Science (New York, N.Y, 291, 1304-1351.

45. TAG. (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature, 408, 796-815.

46. Tanksley, S.D. and McCouch, S.R. (1997) Seed banks and molecular maps: unlocking genetic potential from the wild. Science (New York, N.Y, 277, 1063-1066.

47. Fox, S., Filichkin, S. and Mockler, T.C. (2009) Applications of ultra-high-throughput sequencing. Methods in molecular biology (Clifton, N.J, 553, 79-108.

48. Appleby, N., Edwards, D. and Batley, J. (2009) New technologies for ultra-high throughput genotyping in plants. Methods in molecular biology (Clifton, N.J, 513, 19-39.

49. Oltvai, Z.N. and Barabasi, A.L. (2002) Systems biology. Life's complexity pyramid. Science (New York, N.Y, 298, 763-764.

50. Bach, R., Iwasaki, Y. and Friedland, P. (1984) Intelligent computational assistance for experiment design. Nucleic acids research, 12, 11-29.

51. Miller, W., Makova, K.D., Nekrutenko, A. and Hardison, R.C. (2004) Comparative genomics. Annual review of genomics and human genetics, 5, 15-56.

52. Mardis, E.R. (2008) The impact of next-generation sequencing technology on genetics. Trends Genet, 24, 133-141.

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

54. Livingstone, K.D., Lackney, V.K., Blauth, J.R., van Wijk, R. and Jahn, M.K. (1999) Genome mapping in capsicum and the evolution of genome structure in the solanaceae. Genetics, 152, 1183-1202.

55. Doganlar, S., Frary, A., Daunay, M.C., Lester, R.N. and Tanksley, S.D. (2002) Conservation of gene function in the solanaceae as revealed by comparative mapping of domestication traits in eggplant. Genetics, 161, 1713-1726.

56. Jaillon, O., Aury, J.M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., Choisne, N., Aubourg, S., Vitulo, N., Jubin, C. et al. (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature, 449, 463-467.

57. Schuler, G.D. (1997) Pieces of the puzzle: expressed sequence tags and the catalog of human genes. Journal of molecular medicine (Berlin, Germany), 75, 694-698.

58. Jaillon, O., Aury, J.M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., Choisne, N., Aubourg, S., Vitulo, N., Jubin, C. et al. (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature, 449, 463-467.

59. Goff, S.A., Ricke, D., Lan, T.H., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H. et al. (2002) A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). Science (New York, N.Y, 296, 92-100.

60. Tuskan, G.A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A. et al. (2006) The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). Science (New York, N.Y, 313, 1596-1604.

61. Paterson, A.H. (2008) Genomics of sorghum. International journal of plant genomics, 2008, 362451.

62. Woodsmall, R.M. and Benson, D.A. (1993) Information resources at the National Center for Biotechnology Information. Bulletin of the Medical Library Association, 81, 282-284.

63. Mueller, L.A., Solow, T.H., Taylor, N., Skwarecki, B., Buels, R., Binns, J., Lin, C., Wright, M.H., Ahrens, R., Wang, Y. et al. (2005) The SOL Genomics Network: a

comparative resource for Solanaceae biology and beyond. Plant physiology, 138, 1310-1317.

64. Huala, E., Dickerman, A.W., Garcia-Hernandez, M., Weems, D., Reiser, L., LaFond, F., Hanley, D., Kiphart, D., Zhuang, M., Huang, W. et al. (2001) The Arabidopsis Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. Nucleic acids research, 29, 102-105.

65. Duvick, J., Fu, A., Muppirala, U., Sabharwal, M., Wilkerson, M.D., Lawrence, C.J., Lushbough, C. and Brendel, V. (2008) PlantGDB: a resource for comparative plant genomics. Nucleic acids research, 36, D959-965.

66. Childs, K.L., Hamilton, J.P., Zhu, W., Ly, E., Cheung, F., Wu, H., Rabinowicz, P.D., Town, C.D., Buell, C.R. and Chan, A.P. (2007) The TIGR Plant Tran*script* Assemblies database. Nucleic acids research, 35, D846-851.

67. Buschges, R., Hollricher, K., Panstruga, R., Simons, G., Wolter, M., Frijters, A., van Daelen, R., van der Lee, T., Diergaarde, P., Groenendijk, J. et al. (1997) The barley Mlo gene: a novel control element of plant pathogen resistance. Cell, 88, 695-705.

68. Romer, P., Hahn, S., Jordan, T., Strauss, T., Bonas, U. and Lahaye, T. (2007) Plant pathogen recognition mediated by promoter activation of the pepper Bs3 resistance gene. Science (New York, N.Y, 318, 645-648.

69. Meyers, B.C., Kozik, A., Griego, A., Kuang, H. and Michelmore, R.W. (2003) Genome-wide analysis of NBS-LRR-encoding genes in Arabidopsis. The Plant cell, 15, 809-834.

70. Song, W.Y., Wang, G.L., Chen, L.L., Kim, H.S., Pi, L.Y., Holsten, T., Gardner, J., Wang, B., Zhai, W.X., Zhu, L.H. et al. (1995) A rece*pto*r kinase-like protein encoded by the rice disease resistance gene, Xa21. Science (New York, N.Y, 270, 1804-1806.

