Comparative genome analysis of *Solanum lycopersicum* and *Solanum tuberosum*

Rohit Lall¹, George Thomas¹, Satendra Singh², Archana Singh³ & Gulshan Wadhwa⁴*

¹Department of Molecular and Cellular Engineering, SHIATS, Allahabad-211007; ²Department of Computational Biology & Bioinformatics, SHIATS, Allahabad-211007; ³Division of Biochemistry, Indian Agricultural Research Institute, New Delhi-110012; ⁴Department of Biotechnology, Ministry of Science and Technology, New Delhi – 110003; Gulshan Wadhwa - Email: gulshan@dbt.nic.in; Phone: +91-9811301820; Fax: +91-11-24362884; *Corresponding author

Received October 05, 2013; Accepted October 30, 2013; Published November 11, 2013

Abstract: *Solanum lycopersicum* and *Solanum tuberosum* are agriculturally important crop species as they are rich sources of starch, protein, antioxidants, lycopene, beta-carotene, vitamin C, and fiber. The genomes of *S. lycopersicum* and *S. tuberosum* are currently available. However the linear strings of nucleotides that together comprise a genome sequence are of limited significance by themselves. Computational and bioinformatics approaches can be used to exploit the genomes for fundamental research for improving their varieties. The comparative genome analysis, Pfam analysis of predicted reviewed paralogous proteins was performed. It was found that *S. lycopersicum* proteins belong to more families, domains and clans in comparison with *S. tuberosum*. It was also found that mostly intergenic regions are conserved in two genomes followed by exons, intron and UTR. This can be exploited to predict regions between genomes that are similar to each other and to study the evolutionary relationship between two genomes, leading towards the development of disease resistance, stress tolerance and improved varieties of tomato.

**Key words:** *S. lycopersicum, S. tuberosum, genome.*

**Background:**
*Solanaceae* family represent important family in agriculture as it is one of the major source of edible fruits *Solanum lycopersicum, Solanum tuberosum* and *Nicotiana tabacum*. Tomato fruits are the second most consumed vegetable after potatoes, and are a globally important dietary source of lycopene, beta-carotene, vitamin C, and fiber. Potato contributes to dietary intake of starch, protein, antioxidants, and vitamins. In addition to its agricultural value and due to its diploid genetics and inbreeding potential, tomato is a widely used model species for fundamental research on subjects including fruit development and pathogen response [1].

The developments in sequencing technologies are providing genome sequences of different species. Deciphering a genome sequence, that is, determining the linear order of nucleotides for each chromosome in the genome, allows molecular biologists to understand and manipulate this blueprint. For plants in particular, this in turn enables breeders to more efficiently engineer solutions for crop improvement to respond to the growing demand for food and energy from modern society [2].

The genome draft of Tomato and Potato is now available in plant databases. The nuclear genome of potato and tomato consists of twelve chromosomes. Their genomes are expected to measure approximately 840 Mb and 950 Mb in size, respectively [3-5].

The availability of their genome sequences will provide the community with a first glimpse into genome evolution of *Solanaceae* (and Asterids in general) and will impact both fundamental research and breeding strategies in these species for the coming years.

The aim of the present research work was to predict paralogous proteins in Tomato proteome and to carry out comparative
BIOINFORMATION

genome analysis of Tomato and Potato to uncover various genomic features of two genomes and to gain insight the similarity and differences between two genomes.

Figure 1: Pfam comparison of *S. lycopersicum* and *S. tuberosum* protein sequences.

**Prediction of Paralogous Proteins in *S. lycopersicum* and *S. tuberosum* Genome**

The reviewed set of proteins sequences of *S. lycopersicum* and *S. tuberosum* was retrieved from the Uniprot Database in FASTA format. The all against all database searches by using the genomic BLAST-P available at NCBI server was used to predict paralogous protein in the selected set of protein sequences [7-8]. In case of all against all search, a comparison was made in which every predicted protein sequence was used as a query in a similarity search against a database composed of the rest of the self-proteome, and the significant matches were identified by a low E-value. Since many proteins comprise different combinations of a common set of domains, proteins that align more than 80% of their lengths for query and subject were selected. After this filtration only those alignments were selected which give the sequence identity more than 60%.

