

PLANT “OMICS”: ON THE IMPORTANCE OF SUITABLE RESOURCES



CHIARA COLANTUONO

Università degli Studi di Napoli 'Federico II'

PhD in Computational Biology and Bioinformatics
(Cycle XXVII)

TUTOR: Doc. Maria Luisa Chiusano
COORDINATOR: Prof. Sergio Coccozza

April 2015

INDEX

Abstract

Introduction

1.1 Omics sciences.....	6
1.2 Genomics	7
1.2.1 <i>Solanum lycopersicum</i>	7
1.2.1.1SGN	9
1.2.2 <i>Solanum tuberosum</i>	9
1.2.2.1SPUD.DB	10
1.3 Other resources	10
1.3.1 Ensembl Plants	10
1.3.2 RefSeq.....	11
1.4 Transcriptomics.....	12
1.4.1 Microarray.....	12
1.4.2 RNA-seq.....	15
1.5 Gene co-expression analysis.....	18
1.6 <i>Arabidopsis thaliana</i>	19
1.6.1 TAIR	20
1.6.2 NASCarrays	21
1.7 Aims and Scope	21

Materials and Methods

2.1 <i>Solanum lycopersicum</i> and <i>Solanum tuberosum</i> genome and annotation data	24
2.2 Tomato SSU and LSU detection.....	25
2.3 Tomato putative split genes.....	25
2.4 Remapping of tomato mRNA on the tomato genome	25

3.6 iTAG annotation versus RefSeq annotation..... 69

3.7 “Guide” to the tomato annotations..... 71

3.8 *Arabidopsis thaliana* microarray resources..... 80

 3.8.1 Consequences of mutant inclusion in microarray datasets..... 89

3.9 Rna-seq analysis in tomato leaves under drought stress 93

Discussion..... 98

References 101

Abstract

In the “-omics” era bioinformatics plays a crucial role in development of new suitable strategies to face different kind of problems attempting to better exploit the different aspects of biology. Moreover, with the upcoming of the Next Generation Sequencing (NGS), the amount of data produced has increased exponentially as the needs of managing the results obtained, with the aim of making these information exploitable for new and deeper analyses. However, all the available resources related to a species are not always unified, updated or integrated, creating confusion and data heterogeneity.

In this context, we focused on the currently available resources for some plant genomes. In particular, we considered *Arabidopsis thaliana*, organism model for plant genomics, and other two species of relevant interest in crop genomics, as well as in the worldwide economy, such as *Solanum lycopersicum* (tomato) and *Solanum tuberosum* (potato). We considered all the relevant genomics resources for these plants, to get the current available information concerning genome releases and gene annotation versions.

Moreover, we went deep into the tomato genome annotations available, highlighting still present limits being the one considered the first gene annotation release for this recently sequenced genome.

In the last part of the work, we extended the analysis also to transcriptomics data. On one hand, we investigated Arabidopsis online resources for co-expression analysis based on microarray approach comparing the source data, the methods and the results currently achievable. On the other hand, due to microarray heterogeneity data for tomato and potato, we preferred to focus on RNA-seq analysis strategies, setting up an appropriate pipeline, tested in a specific analysis on tomato drought stress, and focusing on possible issues arising from a limited annotation as the one from tomato.

Our work highlighted the lack of uniformity between reference plant collections, probably caused by multiple different aspects in a multifaceted world like the one of Plant Sciences. Nevertheless, the lack of reliable and uniform references for Plants can lead to misinterpretation of biological data, limiting their use by the scientific community especially in plant comparative genomics.

Introduction

1.1 Omics Sciences

Omics technologies, such as genomics, transcriptomics, proteomics, were introduced in 1990s, with the Human Genome Project. By combining 'gene' and 'ome' words, Hans Winkler created the term genome, referring to "the haploid chromosome set, which, together with the pertinent protoplasm, specifies the material foundations of the species [...]." (Winkler 1920). Many years after, in 1987, McKusick and Ruddle added 'genomics' to the scientific lexicon as the title of a journal they founded, meaning linear gene mapping, DNA sequencing and comparison of genomes from different species (McKusick and Ruddle 1987). The omics technologies lead at copious amounts of data at multiple levels, i.e. from gene sequence and expression to protein and metabolite patterns, underlying variability in cellular networks and function of whole organ systems (Nicholson and Lindon 2008, Wilke et al. 2008).

The aims of the omics science is to reach a complete overview of all the molecules contributing to the functionality of an organism. For example, genomics is the science that defined the complete set of genomic elements inside a cell. However, the determination of the genomic sequence is only the starting point of genomics. Therefore, the genomic sequences are used to study the function of the numerous genes (functional genomics), to compare genome in one organism with another one (comparative genomics), to collect genetic material recovered directly from environmental samples (metagenomics) and to study the complete set of epigenetic modifications (epigenomics).

All the data generated from the omics science have to be integrated and interpreted by complex mathematical and computational models. This effort is called System Biology. In the omics and system biology era, bioinformatics plays a crucial role in development of new suitable strategies to face different kind of problems attempting to better exploiting the different aspects of biology. Moreover, with the upcoming of the Next Generation Sequencing (NGS), the amount of data produced has been increased exponentially as the needs of managing the results obtained, with the aim of making exploitable these information for new and deeper analyses and, especially, available for all the scientific community.

1.2 Genomics

1.2.1 *Solanum lycopersicum*

Solanum lycopersicum (tomato) is one of the most important crop in the world and it is considered a model for the fruit development. Tomato belongs to the Solanaceae family and its genome consist of a 12 chromosomes, in a haploid set, with a total of 950 Mb (Mueller et al. 2009). The complete sequence of the tomato genome was released in 2012 by The Tomato Genome Consortium, in which Italy was involved (Tomato Genome Consortium 2012). At the beginning of the project, the tomato genome was sequenced with a BAC-by-BAC approach. However, with the incoming of NGS, in 2008 a whole genome shotgun (WGS) was applied.

The tomato chromosomes consist of an extended heterochromatic region (77% genome), mostly representing the telomeres and extended pericentromeric regions. The euchromatic regions locates in the distal part of the chromosome (Peterson et al. 1996, Peterson et al. 1998), composed of most single copy sequences with only few retrotransposon (Chang et al. 2008) and the 90% of the genes.

The pericentromeric heterochromatic segments were 1.23 times wider than euchromatic segments. They contain a large portion of retrotransposons, repeated sequences and some single-copy sequences, which also include a lower but representative gene content (Di Filippo et al. 2012). Pericentromeric heterochromatin is generally assumed to be gene poor and repeat-rich, where crossing over is severely repressed (Sherman and Stack 1995) (Fig. 1).

The international Tomato Annotation Group (iTAG) carried out the annotation of the tomato genome, releasing different versions. The most recent ones are the version 2.3, based on the genome assembly SL2.40 (Tomato Genome Consortium 2012), and the version 2.4, based on the genome assembly SL2.50 (Shearer et al. 2014).

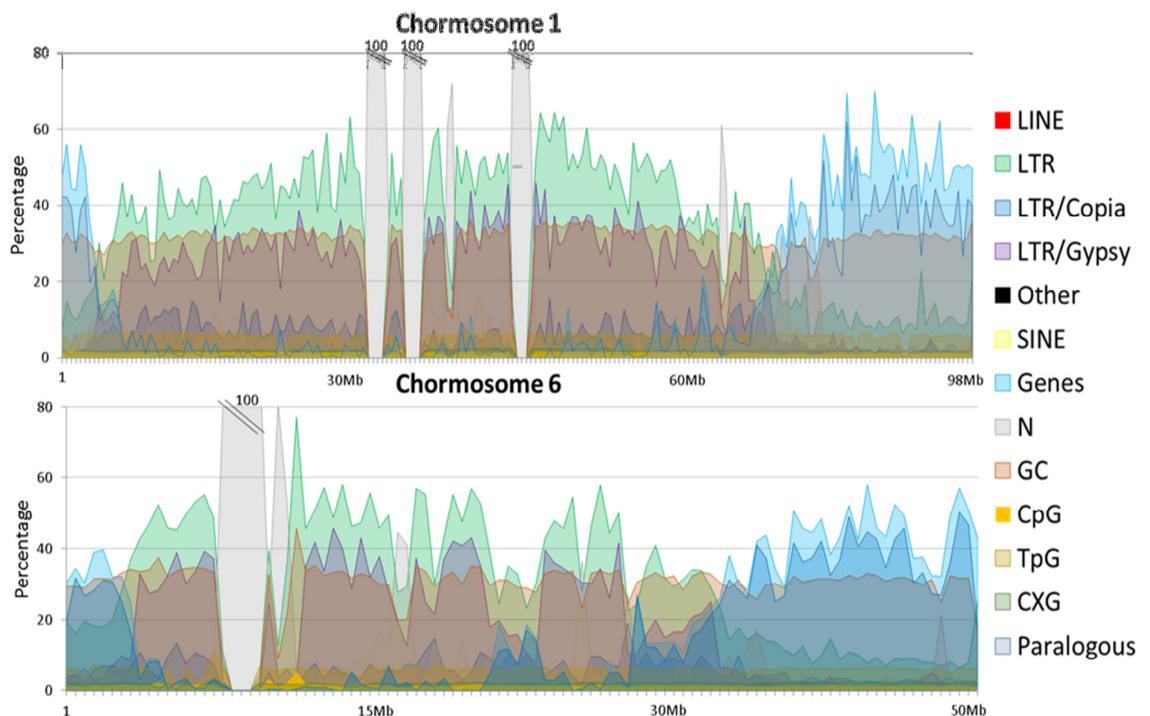


Figure 1 Percentage of genes, paralogues, repeated regions, N, GC, CpG, TpG, CXG on chromosome 1 and 6 is reported

1.2.1.1 SGN

Tomato is considered a model for all the Solanaceae and other species for its fruit development. Many data are available for it and they can be exploited in several online resources.

The reference website for tomato is Sol Genomic Network (SGN) (Bombarely et al. 2011), available at <http://www.sgn.cornell.edu/>. The platform includes all the genomic information about tomato, such as genome assembly versions and annotation versions, downloadable from the FTP page offered by the website. SGN not only includes tomato genomic data, but also genetic, transcriptomic, phenotypic and taxonomic information with the data of other Solanaceae (potato, eggplant, pepper and petunia).

