Abstract

**Aim:** The nucleotide binding site leucine-rich repeat (NBS-LRR) genes are responsive to pathogen strike in plants. This study focused on identifying the loci specific NBS-LRR gene encoding regions in tomato at whole genome level.

**Methodology:** Major computational challenges in analyzing large genomic data using existing analytical tools were limited by the amount of memory used for reading large data. In this study, a specific algorithm was developed to identify a signature pattern associated with stress tolerant coding regions using stream readers for reading whole tomato genome, chromosome wise, to locate NBS-LRR coding sequences. The computer program reads chromosome wise data and extracts the potential stress tolerant coding regions. It was found that more than 300 disease resistance protein coding regions were found across all chromosome, specifically in chromosome 12 more NBS-LRR concentration were found and their respective locus were identified.

**Results:** The computer program reads chromosome wise data and extracts the potential stress tolerant coding regions. It was found that more than 300 disease resistance protein coding regions were found across all chromosome, specifically in chromosome 12 more NBS-LRR concentration were found and their respective locus were identified.

**Interpretation:** The identified disease resistance protein coding regions, specifically the NBS-LRR coding regions and their loci can be useful for plant breeders to select parental lines for developing plants tolerant to disease and pest invasion.

**Key words:** Computational methods, Environmental pathogens, NBS-LRR genes, Tomato genome

Introduction

Plants are prone to suffer from several biotic and abiotic stresses, which affects plant growth and reduces fruit yield and quality. However, plants can adapt to these stresses where different genes are over expressed in plants to improve their tolerance to environmental stresses. Tomato (*Solanum lycopersicum* L.) is highly consumed vegetable and agriculturally significant crop of family Solanaceae, however, highly vulnerable to several diseases that reduces the fruit quality and yield. The recently sequenced tomato genome have helped in the identification of important genes of Solanaceae family. The size of tomato genome is around 948 Mb, distributed across 12 different chromosomes, the smallest being chromosome 6 and biggest chromosome 1; the average is 60 million base pairs in size.

Plants are provided with inherent immune system, which acts as a defense system against pathogen and pest attack. Recent studies on plant research and findings from tomato has helped us to expand our ability to understand the immune system of plant and reveal the mechanism used by pathogens to control its effectiveness (Jones and Dangl, 2006). In plant genomes, the resistant genes (R genes) provide resistance against pathogen invasion by generating R proteins. R protein interacts directly with avirulence (Avr) gene, guards degradation of proteins by Avr gene, detects pathogen associated pattern (Zipfel et al., 2014) and encodes enzyme for toxin degradation.

The essential pathogen defense molecules are wall associated kinase, pattern recognition receptors, receptors with nucleotide-binding domain and leucine-rich repeat. The major portion of disease resistance R genes contain NB-LRR proteins, comprised of NBARC domains (nucleotide-binding adaptor shared by APAF-1, Resistance genes, and CED-4) in central region and leucine-rich repeat (LRR) domains in C-terminal regions (Seo et al., 2016). NBS-LRR protein in plants are large in number and originated long ago and encoded by one of the major gene families in plant system, with several hundred replicas in a genome (He et al., 2004). NBS-LRR proteins are some of the known proteins in plants, ranging from 860 to 1,900 amino-acids. There are in approximately 159 NBS-LRR-encoding genes *Arabidopsis* 400 in *Oryza sativa* (Meyers et al., 2003), 57 in cucumber (Marone et al., 2013). Distribution of NBS-LRR genes are not random, some chromosome have highest number and closely clustered as a result of duplication from common ancestor (Seo et al., 2016). These duplications dominate the expansion of NBS with both genes and replication events across several chromosomes.

Most of the intracellular receptors for immune system in plants belong to NBS-LRR encoding proteins (Eitasand Dangl, 2010; Lee and Yeom, 2015). NBS-LRR proteins play a key role through a network of signaling pathway which induces a chain of plant resistance responses, like activation of oxidative spurt, ion and calcium fluxes, mitogen associated protein-kinase cascade, initiation of pathogen resistance responsive gene and hypersensitivereaction (Belkhadir et al., 2004; Pedley et al., 2005). These responses will destruct pathogen and prevent subsequent infection. At genomic level, understanding the expression of a specific protein under a particular stress can provide a base for recognizing genes. Highly-conserved proteins are needed in fundamental cellular functions, stability or reproduction. Comparative genomics is an important area in genomic research. Comparing the sequences of relative organisms help in locating genes (Mahalakshmi and Ortiz, 2001).

The mutation rate is slower in proteins encoding genes and functional elements than other areas of genome due to natural selection process. Therefore, the changes occurring in functional areas will have less impact compared to the changes in the sequences of other areas, the reason being conservation of sequences due to evolutionary pressure. Thus, genes can be easily detected by comparing the genomic sequences as they are evolutionary conserved, across in Solanaceae crops (Thomas et al., 2013). Tomato diseases are caused by several pathogenic organisms, and to defend against pathogen attack, it has developed an immune system, where pathogen effectors and receptor proteins such as R protein (resistance proteins) play a major role (Bai et al., 2018). Identifying the resistance protein in whole genome is thus essential for developing varieties tolerant to diseases and pathogen invasion.

Materials and Methods

Genome sequence of tomato was completed and published in 2012. The genome sequence of tomato (*S. lycopersicum*) is a valuable resource for analyzing the genomic organization of resistance genes in this crop. Hence, a computer program was developed to search all chromosomes of *S. lycopersicum* for a specific pattern having a signature associated with stress tolerant coding sequences. The program reads the genomic data chromosome wise using stream readers for searching the desired signature pattern. On successful pattern match, the sequences along with signature pattern were extracted for comparative analysis to find similarity with published proteins. The sequences having more than 700 bp of identity with published genes were filtered. These potential sequences which were converted to FASTA format was uploaded to NCBI’s Conserved Domain search program to process as batch files. This interface searches the Conserved Domain Database against multiple protein sequences, and allows to download the results for complete protein sets. The output generated was in tab delimited set of conserved domain matches located on each sequence against input sequences displaying the highest ranking protein sequences with their superfamily. The sequences matching NBS-LRR proteins were grouped together. Further, the locus of matched NBS-LRR protein encoding regions were identified by the program along with its conserved domains.