71. Liu, J., Liu, X., Dai, L. and Wang, G. (2007) Recent progress in elucidating the structure, function and evolution of disease resistance genes in plants. Journal of genetics and genomics = Yi chuan xue bao, 34, 765-776.

72. Toubart, P., Desiderio, A., Salvi, G., Cervone, F., Daroda, L. and De Lorenzo, G. (1992) Cloning and characterization of the gene encoding the endopolygalacturonase-inhibiting protein (PGIP) of Phaseolus vulgaris L. Plant J, 2, 367-373.

73. Gao, H., Narayanan, N.N., Ellison, L. and Bhattacharyya, M.K. (2005) Two classes of highly similar coiled coil-nucleotide binding-leucine rich repeat genes isolated from the Rps1-k locus encode Phytophthora resistance in soybean. Mol Plant Microbe Interact, 18, 1035-1045.

74. Zhang, L., Peek, A.S., Dunams, D. and Gaut, B.S. (2002) Population genetics of duplicated disease-defense genes, hm1 and hm2, in maize (Zea mays ssp. mays L.) and its wild ancestor (Zea mays ssp. parviglumis). Genetics, 162, 851-860.

75. Halterman, D.A. and Wise, R.P. (2004) A single-amino acid substitution in the sixth leucine-rich repeat of barley MLA6 and MLA13 alleviates dependence on RAR1 for disease resistance signaling. Plant J, 38, 215-226.

76. Brueggeman, R., Rostoks, N., Kudrna, D., Kilian, A., Han, F., Chen, J., Druka, A., Steffenson, B. and Kleinhofs, A. (2002) The barley stem rust-resistance gene Rpg1 is a novel disease-resistance gene with homology to rece*pto*r kinases. Proceedings of the National Academy of Sciences of the United States of America, 99, 9328-9333.

77. Taler, D., Galperin, M., Benjamin, I., Cohen, Y. and Kenigsbuch, D. (2004) Plant eR genes that encode photorespiratory enzymes confer resistance against disease. The Plant cell, 16, 172-184.

78. Shen, K.A., Chin, D.B., Arroyo-Garcia, R., Ochoa, O.E., Lavelle, D.O., Wroblewski, T., Meyers, B.C. and Michelmore, R.W. (2002) Dm3 is one member of a large constitutively

expressed family of nucleotide binding site-leucine-rich repeat encoding genes. Mol Plant Microbe Interact, 15, 251-261.

79. Cai, D., Kleine, M., Kifle, S., Harloff, H.J., Sandal, N.N., Marcker, K.A., Klein-Lankhorst, R.M., Salentijn, E.M., Lange, W., Stiekema, W.J. et al. (1997) Positional cloning of a gene for nematode resistance in sugar beet. Science (New York, N.Y, 275, 832-834.

80. Lawrence, G.J., Finnegan, E.J., Ayliffe, M.A. and Ellis, J.G. (1995) The L6 gene for flax rust resistance is related to the Arabidopsis bacterial resistance gene RPS2 and the tobacco viral resistance gene N. The Plant cell, 7, 1195-1206.

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

82. Dodds, P.N., Lawrence, G.J. and Ellis, J.G. (2001) Six amino acid changes confined to the leucine-rich repeat beta-strand/beta-turn motif determine the difference between the P and P2 rust resistance specificities in flax. The Plant cell, 13, 163-178.

83. Sayers, E.W., Barrett, T., Benson, D.A., Bolton, E., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., Dicuccio, M., Federhen, S. et al. (2009) Database resources of the National Center for Biotechnology Information. Nucleic acids research.

84. Stajich, J.E., Block, D., Boulez, K., Brenner, S.E., Chervitz, S.A., Dagdigian, C., Fuellen, G., Gilbert, J.G., Korf, I., Lapp, H. et al. (2002) The Bioperl toolkit: Perl modules for the life sciences. Genome research, 12, 1611-1618.

85. Codd, E.F. (1998) A relational model of data for large shared data banks. 1970. MD Comput, 15, 162-166.

86. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. Journal of molecular biology, 215, 403-410.

87. Iseli, C., Jongeneel, C.V. and Bucher, P. (1999) ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB, 138-148.

88. Hunter, S., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L. et al. (2009) InterPro: the integrative protein signature database. Nucleic acids research, 37, D211-215.

89. Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research, 32, 1792-1797.

90. Knudsen, B. and Miyamoto, M.M. (2003) Sequence alignments and pair hidden Markov models using evolutionary history. Journal of molecular biology, 333, 453-460.

91. Zhang, Z. and Wood, W.I. (2003) A profile hidden Markov model for signal peptides generated by HMMER. Bioinformatics (Oxford, England), 19, 307-308.

92. Sturn, A., Quackenbush, J. and Trajanoski, Z. (2002) Genesis: cluster analysis of microarray data. Bioinformatics (Oxford, England), 18, 207-208.

93. Kerkhoven, R., van Enckevort, F.H., Boekhorst, J., Molenaar, D. and Siezen, R.J. (2004) Visualization for genomics: the Microbial Genome Viewer. Bioinformatics (Oxford, England), 20, 1812-1814.

94. Guindon, S., Lethiec, F., Duroux, P. and Gascuel, O. (2005) PHYML Online--a web server for fast maximum likelihood-based phylogenetic inference. Nucleic acids research, 33, W557-559.

95. Le, S.Q. and Gascuel, O. (2008) An improved general amino acid replacement matrix. Molecular biology and evolution, 25, 1307-1320.

96. Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R. et al. (2007) Clustal W and Clustal X version 2.0. Bioinformatics (Oxford, England), 23, 2947-2948.

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

98. Kall, L., Krogh, A. and Sonnhammer, E.L. (2007) Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server. Nucleic acids research, 35, W429-432.

99. Carver, T., Berriman, M., Tivey, A., Patel, C., Bohme, U., Barrell, B.G., Parkhill, J. and Rajandream, M.A. (2008) Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. Bioinformatics (Oxford, England), 24, 2672-2676.

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

101. Curtis, M.D. and Grossniklaus, U. (2003) A gateway cloning vector set for high-throughput functional analysis of genes in planta. Plant physiology, 133, 462-469.

102. Means, T.K., Golenbock, D.T. and Fenton, M.J. (2000) The biology of Toll-like rece*ptor*s. Cytokine & growth factor reviews, 11, 219-232.

103. DeYoung, B.J. and Innes, R.W. (2006) Plant NBS-LRR proteins in pathogen sensing and host defense. Nature immunology, 7, 1243-1249.

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

105. Morillo, S.A. and Tax, F.E. (2006) Functional analysis of rece*ptor*-like kinases in monocots and dicots. Current opinion in plant biology, 9, 460-469.

106. Fritz-Laylin, L.K., Krishnamurthy, N., Tor, M., Sjolander, K.V. and Jones, J.D. (2005) Phylogenomic analysis of the rece*ptor*-like proteins of rice and Arabidopsis. Plant physiology, 138, 611-623.

107. Biron, D.G., Brun, C., Lefevre, T., Lebarbenchon, C., Loxdale, H.D., Chevenet, F., Brizard, J.P. and Thomas, F. (2006) The pitfalls of proteomics experiments without the correct use of bioinformatics tools. Proteomics, 6, 5577-5596.

108. Searls, D.B. (2000) Using bioinformatics in gene and drug discovery. Drug Discovery Today, 5, 8.

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

110. Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., Gavin, O.L., Gunasekaran, P., Ceric, G., Forslund, K. et al. (2009) The Pfam protein families database. Nucleic acids research.

111. Masle, J., Gilmore, S.R. and Farquhar, G.D. (2005) The ERECTA gene regulates plant transpiration efficiency in Arabidopsis. Nature, 436, 866-870.

112. Frank, S.A. (1992) Models of plant-pathogen coevolution. Trends Genet, 8, 213-219.

113. Ausubel, F.M. (2005) Are innate immune signaling pathways in plants and animals conserved? Nature immunology, 6, 973-979.

114. Belvin, M.P. and Anderson, K.V. (1996) A conserved signaling pathway: the Drosophila toll-dorsal pathway. Annual review of cell and developmental biology, 12, 393-416.

115. Pandey, J.P. (2007) Genetics of immunoglobulins. Medical immunology, 1, 10.

116. Dunne, A. and O'Neill, L.A. (2003) The interleukin-1 rece*ptor*/Toll-like rece*ptor* superfamily: signal transduction during inflammation and host defense. Sci STKE, 2003, re3.

117. Kinoshita, T., Cano-Delgado, A., Seto, H., Hiranuma, S., Fujioka, S., Yoshida, S. and Chory, J. (2005) Binding of brassinosteroids to the extracellular domain of plant rece*ptor* kinase BRI1. Nature, 433, 167-171.

118. Chisholm, S.T., Coaker, G., Day, B. and Staskawicz, B.J. (2006) Host-microbe interactions: shaping the evolution of the plant immune response. Cell, 124, 803-814.

119. Brenner, S.E. (1999) Errors in genome annotation. Trends Genet, 15, 132-133.

120. Brandwagt, B.F., Mesbah, L.A., Takken, F.L., Laurent, P.L., Kneppers, T.J., Hille, J. and Nijkamp, H.J. (2000) A longevity assurance gene homolog of tomato mediates resistance to Alternaria alternata f. sp. lycopersici toxins and fumonisin B1. Proceedings of the National Academy of Sciences of the United States of America, 97, 4961-4966.

121. Axelsen, K.B. and Palmgren, M.G. (2001) Inventory of the superfamily of P-type ion pumps in Arabidopsis. Plant physiology, 126, 696-706.

122. Wortman, J.R., Haas, B.J., Hannick, L.I., Smith, R.K., Jr., Maiti, R., Ronning, C.M., Chan, A.P., Yu, C., Ayele, M., Whitelaw, C.A. et al. (2003) Annotation of the Arabidopsis genome. Plant physiology, 132, 461-468.

123. Brunner, A.M., Busov, V.B. and Strauss, S.H. (2004) Poplar genome sequence: functional genomics in an ecologically dominant plant species. Trends Plant Sci, 9, 49-56.

124. Griffiths, S., Dunford, R.P., Coupland, G. and Laurie, D.A. (2003) The evolution of CONSTANS-like gene families in barley, rice, and Arabidopsis. Plant physiology, 131, 1855-1867.

125. Schauser, L., Wieloch, W. and Stougaard, J. (2005) Evolution of NIN-like proteins in Arabidopsis, rice, and Lotus japonicus. Journal of molecular evolution, 60, 229-237.

126. M. Ziemann, M.B.a.S.Z. (2009) Origin and Diversification of Land Plant CC-Type Glutaredoxins. Genome Biology and evolution, 2009.