1.2.2 *Solanum tuberosum*

Solanum tuberosum (potato) is the most important crop in the world, after wheat, rice and maize and it belongs to the Solanaceae family. Potato genome was the first Solanaceae genome to be sequenced in 2011 (Potato Genome Sequencing Consortium 2011) and it is the first asterid genome, representing a major clade of eudicots. As almost all the Solanaceae family members, potato have 12 chromosomes (Wikstrom et al. 2001) and its genome size is about 844 Mb. The Potato Genome Sequence Consortium (PGSC) carried out the sequencing of two varieties: RH89-039-16 (RH), a diploid, heterozygous potato variety, and DM1-3 516R44 (DM), a doubled monoploid. The PGSC originally started with the sequencing of RH variety. The project built a diploid potato genomic Bacterial Artificial Chromosome (BAC) clone library of 78,000 clones. In addition, the BAC-ends were sequenced and publicly available. From the genetic-physical map, between 50 to 150 seed BACs were identified for each chromosome and fluorescent in situ hybridization (FISH) experiments on selected BAC clones confirmed these anchor points.

The seed clones provided the starting point for a BAC-by-BAC sequencing strategy.

The sequencing of the DM variety was started because the overall progress in the sequencing of RH one was slow. The heterozygosity of RH limited the progress of physical mapping and made the assembly of the genome sequence difficult. Therefore, the sequencing of DM variety done by whole genome shotgun (WGS).

The genome released in 2011 was at scaffolds level, and only one year after, in 2012, the 12 potato pseudomolecules were available (potato genome assembly version 2.1.10). In 2013 the last genome assembly version based on pseudomolecules was released (Sharma et al. 2013) and a new annotation was available.

1.2.2.1 SPUD.DB

SPUD.DB (Hirsch et al. 2014), available at <http://potato.plantbiology.msu.edu/>, is the reference portal through which it is possible to exploit and to obtain the potato genome and annotation. In fact, in the PGSC download page it is possible to download the fasta files of all the genome assemblies released since 2011. Moreover, GFF3 file of all the annotation version were available. The website allowed the exploitation of the last versions of the genome through a Genome Browser and a query page.

1.3 Other resource

1.3.1 Ensembl Plants

Ensembl Plants (<http://plants.ensembl.org/index.html>) is a section of Ensembl Genomes (Cunningham et al. 2015), developed by EBI

(<http://www.ebi.ac.uk/>) and the Wellcome Trust Sanger Institute (<http://www.sanger.ac.uk/>). It is a huge platform that includes annotation, analysis and display of more than 30 plant genomes.

Even though the genomic information included in the platform are obtained from the official resources, Ensembl creates automated annotation, in some cases also curated manually and applies an automatic gene annotation system, called Genebuild. Genebuilds are performed on high-coverage genomes and the initial set-up involves loading the assembly into an Ensembl databases and then running several analyses across the genome such as repeats masking and ab initio gene predictions. This stage is followed by the similarity stage, in which proteins from closely related species are used to build transcript structure in regions. The next stage in the genebuild is to align species-specific cDNA, EST and, when available, RNA-seq to the genome. The final set of gene predictions is obtained by merging identical transcripts built from different proteins sequences to produce multi-transcript gene predictions, each with a non-redundant set of transcripts models.

Ensembl Plants offers several tools to exploit species data, in particular BioMart

(<http://plants.ensembl.org/biomart/martview/5498a38fd3756d7bdea694466adc5357>) is a powerful platform that allows to download all the information available for a given species, such as annotation, orthologous, GO, in a GFT format.

1.3.2 RefSeq

The NCBI Reference Sequence (RefSeq) (Pruitt et al. 2007) is a dedicated database of non-redundant set of reference standards derived from the International Nucleotide Sequence Database Collaboration databases that includes chromosomes, complete genomic molecules (organelle genomes,

viruses, plasmids), intermediate assembled genomic contigs, curated genomic regions, mRNAs, RNAs, and proteins, and it is part of the NCBI environment (<http://www.ncbi.nlm.nih.gov/>). Each RefSeq record represents a synthesis of the information generated and submitted by others. This collection is an integration of different data types, i.e. sequence, genetic, expression, and functional information, with a uniform set of conventions and standards. The RefSeq collection supports the following activities:

- genome annotation;
- gene characterization;
- comparative genomics;
- reporting sequence variation;
- expression studies.

The pipeline used for gene prediction is in principal based on three complementary approaches: 1) known genes are placed primarily by aligning mRNAs to the assembled genomic contigs; 2) additional genes are located based on alignment of ESTs to the assembled genomic contigs; 3) previously unknown genes are predicted using hints provided by protein homologies. Whenever possible, predicted genes are identified by homology between the protein they encode and other known protein sequences.

The records included in RefSeq database can be queried in all the tools offered by NCBI and can be download in a GeneBank format.

1.4 Transcriptomics

1.4.1 Microarray

Nowadays, microarray technology still remains one of the less expensive and powerful approach to study the transcriptome, i.e. the transcriptional activity,

of a biological sample, whether it is represented by a tissue, cells, or a mixture, in specific conditions, such as physiological, stress or pathological ones (Slonim and Yanai 2009). Since the capability of providing a consistent snapshot of the expression of many different genes, though with some well-known technical limits (Hoheisel 2006), microarrays are still a relevant technology despite the incoming of other techniques. They are widely used in many aspects, such as

- Expression analysis;
- Mutation analysis;
- Comparative genomics analysis;
- Gene discovery
- Disease diagnosis.

In particular, their employment in expression analysis not only permits to detect patterns of high or low expressed genes from comparative experiments, but also enable to describe expression patterns for different tissues/conditions, or in time course experiments. Indeed, the variability of the expression of a multitude of genes from a genome can be traced by this technology.

A typical microarray experiment involves the hybridization of an mRNA molecule to the DNA template from which it is originated. Many DNA samples are used to construct an array. The amount of mRNA bound to each site on the array indicates the expression level of the various genes. This number may run in thousands. All the data is collected and a profile is generated for gene expression in the cell. An array is an orderly arrangement of samples where matching of known and unknown DNA samples is done based on base pairing rules. An array experiment makes use of common assay systems such as microplates or standard blotting membranes. The

sample spot sizes are typically less than 200 microns in diameter usually contain thousands of spots.

Thousands of spotted samples known as probes (with known identity) are immobilized on a solid support. The spots can be DNA, cDNA, or oligonucleotides. These are used to determine complementary binding of the unknown sequences thus allowing parallel analysis for gene expression and gene discovery. An experiment with a single DNA chip can provide information on thousands of genes simultaneously. An orderly arrangement of the probes on the support is important as the location of each spot on the array is used for the identification of a gene.

One of the most exploited microarray chip is from Affymetrix [<http://www.affymetrix.com/>]. It consists of a number of probe cells that contain a unique probe. These are tiled in probe pairs as a Perfect Match (PM) and a Mismatch (MM). PM and MM have the same sequence, except for a change in the middle of the MM, avoiding the perfect match with the target sequence (Fig. 2). MM are included since they are supposed to control for variation in chemical composition and abundance of cross-hybridizing fragments from other genes. By combining PM and MM information from many probes, gene to gene differences should be minimized.

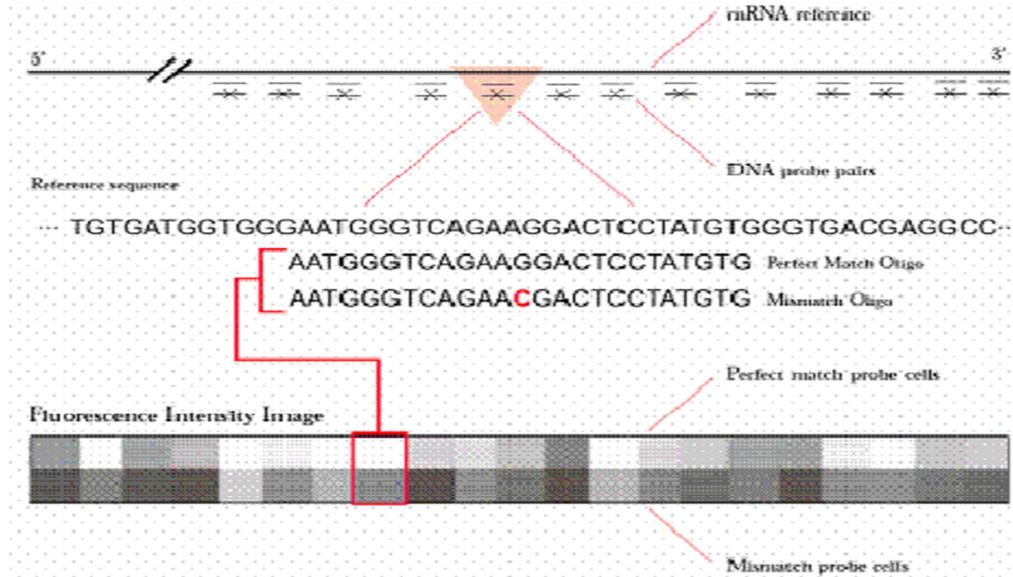


Figure 2 Example of Perfect Match (PM) and Mismatch (MM) sequences. Differences in fluorescence intensity per probe are also shown

1.4.2 RNA-seq

Using Next-Generation Sequencing (NGS) techniques, RNA-seq can reveal the identity of most of the RNA species inside a cell, making a snapshot of their content in a given moment (Chu and Corey 2012).

In principal, a population of RNA (such as mRNA) is converted to a library of cDNA, than the sequences are fragmented and an adaptor is attached, to one or both ends. Each molecule is then sequenced by a high-throughput approach to obtain short sequences from one end (single-end sequencing) or both ends (pair-end sequencing). Pair end reads can, moreover, be overlapping each other, making their assembly easier. The reads are typically 30–400 base pairs, depending on the DNA sequencing technology used (Wang et al. 2009) (Fig 3). The technologies that nowadays allow to perform RNA-seq analysis are Illumina (<http://www.illumina.com/>), Roche 454 (<http://www.454.com/>), Ion Torrent

(<http://www.lifetechnologies.com/it/en/home/brands/ion-torrent.html>), SOLiD (<http://www.lifetechnologies.com/it/en/home/life-science/sequencing/next-generation-sequencing/solid-next-generation-sequencing.html>) and PacBio (<http://www.pacificbiosciences.com/>). They differ in the way of sequencing DNA but also in reads length, coverage and quality.