**Families, domain and repeats for paralogous protein sequences in *S. lycopersicum* and *S. tuberosum***

For the purpose of functional annotation and to investigate the gene family expansion, the identified set of paralogous proteins was used to search the protein families by using the Pfam search. Each family is represented by multiple sequence alignments and Hidden Markov models (HMMs) [9]. The paralogous protein dataset was submitted at Pfam server which predicted the protein families, motifs, repeats and clans at the default pfam parameter (http://pfam.sanger.ac.uk/).

**Results and Discussion**: After performing the all against all searches for all reviewed protein sequences of *S. lycopersicum* and *S. tuberosum* it was found that 60 paralogous proteins present in *S. lycopersicum* and while 110 were present in *S. tuberosum*. All predicted paralogous proteins of *S. lycopersicum* and *S. tuberosum* can be retrieved by using accession number given in Table 1 & 2 (see supplementary material). The predicted paralogous proteins belong to different family having different domain and repeats. For the purpose of functional annotation and to investigate the gene family expansion, the identified set of paralogous proteins was used to search the protein families by using the Pfam search.
Pfam analysis of S. lycopersicum and S. tuberosum protein sequences

It was found that most of the identified proteins belong to different families, domains and clans in S. lycopersicum and S. tuberosum protein sequences Table 3 (see supplementary material). But also there are proteins having no clans (Figure 1). Proteins contain functional units known as domains and various combinations of domains results in different protein formations. Therefore identification of domains in proteins is essential for giving insights into their function. Pfam also generates higher-level groupings of related families, known as clans. A protein belongs to different families, domains and clans may be due to proteins family expansion and adaptations by the genomes [10].

It was found that S. lycopersicum proteins belong to more families, domains and clans in comparison with S. tuberosum. But also there are proteins having no clans.

Comparative genomics Solanum lycopersicum and Solanum tuberosum

The comparison of the genomic regions of S. lycopersicum and S. tuberosum was performed. It was found that the genome of two selected plants have conserved, non conserved and also different genomic compassions and different levels. But there are other areas also where difference in conservation was noted.

It was found that mostly intergenic regions are conserved in two genomes followed by exons, intron (they are found in the genes of most organisms and many viruses, and can be located in a wide range of genes) and UTR (untranslated region) (Figure 2).

An Intergenic region (sometimes also referred to as junk DNA) represent stretch of DNA sequences located between genes. Their function is still unknown but sometime they are involve in regulation of gene expressions (these regions do contain functionally important elements such as promoters and enhancers).

The comparative alignment of genomic regions of S. lycopersicum and S. tuberosum revealed that it was found that there are regions where only conserved part is present in two genomes (Figure 3). Along with this there are regions were conserved regions, untranslated region (UTR) exons present together without any non aligned region (Figure 4). Non aligned Genomic region are also found in the alignment two genomes (Figure 5).

Once the elements in a genome sequence have been identified, the next step is to assign to them a plausible biological function. Computational inference of the function of a particular sequence can be achieved either directly through sequence similarity searches, or indirectly through the identification of common motifs or domains between groups of functionally related sequences.

Presence of Intergenic region in large number may be due to a higher repeat content in tomato genome than the potato genome. There are many protein families that represent a large gene superfamily in plants, these genes are involved in the biosynthesis of secondary metabolites [11-12].

Alignments between genome sequences of multiple accessions or varieties of a single species allow for the study of genome diversity, evolution and insertion/deletion polymorphisms (InDels). Moreover, alignments between the genomes of related species, for example from the same genus, can be generated to identify structural variation such as translocations, inversions, The identified sequence variation from both approaches can be utilized to study the evolution of genomes, and to generate molecular markers that can be exploited to screen large populations [13-14].

The general availability of genome sequences for crop plant species is having a tremendous impact on the genetics and breeding of these organisms. Future comparative sequence analyses of the completed tomato and potato genome sequences will address many of the unresolved questions related with genome-wide profiles of specific multigene families [15].
Conclusion:
The large scale analysis of tomato and potato revealed many interesting structural and functional differences between two genomes. It was found that tomato genome is more repetitive than the potato genome also the composition of repeat is different in these genomes. Taken together, the present will help in understanding the contents, structure and organization of the tomato and potato genomes, which will be of great value to plant breeders and researchers in the years to come.