127. Fulton, T.M., Van der Hoeven, R., Eannetta, N.T. and Tanksley, S.D. (2002) Identification, analysis, and utilization of conserved ortholog set markers for comparative genomics in higher plants. The Plant cell, 14, 1457-1467.

128. Michelmore, R.W. and Meyers, B.C. (1998) Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. Genome research, 8, 1113-1130.

129. Ballvora, A., Ercolano, M.R., Weiss, J., Meksem, K., Bormann, C.A., Oberhagemann, P., Salamini, F. and Gebhardt, C. (2002) The R1 gene for potato resistance to late blight (Phytophthora infestans) belongs to the leucine zipper/NBS/LRR class of plant resistance genes. Plant J, 30, 361-371.

130. Gomez-Gomez, L. and Boller, T. (2000) FLS2: an LRR rece*pto*r-like kinase involved in the perception of the bacterial elicitor flagellin in Arabidopsis. Molecular cell, 5, 1003-1011.

131. Zhou, J., Tang, X. and Martin, G.B. (1997) The *Pto* kinase conferring resistance to tomato bacterial speck disease interacts with proteins that bind a cis-element of pathogenesis-related genes. The EMBO journal, 16, 3207-3218.

132. Salmeron, J.M., Oldroyd, G.E., Rommens, C.M., Scofield, S.R., Kim, H.S., Lavelle, D.T., Dahlbeck, D. and Staskawicz, B.J. (1996) Tomato Prf is a member of the leucine-rich repeat class of plant disease resistance genes and lies embedded within the *Pto* kinase gene cluster. Cell, 86, 123-133.

133. Ohlrogge, J. and Benning, C. (2000) Unraveling plant metabolism by EST analysis. Current opinion in plant biology, 3, 224-228.

134. Ouzounis, C.A. and Karp, P.D. (2002) The past, present and future of genome-wide re-annotation. Genome biology, 3, COMMENT2001.

135. Devos, D. and Valencia, A. (2001) Intrinsic errors in genome annotation. Trends Genet, 17, 429-431.

136. Stein, L. (2001) Genome annotation: from sequence to biology. Nature reviews, 2, 493-503.

137. Gilroy, S. and Trewavas, A. (2001) Signal processing and transduction in plant cells: the end of the beginning? Nat Rev Mol Cell Biol, 2, 307-314.

7. Appendice A

Mappa R del genoma del pomodoro divisa in cromosomi.