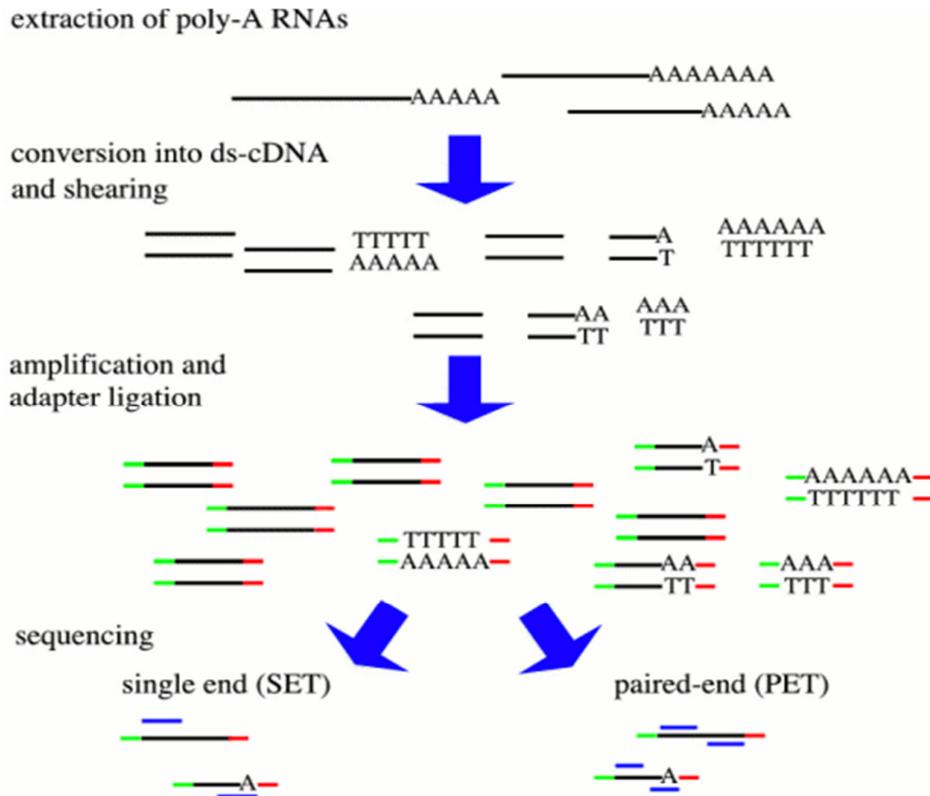


Figure 3 Schematic view of the steps that lead to RNA-seq reads

For example, in Illumina technology, after the cDNA fragmentation, both ends of the double strand are ligated to adaptors. Therefore, single strand sequences are introduced into flowcells, where the complementary sequences of the adaptors are present, allowing the hybridization. The anchored fragments then bend toward the surface and hybridized to a second complementary sequence which contains a primer that allowed DNA polymerase to replicate the fragment. The double-stranded DNA is then denatured, leaving two

complementary fragments attached to the flowcell. This process of hybridization, DNA synthesis and denaturation, is repeated many times to create a cluster of fragments. In the end, complementary fragments are removed and fluorescently-labeled, reversibly terminated nucleotides were added together with primers and DNA polymerase, beginning the read sequencing (Fig. 4).

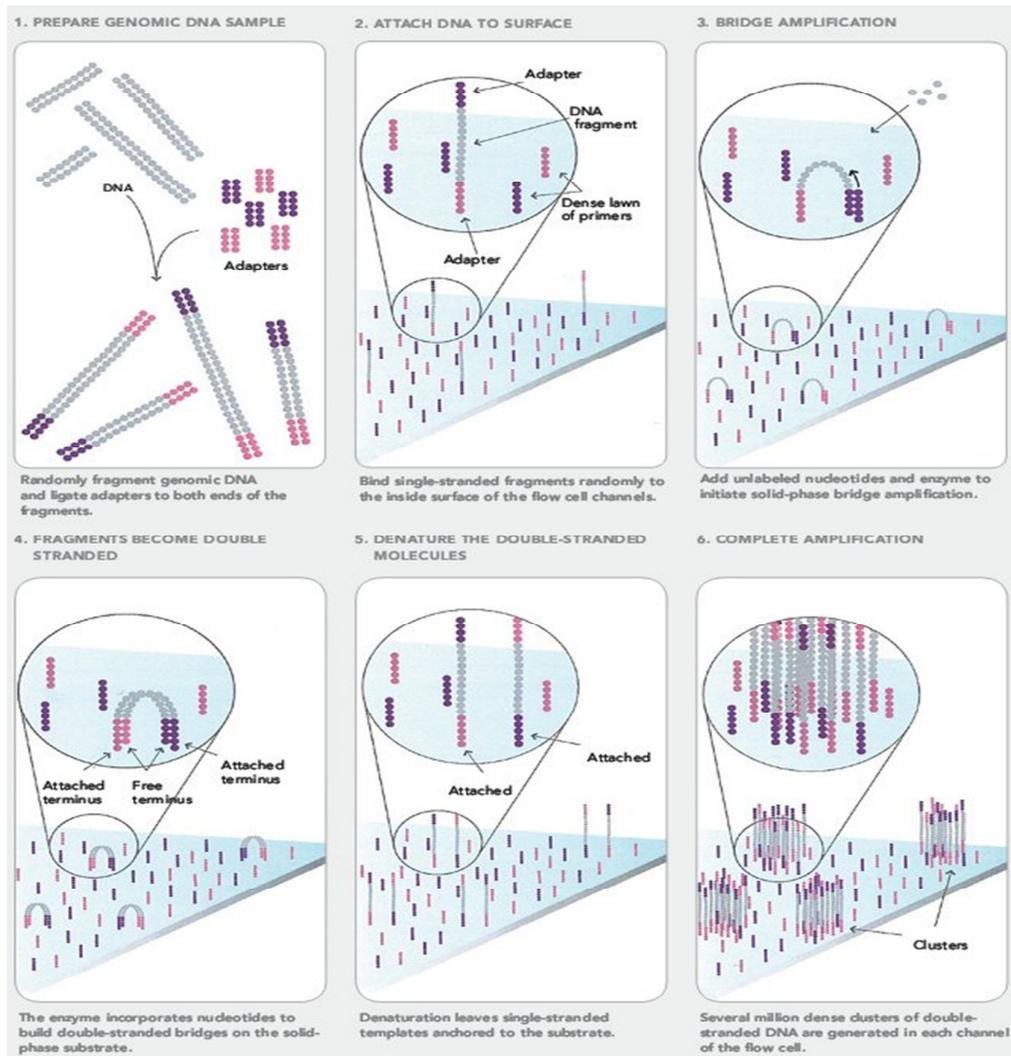


Figure 4 Steps of Illumina technology leading to RNA-seq reads (<http://www.illumina.com/>)

The large spreading of RNA-seq technologies is related to the fact that in this kind of analysis the knowledge of the sequence analyzed is not mandatory, as in a microarray analysis. Overcoming this limit, the applications of this technique are several, such as gene expression (Weber et al. 2007, Sugarbaker et al. 2008, Torres et al. 2008), gene annotation, the investigation of genetic variation (Korbel et al. 2007) and DNA methylation (Cokus et al. 2008).

However, the amount of data that can be generated from a RNA-seq analysis can be only processed by suitable bioinformatics pipelines. For example, a typical employment of RNA-seq is to identify genes differentially expressed genes (DEGs) in a certain conditions, as an example: physiological and stress conditions. In this case, the identification of is made by several steps. The first one is the mapping of the reads on the reference genome, where presents, or their de novo assembly, where the genome was not available. After the mapping, it is necessary to count the reads number inside the gene (or exons) area. Only after these steps, DEGs can be called through several bioinformatics tools, such as DEseq (Anders and Huber 2010) or edgeR (Robinson et al. 2010), that are R packages, or Cufflinks (Trapnell et al. 2010), that works on UNIX system.

1.5 Gene co-expression

The amount of data product by the techniques cited above is an immense amount of biological information that can be used to obtain genes expression profiles. In fact, the analysis of those profiles, derived from a sufficient number of experiments that support a statistically significant results, can support the detection of co-expressed genes from a species, i.e. genes with positively correlated profiles. As defined by the Guilt by Association (GbA) principle, genes sharing the same expression patterns in several experiments

may be studied as candidates involved in the same functional gene network (Quackenbush 2003). Co-expressed genes in general are showed as networks (GCN). The GCN are undirected graph in which each node represents a gene and the edge represent the relationship between them. To evaluate if the relationship between genes, several methods can be applied, such as Pearson correlation coefficient or Spearman's rank correlation coefficient.

Co-expression analysis is a powerful tool that give the possibility to may establish many functionally related genes and reveal about genes regulatory systems (Eisen et al. 1998, Spellman et al. 1998).

Data for co-expression analyses can be obtained with transcriptomics approaches, such as microarrays. However, even though the same approach is used, the comparison among different microarray dataset is not always possible, also after normalization methods. In fact, it is necessary the use of the same technology since probes specificity can be affected by the different way of sample preparation, influencing the measurements (Kuo et al. 2002).

1.6 Arabidopsis thaliana

Arabidopsis thaliana, is a small annual or biennial plant belonging to the Brassicaceae family. It is diploid, it have 5 chromosome, in its haploid form and it was the first plant species whose genome was completely sequenced in 2000 (Arabidopsis Genome Initiative 2000).

Arabidopsis sequence genome was the third one released after the one of *Caenorhabditis elegans* (C. elegans Sequencing Consortium 1998) and *Drosophila melanogaster* (Adams et al. 2000) giving for the first time full access to the genome structure and organization of a vegetal organism (Bevan et al. 2001).

This plant was studied for a long time because its peculiarity, such as a small diploid genome, only 125 Mb, and its cultivation properties: small size, short life cycle and the high seeds production through self-pollination. All these attributes have led to consider this plant as model organism for plants (Koornneef and Meinke 2010). However, beyond these positive aspects, its genome has showed an unexpected complexity: probably, three rounds of whole genome duplications (α , β and γ , where α is the most recent one) have occurred during its evolution, followed by a loss of genomic content (Blanc et al. 2003, Bowers et al. 2003, Tang et al. 2008). All these genomic reshuffling have led to a lacking of conserved gene order that made difficult the exploitation of this species for studies of comparative analyses among species and moreover, the lacking of an exhaustive annotation underlines how *Arabidopsis thaliana* is still far to be the perfect model organism.

1.6.1 TAIR

Being nowadays one of the most studied species, many resources are available for this *Arabidopsis thaliana*. In particular, The Arabidopsis Information Resource (TAIR) (Rhee 2003), is the reference website of all the genomic data related to this plant. Browsing the platform, it is possible to exploit Arabidopsis gene function, expression patterns, genome assembly and annotation data. In this latter are present all the information about the Arabidopsis genes, such as their positions on the chromosomes and their predicted functions. Several genome releases were published in the last years, and the most recent one is version 10.