Acknowledgement:
The authors are grateful to the Sam Higginbottom Institute of Agriculture, Technology & Sciences, Deemed University, Allahabad for providing the facilities and support to complete the present research work.

References:
[1] Kimura S & Sinha N, CSH Protoc. 2008 doi: 10.1101/pdb.emo105 [PMID: 21356708]
[2] Fray RG & Grierson D, Trends Genet. 1993 9: 438 [PMID: 8122312]
[3] Tomato Genome Consortium, Nature. 2012 485: 635 [PMID: 22660326]
[4] Potato Genome Sequencing Consortium and Xu X et al. Nature. 2011 475: 189 [PMID: 21743474]
[5] Dolezel J et al. Nat Protoc. 2007 2: 2233 [PMID: 17853881]
[6] Frazer KA et al. Nucleic Acids Res. 2004 32: W273 [PMID: 15215394]
[7] Altschul SF et al. Nucleic Acids Res. 1997 25: 3389 [PMID: 9254694]
[8] Singh S et al. Bioinformation 2011 6: 31 [PMID: 21464842]
[9] Finn RD et al. Nucleic Acids Res. 2008 36: D281 [PMID: 18039703]
[10] Gogarten JP & Olendzenski L, Curr Opin Genet Dev. 1999 9: 630 [PMID: 10607614]
[11] Zhu W et al. BMC Genomics. 2008 9: 286 [PMID: 18554403]
[12] Datema E et al. BMC Plant Biol. 2008 8: 34 [PMID: 18405374]
[13] Pujar A et al. Database (Oxford). 2013 doi: 10.1093/database/bat028 [PMID: 23681907]
[14] Väli U et al. BMC Genet. 2008 9: 8 [PMID: 18211670]
[15] Moyle LC, Evolution. 2008 62: 2995 [PMID: 18752600]

License statement: This is an open-access article, which permits unrestricted use, distribution, and reproduction in any medium, for non-commercial purposes, provided the original author and source are credited

Edited by P Kangueane
Citation: Lall et al. Bioinformation 9(18): 923-928 (2013)
Supplementary material:

Table 1: Paralogous proteins predicted in *S. lycopersicum*

| S. No. | Reviewed proteins | Paralogous proteins | S. No. | Reviewed proteins | Paralogous proteins |
|-------|------------------|---------------------|-------|------------------|---------------------|
| 1     | Q43495           | XP_004228451.1      | 31    | Q42876           | NP_001238841.1      |
| 2     | P93207           | NP_001234278.1      | 32    | P58905           | XP_004247774.1      |
| 3     | P93206           | XP_004250139.1      | 33    | O04161           | NP_001234253.1      |
| 4     | P93208           | XP_004252529.1      | 34    | Q9FVNO           | NP_001234216.1      |
| 5     | P93209           | NP_001234677.1      | 35    | P10748           | XP_004229803.1      |
| 6     | P42652           | NP_001234007.1      | 36    | Q42884           | NP_001234422.1      |
| 7     | P93210           | NP_001234107.1      | 37    | Q42885           | NP_001234411.1      |
| 8     | P93211           | NP_001234097.1      | 38    | P37215           | NP_001234415.1      |
| 9     | P93212           | NP_001234637.1      | 39    | P37216           | NP_001234418.1      |
| 10    | P93213           | NP_001234267.1      | 40    | Q9M6A3           | NP_001234320.1      |
| 11    | P93214           | NP_001234272.1      | 41    | Q8WY5            | NP_001233968.1      |
| 12    | P18485           | NP_001234178.1      | 42    | Q08655           | NP_001234632.1      |
| 13    | Q42881           | NP_001234026.1      | 43    | P37219           | XP_004237806.1      |
| 14    | P29535           | NP_001234280.1      | 44    | P37220           | XP_004237805.1      |
| 15    | P93236           | XP_004232029.1      | 45    | Q2MIK2           | YP_514837.1         |
| 16    | Q2MIIH           | YP_514861.1         | 46    | Q2MIK5           | YP_514837.1         |
| 17    | Q2MIIH           | YP_514861.1         | 47    | Q2MI10           | YP_514859.1         |
| 18    | P10967           | XP_004247700.1      | 48    | Q2MI93           | YP_514859.1         |
| 19    | P05116           | NP_001234024.1      | 49    | Q2MI11           | YP_514858.1         |
| 20    | P07920           | XP_004251527.1      | 50    | Q2MI94           | YP_514858.1         |
| 21    | P49297           | NP_001233878.1      | 51    | Q2MIK1           | YP_514838.1         |
| 22    | Q96482           | XP_004236747.1      | 52    | Q2MIK4           | YP_514838.1         |
| 23    | Q96483           | XP_004249866.1      | 53    | Q2MIK0           | YP_514839.1         |
| 24    | Q96484           | XP_004249286.1      | 54    | Q2MIK3           | YP_514839.1         |
| 25    | P28032           | NP_001234099.1      | 55    | Q2MIJ9           | YP_514840.1         |
| 26    | Q40170           | NP_001234543.1      | 56    | Q2MI2B           | YP_514840.1         |
| 27    | Q42464           | NP_001234384.1      | 57    | E7DN63           | NP_001234604.1      |
| 28    | Q9XG19           | NP_001234385.1      | 58    | P48980           | NP_001234465.1      |
| 29    | Q40168           | XP_004232995.1      | 59    | P49118           | NP_001234636.1      |
| 30    | Q10712           | XP_004253490.1      | 60    | Q8GUQ5           | XP_004237477.1      |

