An integrated multi-level comparison highlights common aspects and specific features between distantly-related species: Tomato and Grapevine

Luca Ambrosino[1], Hamed Bostan[1], Valentino Ruggieri[1], Maria Luisa Chiusano[1]

[1] Department of Agriculture, University of Naples Federico II, Portici (NA), Italy

Corresponding Author:

Maria Luisa Chiusano[1]

Via università 100, 80155 Portici (NA), Italy

Email address: chiusano@unina.it

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2208v1 | CC BY 4.0 Open Access | rec: 2 Jul 2016, publ: 2 Jul 2016

Abstract

Motivation. Even after years from the first completion of genomes by sequencing, comparative genomics still remains a challenge, also enhanced by the availability of numerous draft genomes with still poor annotation quality. The detection of ortholog genes between different species is a key approach for comparative genomics. For example, ortholog gene detection may support investigations on mechanisms that shaped the organization of the genomes, highlighting on gain or loss of function and on gene annotation. On the other hand, the detection of paralog genes is fundamental for understanding the evolutionary mechanisms that drove gene function innovation and support gene families analyses. Here we report on the gene comparison between two distantly related plants, Solanum lycopersicum (Tomato) (The Tomato Genome Consortium 2012) and Vitis vinifera (Grapevine) (Jaillon et al. 2007), considered as economically important species from asterids and rosids clades, respectively. The strategy was accompanied by integration of multilevel analyses, from domain investigations to expression profiling, to get to the most reliable results and to offer powerful resources, in order to understand different useful aspects of plant evolution and physiology and to dissect traits and molecular aspects that could provide novel tools for agriculture applications and biotechnologies.

Methods. In order to predict best putative orthologs and paralogs between Tomato and Grapevine, and to overcome possible annotation issues, all-against-all sequence similarity searches between genes, mRNAs and proteins collections of both species were performed. A Bidirectional Best Hit approach was implemented to detect the best orthologs between the two species. Moreover we developed a dedicated algorithm in Python programming language able to define more extended alignments between mRNA sequences. NetworkX package (Hagberg et al. 2008) was used to define networks of paralogs and orthologs. Proteins domain prediction was carried out on the entire Tomato and Grapevine protein collection by using InterProScan program (Jones et al. 2014). The enzyme classification was obtained by sequence similarity searches between Tomato and Grapevine mRNA collections and the entire UniProt reviewed protein collection (UniProt consortium 2015). The metabolic pathways associated to the detected enzymes were identified exploiting the KEGG Database (Kanehisa and Goto

2000). Expression level of three developmental stages of Tomato (2 cm fruit, breaker and mature red) and the corresponding stages of Grapevine (post-setting, veraison, mature berry) was defined on the basis of the iTAG loci (Shearer et al. 2014) and v1 vitis loci, respectively. The expression was normalized by Reads Per Kilobases per Million (RPKM) for each tissue/stage. The identification of similar expression profiles was performed by the K-means clustering method (Soukas et al. 2000), using the Pearson correlation coefficient as distance metric. For each cluster a subsequent clusterization by the Hierarchical Clustering (HCL) (Eisen et al. 1998) using the Euclidean distance grouped genes also on the basis of expression levels. Both the clustering methods used are those from implemented in the MultiExperiment Viewer (MeV) software.

Results. Although Tomato and Grapevine are phylogenetically distant species, they are both model species for understanding fleshy fruit formation. Comparative analyses, though the available annotations are still preliminary, are essential to understand fruit development. We predicted the presence of a strong core of orthologs genes, exploiting an appropriate approach and overcoming the annotation limits. Networks of ortholog/paralog genes were built between the compared species, offering resources to support studies about the organization and the evolution of gene families in different organisms. By this approach, we detected gene families of one species that underwent an expansion/reduction in the number of their elements when compared to the other species. Species-specific genes of Tomato and Grapevine were also detected. The protein domains common to both species, as the ones exclusively detected in Tomato and Grapevine, and the common and the distinctive enzymatic classes associated to related metabolic pathways, were also predicted for the two compared species supporting structure and functional annotations. Furthermore, the association of RNA-seq data offered an additional information level for comparing gene functionalities from the two species. Thanks to this core collection, we report on similarities and peculiarities between the two genomes.