1.6.2 NASCArrays

In this work, particular attention was dedicated to Arabidopsis co-expression analysis using a microarray approach, and NASCArrays database is the reference site for all the public Affymetrix ATH1 and AG ‘GeneChip’ microarrays for *A. thaliana* (Craigon et al. 2004). The platform collects 706 experiments and 5364 slides. All data are described following the MIAME guidelines (Brazma et al. 2001) and the description includes the sample information, hybridization, normalization and scanning protocol exploited (generally based on the MAS5.0 protocol (Pepper et al. 2007)). For each gene in a slide, the expression is defined by the Signal, Stat Pairs Used, Present Call and Detection P-value, and generally the original probe measures of the CEL files are available too. Nascarrays allows the user to search for single microarray experiment. Data mining tools are also offered:

- the spot history shows the expression profile of a gene over all the available experiments;
- the two gene scatter plot shows a scatter plot of the gene specific expression values through all the experiments;
- the gene swinger tool shows the experiments which show a consistent change of the expression value of a desired gene when compared to the overall expression values of that gene in other experiments;
- the bulk gene download enables the user to download all the expression profiles of a gene (or all the genes) over all the experiments included in the database.

1.7 Aims and Scope

All the amount of data produced from different kind of technologies in the omics sciences, are often hosted in dedicated platforms that allow users to exploit them. This platforms are precious in the research work, but not always updated or integrated with other available resources. Sample homogeneity should be a fundamental requirement also to support comparable analyses worldwide. Often, because of fast technological evolution and the lack of unified experimental strategies, homogeneous data collections from different species, tissues, conditions and unified and coherent platforms are not always available. As a result, consistent collections comes from heterogeneous samples, i.e. from the same species, but not necessarily from the same genotype, and similar and comparable technologies.

In the laboratory of Dr. Chiusano where I carried out my PhD thesis work, we focused mainly on plant genomics, specifically on *Solanum lycopersicum* (tomato). Starting from this species, we expanded our investigations on genome resources considering *Arabidopsis thaliana*, a model plant species, and *Solanum tuberosum* (potato), another recently sequenced Solanaceae species. We highlighted the heterogeneity of the resource available for potato and we put particular attention to the problems of the tomato genome annotation.

We then moved to perform gene co-expression analysis and validate possible methodologies in plants. We investigated microarray resources for the plant species considered and we got to the point that exhaustive collections for this approach were only available for *A thaliana*. Interestingly, we faced the multitude of resources for gene co-expression dedicated to *A. thaliana*, and we investigated on the possible advantages/disadvantages of these multiplicity. For tomato and potato, the data available were from

heterogeneous collections to be compared and provide a consistent collection for gene co-expression. We also considered the expansion of RNA-seq based collections for plants. Therefore, to get inside this novel technologies, I set up a pipeline for the management and the analysis of these type of data. I tested it in the analysis of differentially expressed genes (DEGs) in tomato drought stress and I also compared the way the results could be affected by an appropriate gene annotation.

Materials and Methods

2.1 *Solanum lycopersicum* and *Solanum tuberosum* genome and annotation data

The tomato genome sequences version 2.40 and 2.50 were downloaded in fasta format from the FTP section of SGN (<http://solgenomics.net/>) dedicated to *Solanum lycopersicum* data, as well as the GFF3 annotation files versions 2.3 and 2.4.

RefSeq tomato annotation was retrieved from the NCBI website (<http://www.ncbi.nlm.nih.gov/>), querying for “*Solanum lycopersicum* gene” in the Gene database and selecting only for RefSeq sequences. The annotation of the genes was downloaded in GenBank format through the “sent to” option offered by the website. A GFF3 was eventually obtained from the GenBank format with a suitable Perl script.

Solanum tuberosum GFF annotation files were downloaded from different resources. In the SpudDB website, in the page dedicated to the download of PGSC (Potato Genome Sequencing Consortium) data (http://potato.plantbiology.msu.edu/pgsc_download.shtml), the GFF files obtained were:

- PGSC_DM_V403_genes.gff;
- PGSC_DM_v3_2.1.11_pseudomolecule_annotation.gff;
- PGSC_DM_v3_2.1.10_pseudomolecule_annotation.gff;
- PGSC_DM_v3.4_gene.gff.

Another potato GTF annotation was downloaded from Ensembl Plants, in the Biomart section (<http://plants.ensembl.org/>), obtaining GeneID, TranscriptID, Start, End and Strand information.

2.2 Tomato SSU and LSU detection

Small and Large subunit (SSU and LSU) of rRNAs were predicted in tomato using a BLASTn (Camacho et al. 2009), aligning the 12 chromosomes, plus chromosome 0, of the tomato assembly SL2.50 versus SSU (RF01960) and LSU (RF02543) databases, independently. The two datasets of repeated sequences were downloaded from RFAM release 12.0 (Griffiths-Jones et al. 2003). From the results of the alignment, only sequences that were $\geq 98\%$ of coverage were taken in consideration.

2.3 Tomato putative split genes

In order to verify if there were missannotated genes into the tomato genome, a BLASTx (Camacho et al. 2009) analysis was performed aligning the tomato mRNA (iTAG vers. 2.3) versus the UNIPROT reviewed database ver. 2013_06 (UniProt Consortium 2015), with an e-value cut-off of 10^{-3} . From the BLASTx result, all the mRNA codified by genes annotated in close positions on the genome (with a maximum of 6 genes between them) and matching the same proteins, were extracted. Afterwards, genes that matched the same protein but in different consecutive positions were called as *split genes*.

2.4 Remapping of tomato mRNA on the tomato genome

The remapping of the tomato mRNAs (iTAG vers. 2.3) on the tomato genome (SL2.40) was performed by GenomeThreader (Gremme et al. 2005), using the “*cdna*” option and setting a cut-off of 0.80 of coverage and 0.90 of identity.

The results were processed by a home-made pipeline set up in the Dr. Chiusano laboratory: for all the mRNA queried it was assigned a suitable flag in order to clarify their behaviour (see Results).

2.5 *Arabidopsis thaliana* microarray analysis

2.5.2 Microarray dataset

We have downloaded the gene expression values of 79 experiments with samples taken from several tissues, in physiological conditions and repeated in triplicate, for a total number of 237 microarray slides, from The “Developmental Series Expression atlas of *Arabidopsis* development” subfolder (<http://affymetrix.arabidopsis.info/narrays/experimentbrowse.pl>) (Tab. 1). Each slide was based on the ATH1 Affymetrix chipset, able to detect 22810 probes and normalized through MAS 5.0 protocol. Only 21769 probes had signals and from these latter we have removed the following probes: 387 known as multiple genes matching, 53 no gene matching (transposon, miRNA, others), 107 similar to unrelated sequences (x_at probes) (Redman et al. 2004), 27 shared probes (s_at), 3 “sequence family” probes (f_at), 1 “rules dropped” probes. Moreover, the expression signal of 224 genes was defined by more than one probe (totally 458 redundant probes), so we took the average of these ones. The final step was to filter out all the probes with an expression level under the 5th percentile in each sample, in all the experiments, bringing the final number of gene specific probes exploited in this work to 20908. The average signal of each probe in each experiment was calculated in Excel taking the average of the three replicates. A log₂ transformation on all the signals has been applied: in this way only genes with a huge difference in the signal will be treated and considered as differently expressed.

Table 1 Dataset used for the analyses. In grey there are the 16 experiments with mutants, in white 63 experimnts without mutants

Sample ID	Genotype	Tissue	Age	Photoperiod	Substrate
ATGE_11	gl1-T	rosette leaf #4, 1 cm long	21+ days	continous light	Soil
ATGE_18	gl1-T	rosette leaf #12	21+ days	continous light	Soil
ATGE_46	clv3-7	shoot apex, inflorescence (after bolting)	21+ days	continous light	Soil
ATGE_47	lfy-12	shoot apex, inflorescence (after bolting)	21+ days	continous light	Soil
ATGE_48	ap1-15	shoot apex, inflorescence (after bolting)	21+ days	continous light	Soil
ATGE_49	ap2-6	shoot apex, inflorescence (after bolting)	21+ days	continous light	Soil
ATGE_50	ap3-6	shoot apex, inflorescence (after bolting)	21+ days	continous light	Soil
ATGE_51	ag-12	shoot apex, inflorescence (after bolting)	21+ days	continous light	Soil
ATGE_52	ufo-1	shoot apex, inflorescence (after bolting)	21+ days	continous light	Soil
ATGE_53	clv3-7	flower, stage 12; multi-carpel gynoecium; enlarged meristem; increased organ number	21+ days	continous light	Soil
ATGE_54	lfy-12	flower, stage 12: shoot characteristics; most organs leaf-like	21+ days	continous light	Soil
ATGE_55	ap1-15	flower, stage 12: sepals replaced by leaf-like organs, petals mostly lacking, 2° flowers	21+ days	continous light	Soil

Materials and Methods

ATGE_56	ap2-6	flower, stage 12: no sepals or petals	21+ days	continous light	Soil
ATGE_57	ap3-6	flower, stage 12: no petals or stamens	21+ days	continous light	Soil
ATGE_58	ag-12	flower, stage 12: no stamens or carpels	21+ days	continous light	Soil
ATGE_59	ufo-1	flower, stage 12; filamentous organs in whorls two and three		continous light	Soil
ATGE_1	wild type	cotyledon	21+ days	continous light	Soil
ATGE_2	wild type	hypocotyl	21+ days	continous light	Soil
ATGE_3	wild type	Root	21+ days	continous light	Soil
ATGE_4	wild type	shoot apex, vegetative + young leaves	21+ days	continous light	Soil
ATGE_5	wild type	leaves 1+2	21+ days	continous light	Soil
ATGE_6	wild type	shoot apex, vegetative	21+ days	continous light	Soil
ATGE_7	wild type	seedling, green parts	21+ days	continous light	Soil
ATGE_8	wild type	shoot apex, transition (before bolting)	21+ days	continous light	Soil
ATGE_9	wild type	roots	21+ days	continous light	Soil
ATGE_10	wild type	rosette leaf #4, 1 cm long	21+ days	continous light	Soil
ATGE_12	wild type	rosette leaf #2	21+ days	continous light	Soil
ATGE_13	wild type	rosette leaf #4	21+ days	continous light	Soil
ATGE_14	wild type	rosette leaf #6	21+ days	continous light	Soil
ATGE_15	wild type	rosette leaf #8	21+ days	continous light	Soil
ATGE_16	wild type	rosette leaf #10	21+ days	continous light	Soil
ATGE_17	wild type	rosette leaf #12	21+ days	continous light	Soil
ATGE_19	wild type	leaf 7, petiole	21+ days	continous light	Soil
ATGE_20	wild type	leaf 7, proximal half	21+ days	continous light	Soil
ATGE_21	wild type	leaf 7, distal half	21+ days	continous light	Soil
ATGE_22	wild type	develompental drift, entire rosette after transition to flowering, but before bolting	21+ days	continous light	Soil
ATGE_23	wild type	as above	21+ days	continous light	Soil
ATGE_24	wild type	as above	21+ days	continous light	soil