Figura 44. Leggenda

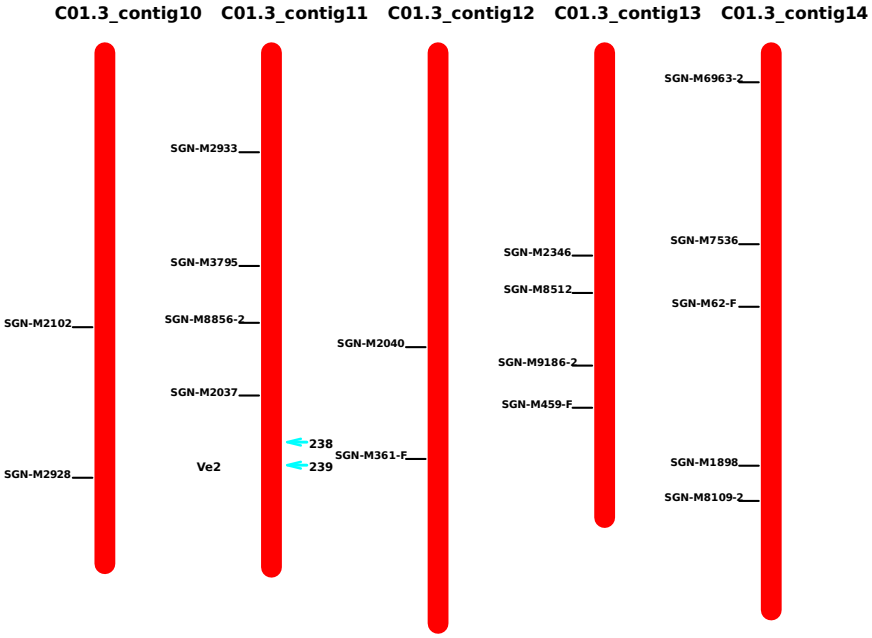


Figura 45. Cromosoma 1

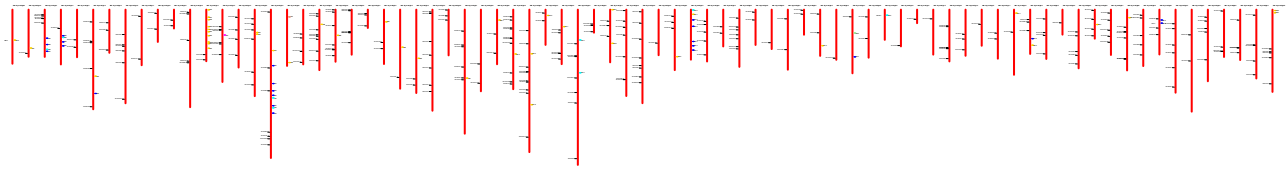


Figura 46. Cromosoma 2

Figura 47. Cromosoma 3