Results and Discussion

Searching the whole genome sequence and identifying the desired sequences is cumbersome. Large DNA sequence
files require memory resources and significant time for processing and analyzing, which can be a real problem for several operating systems (Gagniuc et al., 2012), further increasing pressure on the computing system. However, efficient computational methods can significantly reduce the volume of laboratory work (Thomas et al., 2017). Genes with similar expression patterns are found to be grouped together in genome, as evident by a signature pattern (Yang et al., 2014).

Developed computer program searches specific pattern associated with stress tolerant coding regions. The Conserved Domain Database search resulted in more than 300 plus disease resistance protein coding regions across all chromosomes (Table 1). Further, it was classified according to specific disease resistance, viz., Tospovirus resistance, Nematode resistance, NBS-LRR-resistance, Late Blight resistance, Disease resistant genes etc. However, highly concentrated NBS-LRR coding sequences were found in chromosome 12, from which the sequences having more than 700 bp identity with published NBS_LRR proteins were searched and 36 numbers of NBS-LRR regions with high similarity were found (Fig. 1). It was surprising that NBS-LRR coding sequences were found to be closely clustered in chromosome 12 of S. lycopersicum whose length was 64 million bp which is equivalent to 64.8 megabytes (MB) of

![Fig. 1: Graphic representation of distribution of NBS-LRR coding sequences in close clusters of chromosome 12.](image)

![Fig. 2: Graphic chart displays NBS-LRR coding sequences matching length (y-axis) and its location (x-axis) on chromosome 12 of S. lycopersicum.](image)
The reason for the closely spaced NBS-LRR regions could be tandem duplication events occurred during evolution. Luo et al. (2012) reported that most NBS-LRRs are grouped as clusters in genome. The chromosomal positioning in soya bean suggest duplication events happening at varying time periods, and an extra pericentromeric positioning may enable their rapid evolution (Nepal et al., 2015). The amount of NBS LRR genes in different genomes of plant differs significantly, from less than 100 to greater than 1000 (Yue et al., 2012; Xia et al., 2015). Nevertheless, many exceptions are reported for low copy number of NBS-LRRs in papaya (54), cucumber (70), melon (75) and watermelon (55) (Porter et al., 2009; Lin et al., 2013; González et al., 2015). However, many times large number of NBS-LRRs proteins are fatal to plant cells.

Fig. 3: CIRCOS image shows 36 NBS-LRR protein encoding sequences found in chromosome 12 of S. lycopersicum along with other stress tolerant proteins with its respective superfamily.
Table 1: Potential disease tolerant proteins and its presence in chromosome

<table>
<thead>
<tr>
<th>Name of protein</th>
<th>Presence in chromosome</th>
<th>Total number of potential genes in all chromosomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tospovirus Resistance Protein C</td>
<td>All Chromosomes 1 to 12</td>
<td>162</td>
</tr>
<tr>
<td>Nematode-Resist-Protein</td>
<td>All chromosomes 1 to 12</td>
<td>76</td>
</tr>
<tr>
<td>NBS-LRR-Resist-Protein</td>
<td>CHR-12</td>
<td>36</td>
</tr>
<tr>
<td>Late Blight Resist-Protein</td>
<td>CHR-2,3,4,5,6,8,9,10 &amp; 11</td>
<td>26</td>
</tr>
<tr>
<td>Tospovirus-Resist-Protein-A</td>
<td>CHR-7,8,9,10 &amp; 12</td>
<td>14</td>
</tr>
<tr>
<td>Disease-Resist-Gene</td>
<td>CHR-12</td>
<td>11</td>
</tr>
<tr>
<td>Disease-Resist-prot-R3A</td>
<td>CHR-11</td>
<td>6</td>
</tr>
<tr>
<td>Disease-Resist-Prot1-RPP13</td>
<td>CHR-11</td>
<td>3</td>
</tr>
<tr>
<td>Disease-Resist-Prot-RPM1</td>
<td>CHR-11</td>
<td>2</td>
</tr>
<tr>
<td>Resistance-Protein-F</td>
<td>CHR-9</td>
<td>2</td>
</tr>
<tr>
<td>Resistance-Protein-F-Var-1,2</td>
<td>Total</td>
<td>340</td>
</tr>
</tbody>
</table>

Considerable progress has been made in understanding the inherent immune system of plants. However, the NBS-LRR protein regulatory network and physiological basis of NBS–LRR protein activities is not completely understood. Comprehensive analyses are required to understand the presence and distribution of disease resistance protein throughout genome, specifically NBS-LRR protein coding regions as these proteins play a vital role in protecting plants from invasive pathogens. Hence, study at whole genome level and distribution of these proteins using computational techniques is rapid and cost effective. This will definitely provide better opportunity for plant breeders to locate these encoding proteins in crops to select them as parents for developing varieties tolerant to diseases and pest invasion.

Acknowledgment

The authors wish to thank PI of ICAR-Network project on Agricultural Bioinformatics and Computational Biology for funding this research work under “CAB in project”, ICAR-IASRI, New Delhi.

References


Lin, X., Y. Zhang, H Kiangs and J. Chen: Frequent loss of lineages and
deficient duplications accounted for low copy number of disease resistance genes in Cucurbitaceae. BMC Genomics, 14, 335 (2013).