Materials and Methods

ATGE_25	wild type	senescing leaf	21+ days	continous light	soil
ATGE_26	wild type	cauline leaf	21+ days	continous light	soil
ATGE_27	wild type	stem, 2nd internode	21+ days	continous light	soil
ATGE_28	wild type	stem, 1st node	21+ days	continous light	soil
ATGE_29	wild type	shoot apex, inflorescence (after bolting)	21+ days	continous light	soil
ATGE_31	wild type	flower, stage 9	21+ days	continous light	soil
ATGE_32	wild type	flower, stage 10/11	21+ days	continous light	soil
ATGE_33	wild type	flower, stage 12	21+ days	continous light	soil
ATGE_34	wild type	flower, stage 12, sepals	21+ days	continous light	soil
ATGE_35	wild type	flower, stage 12, petals	21+ days	continous light	soil
ATGE_36	wild type	flower, stage 12, stamens	21+ days	continous light	soil
ATGE_37	wild type	flower, stage 12, carpels	21+ days	continous light	soil
ATGE_39	wild type	flower, stage 15	21+ days	continous light	soil
ATGE_40	wild type	flower, stage 15, pedicels	21+ days	continous light	soil
ATGE_41	wild type	flower, stage, 15, sepals	21+ days	continous light	soil
ATGE_42	wild type	flower, stage, 15, petals	21+ days	continous light	soil
ATGE_43	wild type	flower, stage, 15, stamen	21+ days	continous light	soil
ATGE_45	wild type	flower, stage, 15, carpels	21+ days	continous light	soil
ATGE_73	wild type	mature pollen	6wk	long day(16/8)	soil
ATGE_76	wild type	siliqua, with seeds stage 3; mid globular to early heart embryo	8wk	long day(16/8)	soil
ATGE_77	wild type	siliqua, with seeds stage 4;early to late heart embryo	8wk	long day(16/8)	soil
ATGE_78	wild type	siliqua, with seeds stage 5	8wk	long day(16/8)	soil
ATGE_79	wild type	seed, stage 6; mid to late torpedo embryos	8wk	long day(16/8)	soil
ATGE_81	wild type	seed, stage 7; late torpedo to early walking-stick embryo	8wk	long day(16/8)	soil

ATGE_82	wild type	seed, stage 8; walking-stick to early curled-cotyledons embryo	8wk	long day(16/8)	soil
ATGE_83	wild type	seed, stage 9; curled-cotyledons to early green-cotyledons embryo	8wk	long day(16/8)	soil
ATGE_84	wild type	seed, stage 10; green cotyledons embryo	8wk	short day (10/14)	soil
ATGE_87	wild type	vegetative rosette	7 days	short day (10/14)	soil
ATGE_89	wild type	vegetative rosette	14 days	short day (10/14)	soil
ATGE_90	wild type	vegetative rosette	21 days	short day (10/14)	soil
ATGE_91	wild type	Leaf	15 days	long day (16/8)	soil
ATGE_92	wild type	flower	28 days	long day (16/8)	soil
ATGE_93	wild type	Root	15 days	long day (16/8)	soil
ATGE_94	wild type	Root	8 days	continuos light	soil
ATGE_95	wild type	Root	8 days	continuos light	soil
ATGE_96	wild type	seedling, green parts	8 dyas	continuos light	soil
ATGE_97	wild type	seedling, green parts	8 days	continuos light	soil
ATGE_98	wild type	Root	21 days	continuos light	soil
ATGE_99	wild type	Root	21 days	continuos light	1x MS agar, 1% sucrose
ATGE_100	wild type	seedling, green parts	21 days	continuos light	soil
ATGE_101	wild type	seedling, green parts	21 days	continuos light	1x MS agar, 1% sucrose

2.5.2 Mutant inclusion/exclusion from the dataset

To evaluate the possible effect of mutant inclusion/exclusion, we purposely selected two pools of genes as case examples, including genes most and least affected by co-expression instability (i.e. variation in co-expression due to the samples included in the dataset), respectively. For each gene pool, 200 genes

were tested for pair wise co-expression (39800 gene pairs) as measured by Pearson's correlation coefficient (r) and associated p-value, based on the two datasets of samples described, either with (mut+) or without (mut-) mutants (79 and 63 samples, respectively, Tab. 1). In a preliminary analysis on both stable and unstable gene pools, we extensively tested the existence of a relationship between the frequency of gene co-expression and presence/absence of mutants in the dataset. We used the Chi-square test for independence on 2x2 contingency tables reporting, for each gene and for all data pooled, the observed occurrences of either co-expressed ($r \geq 0.7$ or $r \leq -0.7$) or not co-expressed ($-0.7 < r < 0.7$) gene pairs, for either mut+ or mut-datasets (398 pairwise comparisons for each gene). A significant Chi-square statistic indicated the dependence of the observed co-expression patterns from the inclusion or exclusion of mutants in the reference dataset.

Then, for each gene pair, we assessed the effect of mutants on gene co-expression by testing the significance of the difference in correlations with or without mutants. Occurrences of significant ($P < 0.05$) and not significant differences in correlations were calculated both separately for each tested gene and for all genes pooled. In the case of significant correlation differences (i.e. gene pairs with co-expression significantly affected by mutant inclusion or exclusion), the type, the occurrence and the significance of the effect was assessed. Effect types were defined on the base of the possible values of r_{mut+} and r_{mut-} ("+", positive and statistically significant at $P < 0.05$; "-", negative and statistically significant; "n.s." not statistically significant). The types of effects after mutant exclusion are the follows: gene co-expression inhibition (from statistically significant r_{mut+} to not significant r_{mut-}), induction (from not statistically significant r_{mut+} to significant r_{mut-}), inversion (from positive to negative correlation or viceversa) and changes of magnitude not affecting r sign and significance. For each type of effect, mean and 95% confidence interval of occurrence in

the gene pairs tested for each gene (N=199) were calculated. To assess the relevance of each effect type, t-tests for single samples were used to assess significant deviation of the effect occurrence from zero. To evaluate the relative prevalence of different types of effect, occurrences were expressed as percentage of the total number of gene pairs significantly affected by all types of effect.

2.6 Resources for *A. thaliana* gene co-expression

Many available platforms allow to perform gene co-expression analyses based on microarray, we focused on the 11 resources available for *A. thaliana*.

2.6.1 ATCOECIS

AtCOECiS (Vandepoele et al. 2009) is an online platform exclusively dedicated to *Arabidopsis thaliana*. It allows the user not only to identify co-expressed genes but also gene co-expression neighborhoods associated by cis-regulatory motifs or GO categories.

With the aim of verifying the guilty-by-association (GbA) relationship on a predefined set of genes, which establishes a link between gene expression trend and the gene function, they quantified the level of expression similarity using the expression coherence (EC). EC is a measure of expression similarity levels in a gene set, ranking between 0 and 1, and reporting the fraction of gene pairs per Gene Ontology (GO) category (Gene Ontology Consortium 2004) that shows elevated co-expression. Hereafter, the Pearson Correlation (PC) coefficient has been used as a measure to describe the similarity between expression profiles and three different thresholds, higher than 0.63, 0.72 and 0.83. The resulting output provides the gene annotation of

the query followed by the associated GO categories, the properties of co-expression neighborhoods, the cluster size, the clustering coefficient and the complete co-expressed genes list.

2.6.2 ATTED-II

ATTED-II (Obayashi and Kinoshita 2010) was released in 2007 and is a co-expression database expected to include *Arabidopsis thaliana*, rice, soybean, maize, grape, medicago and poplar. However only *Arabidopsis* genes may be currently queried. ATTED-II is organized primarily in two sections called “Search” and “Draw”.

“Search” section offers four tools to obtain information about gene functions and about their expression variations using different and global microarray datasets or user defined correlations list. These correlations are ordered by the Mutual Ranking (MR) algorithm: in this way, the result of a co-expression query for a specific gene in ATTED-II is the list of the first 300 genes ordered by their decreasing MR. The main benefit of Mutual Ranking value, in comparison with the most used PC values, is its lower sensitiveness to the differences within the tissues and experimental conditions of microarrays dataset.

In the “Draw” section are available four tools to visualize gene relations networks, hierarchical clustering, gene-to-gene co-expression and GO classification.

2.6.3 BAR

The BAR (Bio-Array Resource for Plant Biology) (Toufighi et al. 2005), from University of Toronto, is an on-line platform that offers several tools for

the management and exploration of the expression data, primarily in *A. thaliana*. The philosophy beyond BAR website is to offer simple and smart tools, developed with a user friendly interface.

The Expression Angler tool shows the best (or the worst) correlated genes with the query one, according to their PC value, calculated using one of the dataset described in table 2 or exploiting a customized one. Query results are available also with a heat mapping visualization format which let the user to have genes ranked by their PC values. Sample Angler is a tool aimed to detect a shared expression trend between two or more samples, chosen from a particular dataset or from a self-made one. Microarray slides similarity is expressed with PC value too and, according to this latter, a short ranking list with the heat mapping graphic is shown.

Arabidopsis Interaction Viewer tool offers a really detailed landscape of protein interactions, showing in one graph all the relations within a protein query list, defined by PC, experimental results and computational predictions obtained by associating interaction behaviors of orthologue proteins in yeast (*Saccharomyces cerevisiae*), nematode worm (*Caenorhabditis elegans*), fruitfly (*Drosophila melanogaster*), and human (*Homo sapiens*).

2.6.4 COP

CoP (Co-expressed biological Process) (Ogata et al. 2010) is an online platform with the main proposal of associating genes with similar expression profiles and biological information. This database contains the expression data from several plants, included Arabidopsis.

CoP takes into account only positive gene-to-gene correlation, exploiting the cosine correlation (CC), which considers only correlation between 0 and 1. The main approach to analyze gene co-expression on the website is choosing

“the gene-co-expression information” from the main page, using AGI code, probe id or gene name in the query form. In the result page, genes are listed not only by CC, but primarily by their Vertex F-measure (VF) [<http://webs2.kazusa.or.jp/kagiana/cop0911/pages/terms.html>], ranged 0-1, which indicates represents the stronger co-expression to a group of genes.. So genes with the highest VF values are chosen as the most co-expressed ones.

In the Cop website Network, modules of co-expression are identified through the “Confeito” algorithm which produces and ranks network modules according to the Network F-measure (NF), which is the harmony mean between the Network Recall (NR) and the Network Precision (NP) [<http://webs2.kazusa.or.jp/kagiana/cop0911/pages/terms.html>]

2.6.5 CORNET

CORNET (CORrelation NETworks) (De Bodt et al. 2010), released in 2009, is another on line microarray platform specific for *Arabidopsis thaliana*. The site offers two tools, namely co-expression and PPI tool.

The co-expression tool allows identifying genes with similar expression profiles with the query gene, exploiting one or more predefined expression datasets or a user-defined one. The correlations can be calculated either with Pearson or Spearman test, and the threshold can be fixed by the user. It is also possible to know the localizations of the proteins translated by the genes co-expressed and the output of this tool is a Cytoscape view of the correlations.