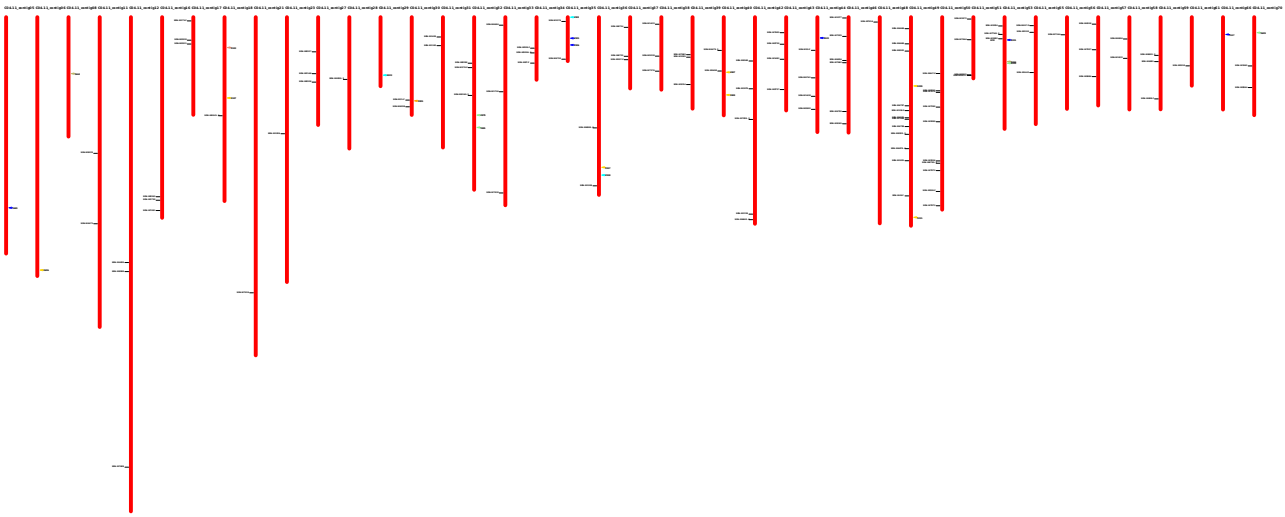


Figura 48. Cromosoma 4

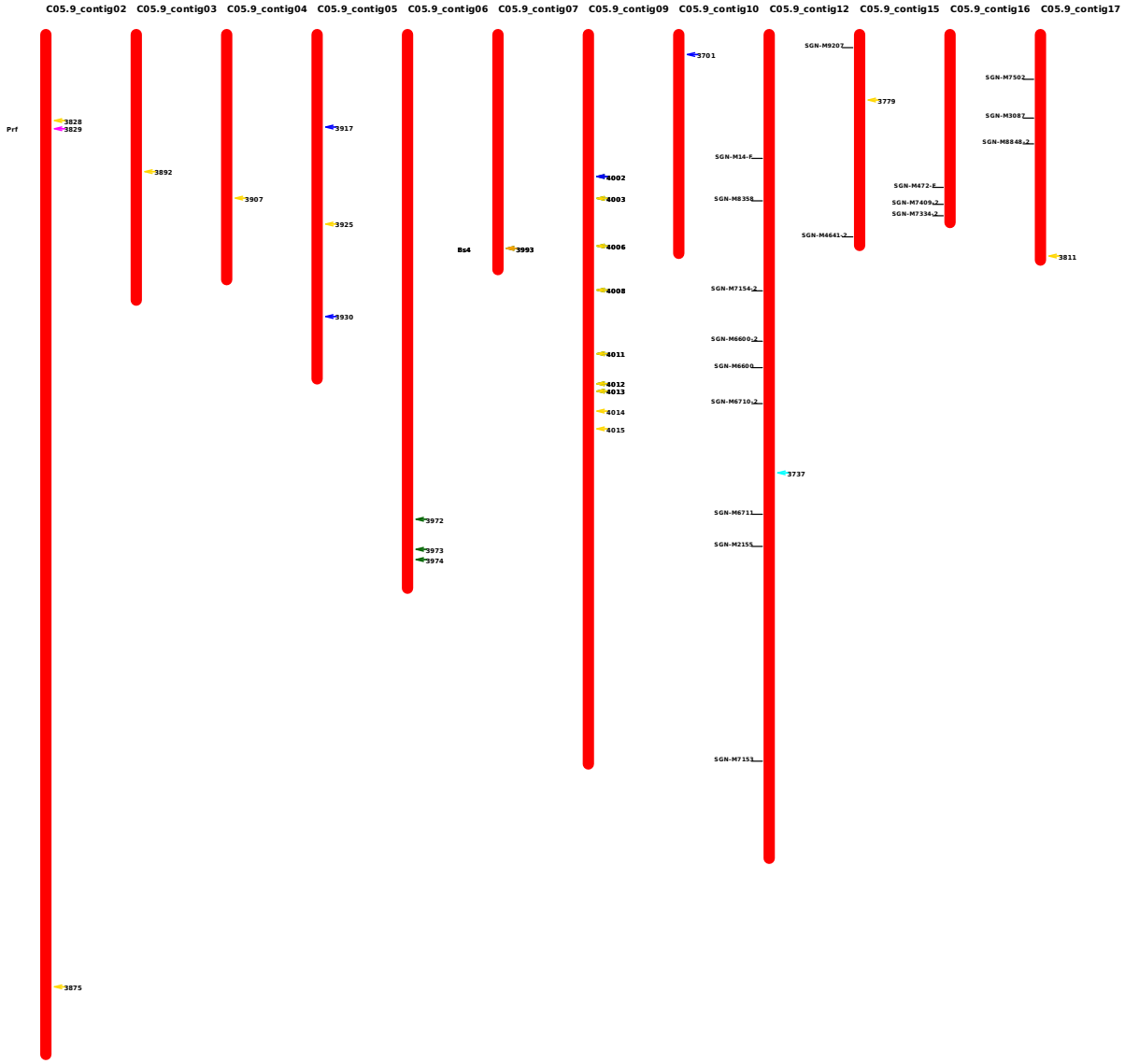


Figura 49. Cromosoma 5

Figura 50. Cromosoma 6

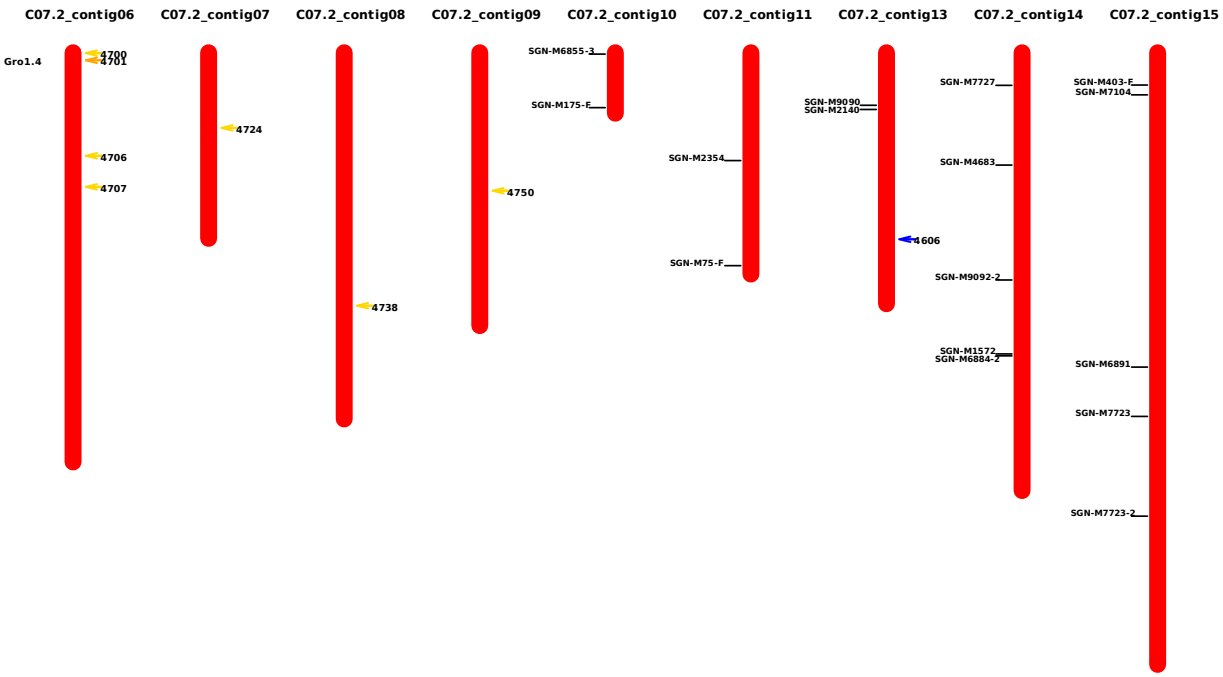