Table 2: Paralogous proteins predicted in *S. tuberosum*

| S. No. | Reviewed proteins | Paralogous proteins | S. No. | Reviewed proteins | Paralogous proteins |
|-------|------------------|---------------------|-------|------------------|---------------------|
| 1     | Q43643           | CA0X03822.1         | 56    | P54260           | P54260.1           |
| 2     | Q2VEG8           | AAC23997.1          | 57    | P30924           | P30924.2           |
| 3     | P30170           | P30170.1            | 58    | P55243           | CAW47336.1         |
| 4     | P30171           | P30171.1            | 59    | P23509           | CAW47343.1         |
| 5     | P30173           | P30173.1            | 60    | P50433           | CAW06997.1         |
| 6     | P30167           | P30167.1            | 61    | P393564          | P393564.1          |
| 7     | P30168.2         | P30171.1            | 62    | P48020           | P48020.1           |
| 8     | P14674           | P14674.1            | 63    | Q41437           | BAB20771.1         |
| 9     | P14675           | P14675.1            | 64    | Q41438           | AXE26933.1         |
| 10    | Q42429           | Q42429.1            | 65    | Q08276           | CAW64031.1         |
| 11    | P31427           | P31427.2            | 66    | P50217           | CBN63628.1         |
| 12    | Q41480           | Q41480.2            | 67    | P22200           | P22200.1           |
| 13    | Q43646           | Q43646.1            | 68    | Q9S8M0           | Q9S8M0.2           |
| 14    | P58519           | P58519.1            | 69    | P29696           | P29696.1           |
| 15    | Q41448           | Q41448.1            | 70    | P80471           | P80471.2           |
| 16    | P80595           | P80595.2            | 71    | P37831           | P37831.1           |
| 17    | P21357           | P21357.2            | 72    | O23479           | O23479.1           |
| 18    | Q27655           | Q27655.1            | 73    | Q41238           | Q41238.1           |
| 19    | Q2VEH0           | CA05947.1           | 74    | Q2370            | Q2370.1            |
| 20    | Q2VEI8           | Q2VEI8.1            | 75    | P37225           | P37225.1           |
| 21    | Q819H6           | Q819H6.1            | 76    | P32088           | P32088.2           |
| 22    | P8916            | CAW55661.1          | 77    | Q9FS88           | Q9FS88.2           |
| 23    | Q76MX2           | CB70727.1           | 78    | Q307Y9           | Q307Y9.1           |
Table 3: *S. lycopersicum* and *S. tuberosum* Pfam comparison

| Classification | S. lycopersicum | S. tuberosum |
|----------------|-----------------|--------------|
| Family         | 304             | 194          |
| Domain         | 191             | 113          |
| No_clan        | 788             | 369          |
| CLO            | 590             | 338          |