The two tools of the site can be exploited together, with only one query.

2.6.6 CressExpress

While the major part of co-expression databases provides a single oriented dataset viewpoint, CressExpress (Srinivasasainagendra et al. 2008) offers a more customizable approach in this field. Available since 2008, this resource allows to choose not only different microarray datasets collected from NASC website, but it offers the chance to select also the preferred chip platform and normalization method. As reported in the table, 4 microarrays dataset releases are selectable for co-expression analyses.

Co-expression among genes is expressed through r^2 , the square of the common used PC: its out coming p-values and slope numbers show the positive or negative nature of the correlation. In addition, CressExpress offers a pathway co-expression density analysis, defined as PLC (pathway level co-expression), which allows the user to have a ranking of the most co-expressed genes in an Aracyc pathway, with the ones chosen in the query, according to an user defined r^2 threshold, p-value and numbers of connections established by each gene

2.6.7 CSB.DB

The Comprehensive Systems Biology Project (CSB) (Steinhauser et al. 2004) website hosted at the Max Planck Institute of Molecular Plant Physiology was developed with the purpose of containing transcriptional correlations databases of key model organisms as *A. thaliana*, *Saccharomyces cerevisiae* and *Escherichia coli*. The first one, AthCoR@CSB.DB offers a co-expression querying platform based on four base 2 log normalized primary microarrays collection, one from the NASC's International Affymetrix Service and the other three from the AtGenExpress consortium. Three tools are available on AthCoR@CSB.DB. The first one is the Single Gene Query (sGQ) that allows to obtain the most correlated genes for a query one, according to the

expression profiles of one of the dataset chosen. Correlations among genes can be defined with a Pearson correlation test, or with a Spearman or Kendall one, while the query output can be customized in order to have genes shown according to their correlation value, statistical meaning or if belonging to a particular cell process or categories. The second tool, the multiple gene query, follows the same interface of sGQ but, however, it shows the correlations established only among a list of 60 genes of interest at most. The last tool of AthCoR@CSB.DB is the Intersection Gene Query (isGQ), that allows to choose two or three genes of interest and identify the most co-expressed ones with the ones stored in one of the four dataset defined during the query. Two or three lists of genes (according to the number of inputs inserted in the query form) ranked by their shared correlation degree are shown in the result page, each one with all the statistical and biological information described as in SGQ. Moreover if only two genes are selected as input query, results can be customized in order to have the best positive correlations with the first gene and the most negative ones with the second gene, and vice versa.

2.6.8 GeneCAT

GeneCAT (Gene Co-expression Analysis Toolbox) (Mutwil et al. 2008) is a multispecies database released in 2008, containing the gene expression values of *Arabidopsis thaliana*. After choosing one or more genes to query, it is possible to analyze the desired genes using 5 different tools.

“Co-expression analysis” tool is the core of co-expression investigation in GeneCAT: it compares the expression profile of the query gene to every other gene in the database, ranked by PC, which can be further filtered by a specific r-value threshold too. In order to point out a common biological role among the co-expressed genes shown in the list, the result page offers also

some facilities such as a co-expressed gene network built by measuring mutual co-expression ranks in a pair-wise manner between the 50 most correlated with the query term genes. Another tool of GeneCAT is Map-omatic which, after declaring a dataset of defined genes, allows the visualization of the Pearson values distribution of the correlations between these latter and the genes chosen for the query.

2.6.9 Genemania

Genemania (Mostafavi et al. 2008) released in the 2010, includes protein-protein interaction (PPI), literature, genomic and proteomic information from several on line datasets. All the data are integrated with the purpose to develop, or to define by the novo, the functional roles, the relations and the possible interactions of a single or multiple genes in several organisms, such as *A. thaliana*. The result of this investigation collapses in a graphical representation of a gene network built by different edges, each one describing the nature and the weight of the relation shared by two or more elements. The first step of Genemania query form is the definition of the dataset(s) to exploit in order to infer the relations among a group of genes. About 215 resources are available for Arabidopsis. The next step is to define the network weighting and Genemania offers 3 different set of choices: a query dependent weighting, a GO based method or a “based on equal weighting” set of preference. Results page offers the previously described gene network with each edge colored according to its criteria of relation and with a percent value describing its contribution in gene-to-gene association. Genes tab on the right shows the cellular function(s) associated to each element of the network, with a list of possible synonymous genes, while the function tab allows to visualize globally in the graph all the genes associated with one or more cellular process. Gene function association is statistically supported

with an FDR value and for each process a coverage indication is available too, which is equal to the number of query elements found in the network compared to the size of the full list of genes associated to that particular function.

2.6.10 Geninvestigator

Geninvestigator (Zimmermann et al. 2004), one of the most exploited bioinformatics resources since 2004, collects biomedical and plant biology genomics data of the most studied organisms, such as *Arabidopsis thaliana*. The first step of a co-expression analysis in the Geninvestigator query is the definition of a fully customizable list of platforms and datasets assortments, with the possibility to choose and relate single tissue or experiment combinations too if preferred.

The similarity search tools set on Geninvestigator allows to identify group of genes gathered by their expression profiles. The hierarchical clustering i.e. tool offer several ways to visualize genes association according to the distribution of these latter among samples, tissues and development stages or perturbations schemes. In a similar manner, user can cluster factors instead of gene, in order to identify expression trend shared by two or more samples and, moreover, genes and factors can be clustered together to obtain the elements with the most similar expression profile for both aspects. A sharper approach to cluster query genes in relation to the biological aspect considered is available in the biclustering tool and it is based on the Bimax algorithm. After choosing the factor to investigate in a user defined dataset, it is possible to search for cluster able to satisfy desired conditions as the smallest number of genes to hold within (min. probe sets), the smallest number of samples or factor elements to consider (min. factors), a minimum expression value and a

minimum up or down regulation degree if a perturbation scheme is the chosen factor at the beginning.

Co-expression tool completes the similarity search suite with the aim to identify the most co-expressed genes with a query one. Co-expression is measured with the Pearson correlation on the \log_2 transformed values of the dataset chosen and the results are depicted with a circular hierarchical clustering collecting the query gene in the center and the co-expressed ones around, with distances from the former defined by their Pearson value. Moreover, as for the clustering tool, a factor defined subset can be chosen to restrict co-expression analyses only to genes characterizing specific tissues, samples or conditions, and another added values it is the chance to filter out co-expressed genes according to their mutual correlation value.

2.6.11 PlaNet

PlaNet (Planet Network) (Mutwil et al. 2011) is a network website for *Arabidopsis thaliana* and other eight species.

On this website, there are a lot of useful features to evaluate *Arabidopsis* co-expressions: after choosing one or more genes for the query, as already seen in the previous databases, it is possible to observe the expression values variation among several tissues and/or experimental conditions. But the core of PlaNet database is its network tools package, based on the Highest Reciprocal Ranking (HRR) and on the Heuristic Cluster Chiseling Algorithm (HCCA). HRR (Highest Reciprocal Rank) is a variant of the Mutual Ranking algorithm seen in ATTED-II and it expresses the correlation strength between two genes, not through the geometric average between their rank positions in a mutual PC list, but by the highest rank position in these latter..-By keeping in a graph all genes within n steps away from the query gene, PlaNet offers a simple but powerful cluster representation, called node vicinity network

(NVN). This latter offers a quick graph of the most related genes with the query and, together with the HRR, this is the core of the HCCA. The main result of HCCA is the Meta Network page on PlaNet, which shows, in a comprehensive manner, pre-calculated best fitted clusters of correlated genes, in order to explore Arabidopsis transcriptome in the fastest way, or let the user to individuate the best pre calculated cluster which contains a query gene.

2.7 RNA-seq analysis in tomato

RNAs from transcriptome analysis were extracted from tomato leaves, in 4 different conditions, each of them with 3 technical replicates. The reads were sequenced exploiting Illumina technologies [<http://www.illumina.com/>] in paired-ends, with a coverage of 2x7 millions and an average length of 100 bases. Fastq sequences cleaning was performed by Trim Galore [http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/]. In the first step, low-quality bases were trimmed off from the 3' end of the reads. In the second step, Cutadapt (Martin 2011) removed adapter sequences; default parameters for paired end were used. Therefore, fastQC [<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>] software were used to check and assess the reads quality. Finally, the output generated was composed by two datasets: one with mate pairs and the other one for single reads. The two dataset generated by Trim Galore were aligned independently along the tomato genome (version 2.40) using Bowtie version 2.1.0 (Langmead and Salzberg 2012) and Tophat version 2.0.8 (Kim et al. 2013). After mapping, only reads one time mapped were counted per gene (iTAG annotation, version 2.3) with HTseq-count [<http://www-huber.embl.de/users/anders/HTSeq/>] version 0.5.4p1, with paired-end and “union” setting, using *Solanum lycopersicum* SL2.40.18 GTF, obtained from

Ensembl Plants Biomart section (<http://plants.ensembl.org/biomart/martview/27a472c92b73ab33ed10af02c668e8e9>).

Differential expressed genes (DEGs) analysis was performed by DESeq package (Anders and Huber 2010) version 1.10.1, one of the available R package that used negative binomial test for DEGs calling (FDR \leq 0.01). In order to define the set of expressed genes, raw read counts were normalized to RPKM (Reads per Kilobase per Million) and genes above the 1 RPKM cut-off were considered expressed and kept for the DEGs calling.

GO enriched analysis was performed by goseq package (Young et al. 2010) (FDR \leq 0.05). Median length per gene was extract with a customized script in R from gene length downloaded from Ensembl Plants BioMart. GO database exploited for the analysis was obtained performing BLAST2GO (Conesa et al. 2005) and GO Molecular Function (MF), Biological Process (BP) and Cellular Component (CC) terms were extracted from GO.db package.

Results

3.1 *Solanum tuberosum* available annotations

An overview of the *Solanum tuberosum* (potato) available data were carried out in order to check the potato annotation versions exploited by the on line resources (Fig. 5).

Potato's genomic, transcriptomics and proteomic data, can be obtained through online website, in particular Spud DB (Hirsch et al. 2014) that is the reference website for the Potato Genome Sequencing Consortium (PGSC). The platform is comprehensive of all the fasta files of the genome assemblies released from 2011 to 2013, in superscaffold and pseudomolecules level. Moreover it includes the annotation versions in GFF3 format and, for the old versions of the annotation, fasta files of genes, CDS and peptides.