Figura 51. Cromosoma 7

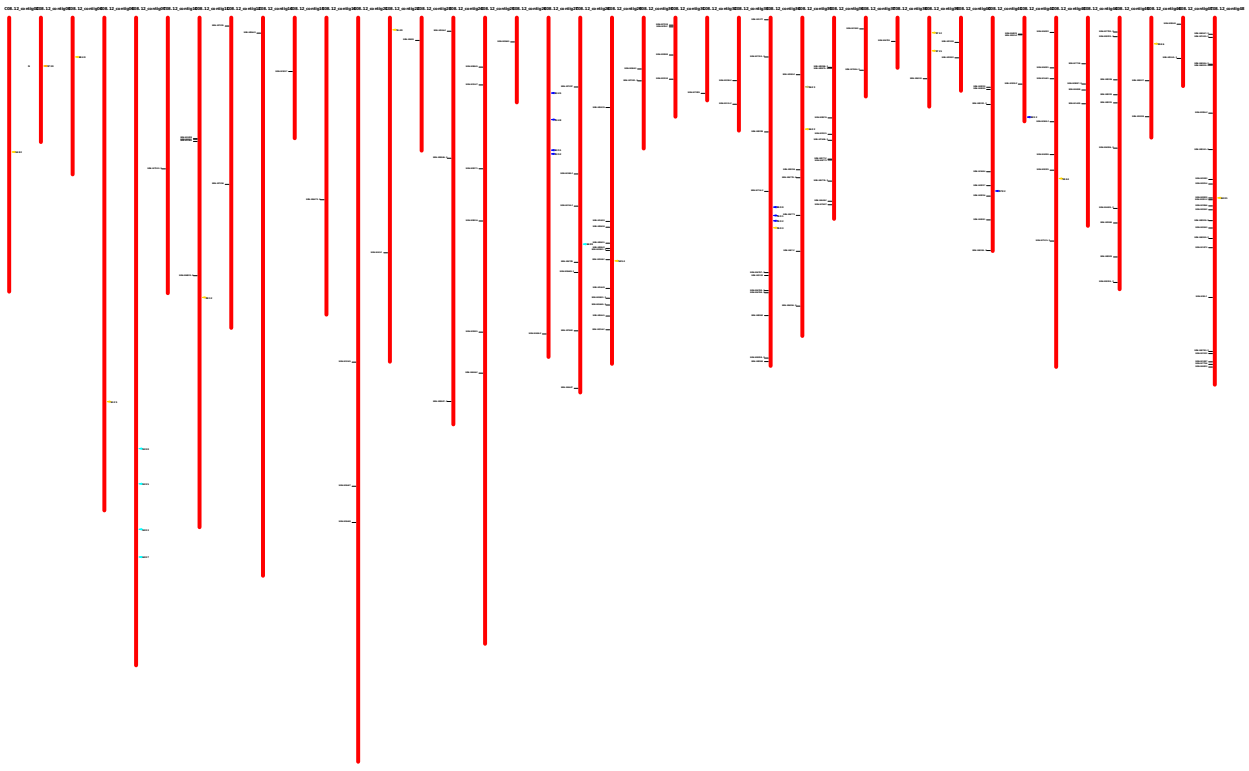


Figura 52. Cromosoma 8

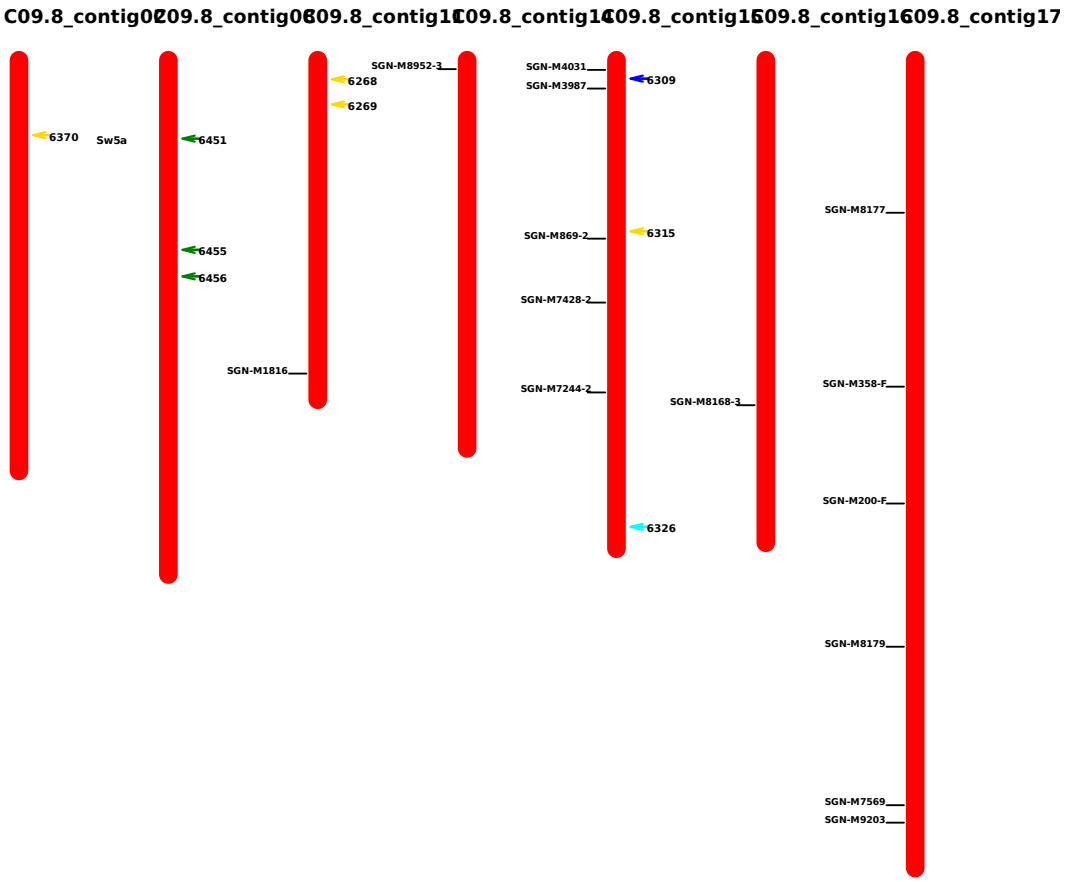


Figura 53. Cromosoma 9