Another resource that includes potato information is Ensembl Plants (<http://plants.ensembl.org/index.html>). In this platform, the assembly version SolTub_3.0 and the related annotation are included. In the description page, it is indicated that the assembly version is the same published in 2011, at scaffold level. However, the comparison between the annotation available on Ensembl and the one stored in SpudDB (version 3.4), is not possible, because the Ensembl version is based on chromosomes while the SpudDB version 3.4 is based on scaffolds.

In order to understand which annotation version was stored in Ensembl Plants, a comparison among this latter and all the chromosome-based annotations available on SpudDB was performed.

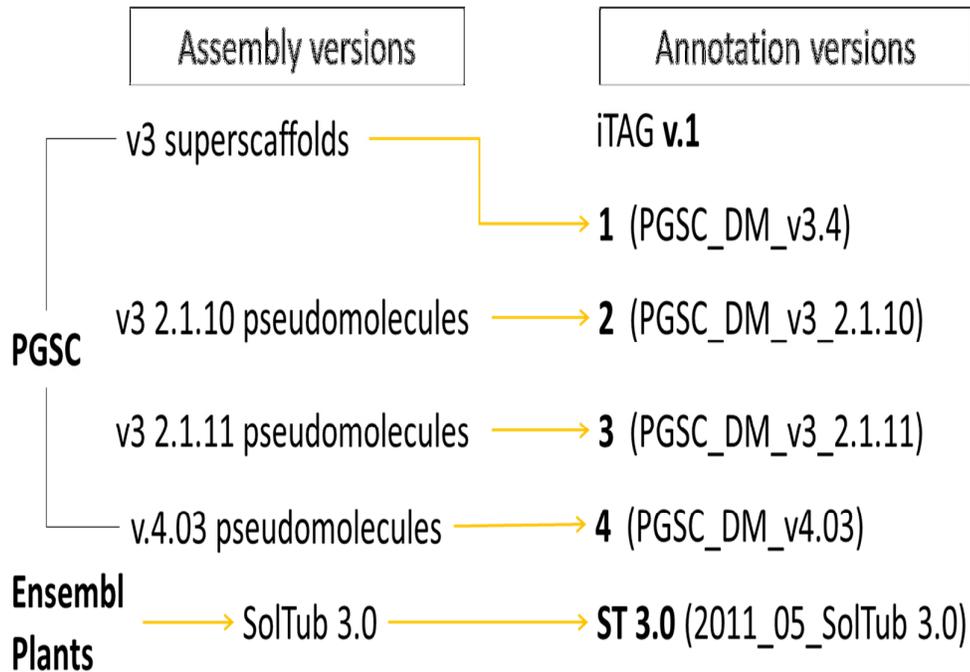


Figure 5 List of the different genome assemblies and their related annotations, available from SPUDdb website (<http://solanaceae.plantbiology.msu.edu/>) and Ensembl Plants (<http://plants.ensembl.org/index.html>)

Even though the gene names among all the annotation version considered were identical, the total number of genes and their genome positions are different, unless for v. 2.1.10 and 2.1.11 (Tab. 2). However, observing carefully to the gene number per annotation, it was evident that the one of EnsemblPlants was similar to the version 4.03, the most recent one. Therefore, a more deeper comparison between the two annotations was performed, highlighting that the 91,5% of the genes have the same start and end positions, suggesting that the version exploited in Ensembl Plants is based on the version 4.03 and not on the version 3.4, as wrongly indicated on the website. Moreover, in Ensembl Plants annotation there were 3953 genes with a different Gene ID nomenclature, that are lacking in all the other versions released by the official resource.

Table 2 Number of the gene from the different potato genome annotations analyzed. Percentage of the identical locus per pair comparison is also reported.

	Ensembl	v. 4.03	v. 2.1.10	v. 2.1.11
# genes	39021	39028	35119	35119
% exact annotation	91,5		78	

Information about potato genome annotation can be also obtained through a reference platform for all the plant genomes: PlantGDB (Duvick et al. 2008). In this platform the potato annotation version stored is v. 2.1.10, indicating that all the information that can be taken out from that website are obsolete.

The importance of knowing the most update version released for a genome and, in particular, knowing which version is used, it's a relevant issue in all the relayed analyses, such as orthologue gene detections.

In this frame, we checked the potato annotation versions exploited in some of the most widely used orthologue platforms, such as Phytozome (Goodstein et al. 2012), Plaza (Proost et al. 2009), GreenPhyl (Conte et al. 2008) and EggNog (Powell et al. 2011) (Tab. 3).

The results reported in Table 3 underlines the information heterogeneity of all the platforms taken into consideration and put a light on the fact that none of the available resource for ortology searches is using the most updated potato annotation version.

Table 3 Potato annotation versions exploited in Phytozome, Plaza, GreenPhyl and EggNog

			
http://www.phytozome.net/	http://bioinformatics.psb.ugent.be/plaza/	http://www.greenphyl.org/v3/cgi-bin/index.cgi	http://eggnog.embl.de/version_4.0.beta/
2 (v2.1.10)	ITAG v. 1	1 (v3.4)	ST 3.0

3.2 *Solanum lycopersicum* annotation

The tomato genome was released in 2012 (Tomato Genome Consortium 2012), and now it is considered a model for other Solanaceae species. However, tomato is still far away to be a real model due to the lacking of information and problems in the official annotation.

In 2012 with the release of the tomato genome the tomato annotation version 2.3 was released as well by the iTAG consortium. In 2014 an update of the genome was released and with the new genome assembly (Shearer et al. 2014) the 2.4 iTAG annotation was available. Comparing the two genome assemblies of tomato, SL2.40 released in 2012 and SL2.50 released in 2014, it is clear that the length of the 12 chromosomes is changed, indicating that new sequences were added to the previous assembly (Tab. 4) and from a dotplot of the twelve chromosomes of SL2.40 vs the ones of SL2.50 (Fig. 6) it is highlighted that some chromosome sequences had different positioning.

Results

Table 4 Number of nucleotides (nt) per chromosome in version SL2.40 and SL2.50. Number of A, T, C, G and N per chromosomes is also specified. In yellow, the number of nt reported in GFF3 of iTAG 2.4 per chromosomes 09 and 10 versus the real number (in bold)

	TOT		A		T		C		G		N	
	vers. SL2.40 (nt)	vers. SL2.50 (nt)	vers. SL2.40 (nt)	vers. SL2.50 (nt)	vers. SL2.40 (nt)	vers. SL2.50 (nt)	vers. SL2.40 (nt)	vers. SL2.50 (nt)	vers. SL2.40 (nt)	vers. SL2.50 (nt)	vers. SL2.40 (nt)	vers. SL2.50 (nt)
chr01	90304244	98543444	28545110	28543252	28527867	28529725	14560935	14550787	14486151	14496299	4184181	12423381
chr02	49918294	55340444	15681575	15671870	15687901	15697606	7941475	7928334	7946061	7959202	2661282	8083432
chr03	64840714	70787664	20077058	20074769	20119209	20121498	10305461	10320842	10360086	10344705	3978900	9925850
chr04	64064312	66470942	20063213	20046332	20021016	20037897	10170517	10173967	10194277	10190827	3615289	6021919
chr05	65021438	65875088	20205055	20186419	20220814	20239450	10403986	10402824	10410775	10411937	3780808	4634458
chr06	46041636	49751636	14385825	14372797	14395274	14408302	7382860	7372715	7408992	7419137	2468685	6178685
chr07	65268621	68045021	20439152	20439152	20390614	20390614	10584159	10584159	10546887	10546887	3307809	6084209
chr08	63032657	65866657	19731518	19732817	19678972	19677673	10208064	10176256	10166134	10197942	3247969	6081969
chr09	67662091	72389422 / 72482091 65509773 /	21388978	21382831	21358659	21364806	11051986	11048464	11068160	11071682	2794308	7614308
chr10	64834305	65527505	20073867	20073867	20078738	20078738	10311029	10311029	10327550	10327550	4043121	4736321
chr11	53386025	56302525	16531010	16552154	16600102	16578958	8553618	8592589	8572555	8533584	3128740	6045240
chr12	65486253	67145203	20336474	20346043	20306478	20296909	10561869	10585944	10601561	10577486	3679871	5338821
TOT	759860590	664128624										

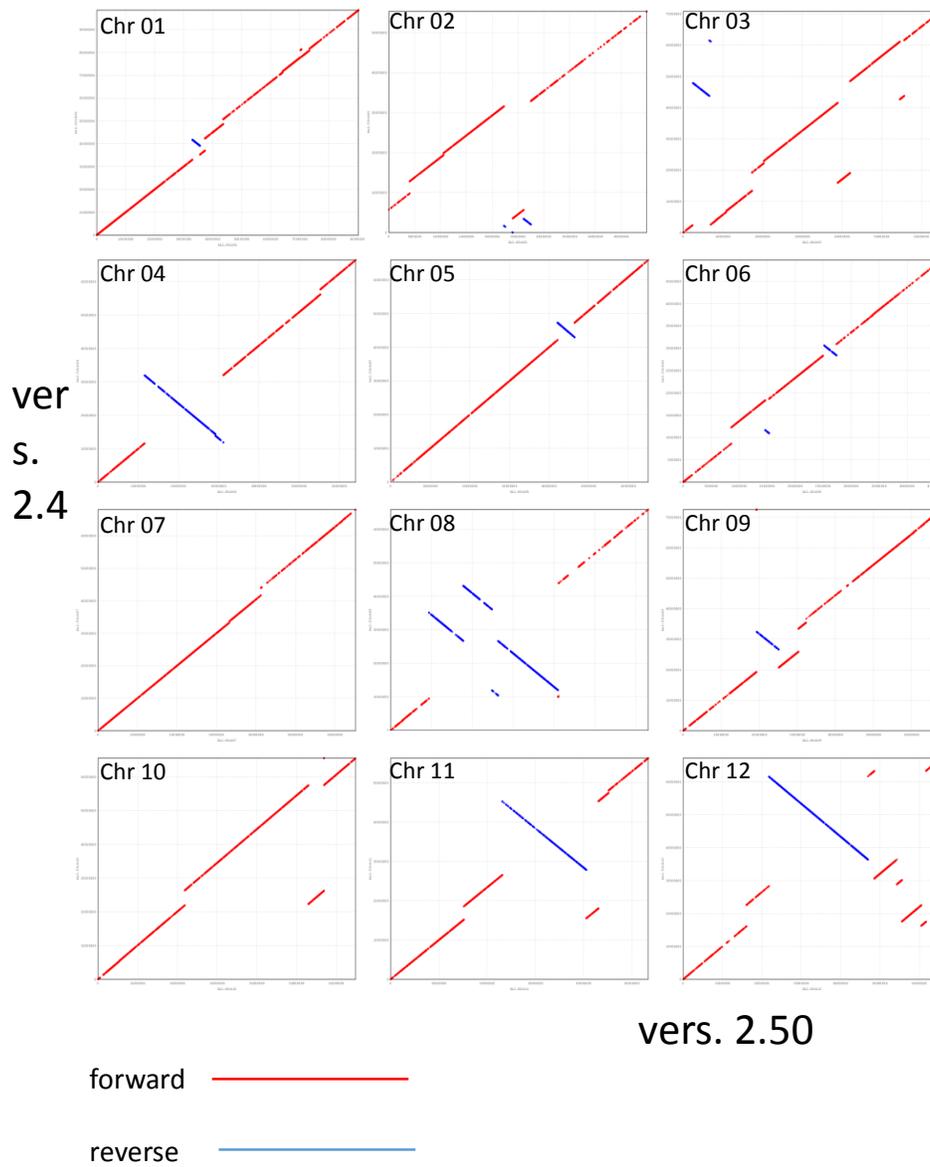


Figure 6 Dotplots between genome assembly version SL2.40 and SL2.50 (Shearer et al. 2014)

Although the new genome assembly seems to be different from the previous one, if we put attention on the number of nucleotide added and the nucleotide of new N, they are exactly the same. This results underlined that although new sequences were added to the new assembly, they were entirely composed by N nucleotides. The variation in the number of the other bases was due to the fact that some chromosome pieces were changed in the orientation (Tab. 5).