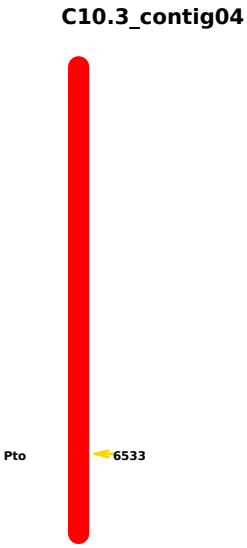


Figura 54. Cromosoma 10

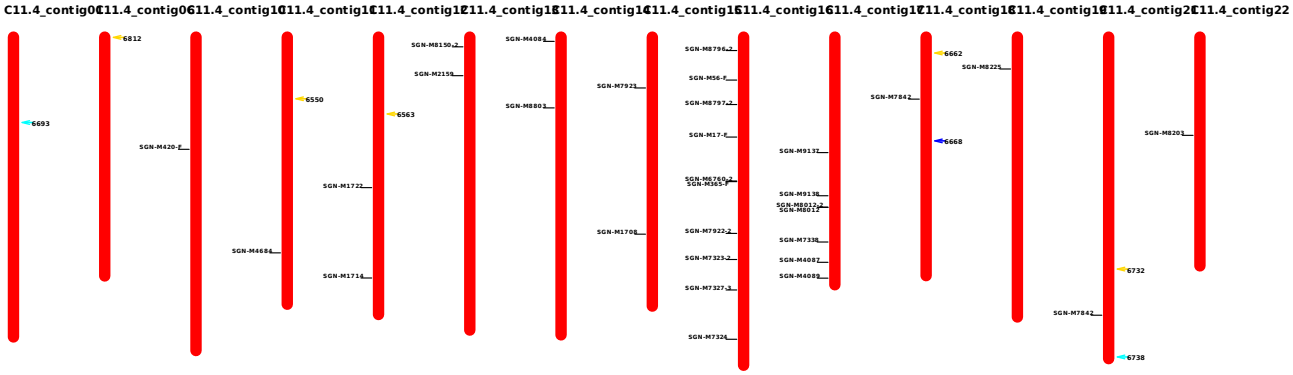


Figura 55. Cromosoma 11

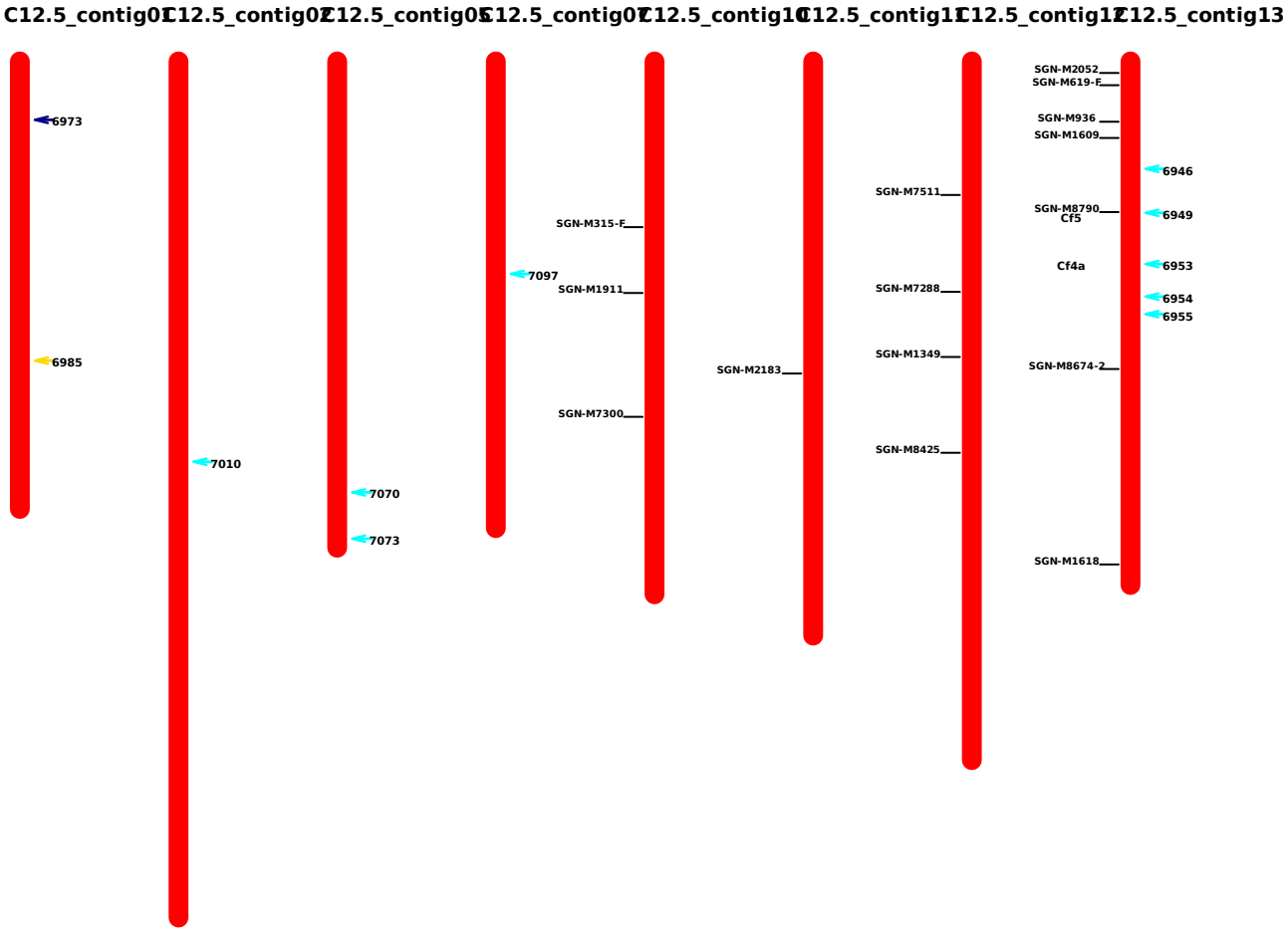


Figura 56. Cromosoma 12