Table 5 Delta of the number of nucleotides (N included) between genome assembly version SL2.40 and SL2.50

	Delta 2.50 - 2.40					
	TOT	A	T	C	G	N
chr01	8239200	-1858	1858	-10148	10148	8239200
chr02	5422150	-9705	9705	-13141	13141	5422150
chr03	5946950	-2289	2289	15381	-15381	5946950
chr04	2406630	-16881	16881	3450	-3450	2406630
chr05	853650	-18636	18636	-1162	1162	853650
chr06	3710000	-13028	13028	-10145	10145	3710000
chr07	2776400	0	0	0	0	2776400
chr08	2834000	1299	-1299	-31808	31808	2834000
chr09	4820000	-6147	6147	-3522	3522	4820000
chr10	693200	0	0	0	0	693200
chr11	2916500	21144	-21144	38971	-38971	2916500
chr12	1658950	9569	-9569	24075	-24075	1658950

In addition, we compared the two released annotations. The version 2.3 and the 2.4 one were different only in gene numbers: version 2.3 had 34727 genes while the 2.4 one had only 34725 genes. However, this is the only difference between them. In fact, the structure of the 34725 genes in common was exactly the same: total gene length, mRNA length, exon structures and lengths were not changing through the annotations. Only gene positions in the version 2.4 were in part different, due to the added pieces of N sequences (Fig. 7).

The update of the tomato genome assembly and the release of the new annotation was only an adding of N that creates problems in many different analyses.

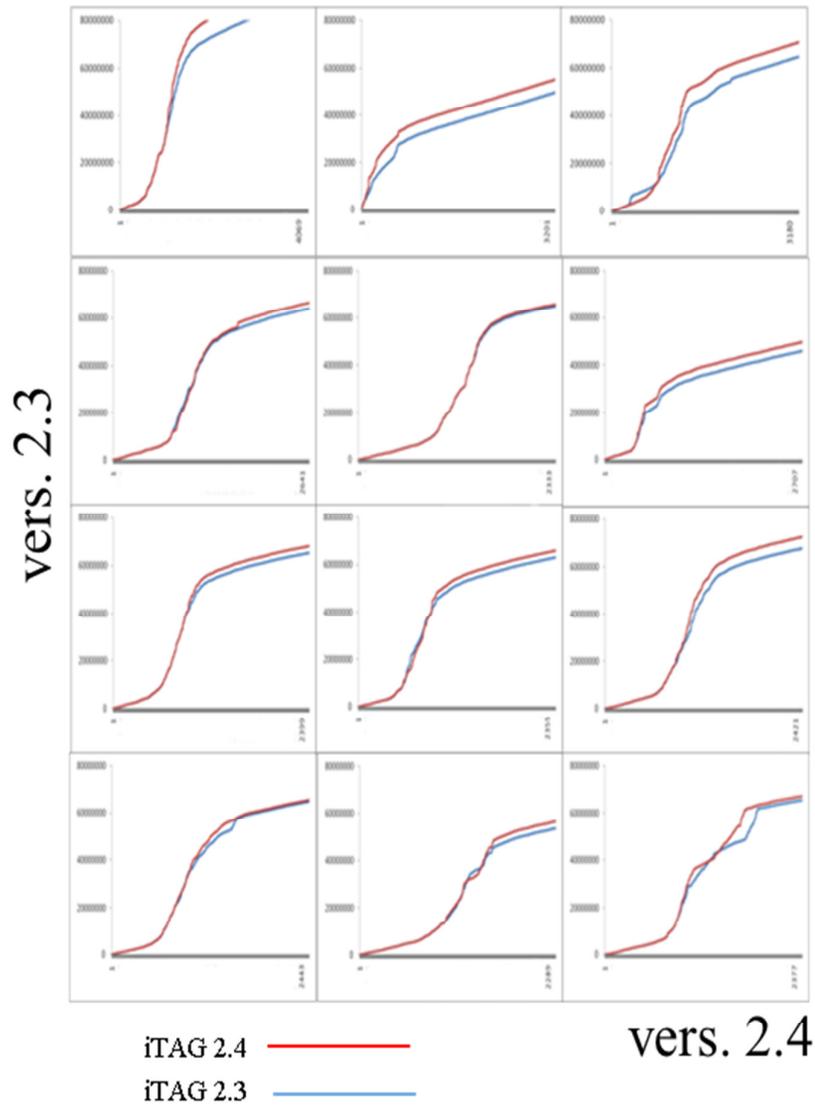


Figure 7 Plots of gene positions on the twelve chromosomes of SL2.40 and SL.50

3.3 Tomato repeated sequences

From the tomato official annotation released, information about some of the repeated regions were lacking. In particular, long rDNAs: large subunit (LSU) and small subunit (SSU) were excluded from the analyses of the tomato annotation by the consortium, because of a specific option that avoids the annotation of these specific regions.

Therefore the analysis resulted to be limited to the identification of 1,853 non-coding RNAs of 90 distinct Rfam families in which almost 48% of all the targets represented tRNAs (RF00005) (Tomato Genome Consortium 2012). To fulfill this limitation, we annotated independently LSU and SSU, enriched the tomato repeats annotation (Tab. 6).

	5.8S RNA	5S rRNA	tRNA	SSU	LSU
chr 00	11	3	16	4	20
chr 01	2	38	109	4	9
chr 02	0	1	76	1	6
chr 03	2	3	83	5	6
chr 04	0	1	71	1	4
chr 05	3	0	60	2	1
chr 06	7	0	102	5	11
chr 07	1	2	52	2	4
chr 08	0	0	70	1	8
chr 09	0	2	44	2	3
chr 10	0	0	90	1	8
chr 11	13	4	48	12	21
chr 12	1	0	64	2	6
sum	40	54	885	42	107

Table 6 *Distribution per chromosome of 5.8S, 5S, tRNA, SSU and LSU RNA*

From Table 6 it was highlighted that 5.8S rRNA genes were listed mainly on chromosome 11 and 6, and eleven genes were still on unassigned sequences collected in chromosome 0. High number of 5S genes on chromosome 1

confirmed the loci identified as repeated in tandem by FISH on pachytene chromosomes on the short arm of chromosome 1 (1S), close to the centromeric region (Vallejos et al. 1986, Lapitan et al. 1991, Xu and Earle 1996).

As well as 5.8S, also LSU had the higher copy numbers on chromosomes 11, 6 and 0. Meanwhile, SSU were concentrated not only on chromosome 11 but also on chromosomes 3 and 6. Finally, tRNA were the larger non coding RNA family annotated. They were 885 located especially on chromosomes 1 and 6.

Even though tRNA were not generally in tandem on the genomes, we found 15 tRNA genes tandemly located on chromosome 1 (Fig. 8.B). Moreover, always on chromosome 1, 37 5S genes were found repeated in tandem (Fig. 8.A).

3.4 Tomato annotation problems

3.4.1 Overlapping genes

In the tomato genome annotation version 2.3 there were 1309 predicted genes that overlapped a consecutive gene. However, out of that, only 664 genes overlapped a gene on the same strand. In this cases, the overlapping can be of few nucleotides to 100% overlap (Fig. 9).

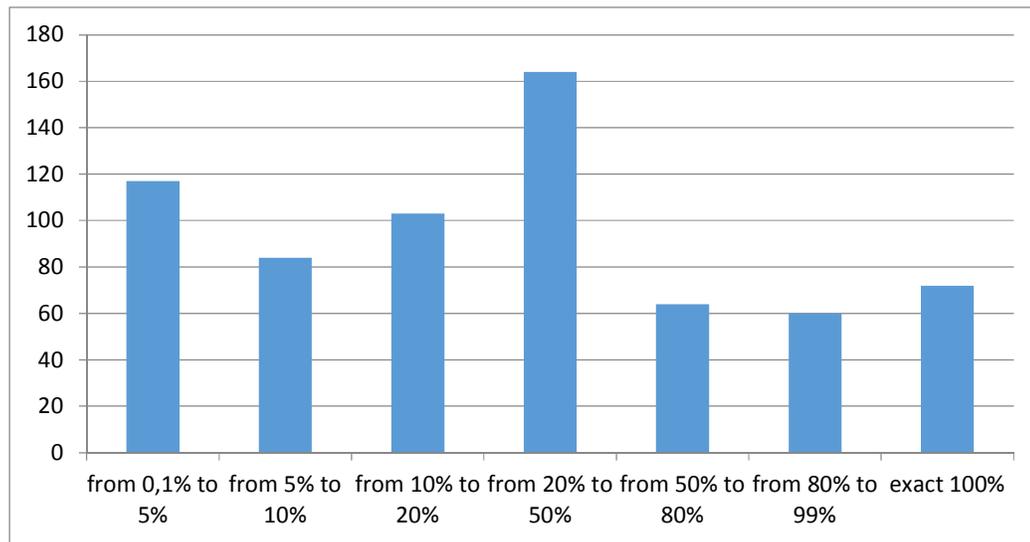


Figure 9 *Number of overlapping genes divided by the percentage of overlap*

The overlapping genes were distributed equally on the chromosomes 0 to 10, but there was no gene overlapping on chromosomes 11 and 12. It is interesting underline that more than 70 genes were completely overlapping with another gene on the same strand. This kind of situation can create several problems in many different kind of genome analyses.

Out of the overlapping genes, we notice that on chromosome 1 there were three genes (Solyc01g088230, Solyc01g088210 and Solyc01g088200) that were exactly overlapping each other's, from the start until the end of their locus. This three genes were annotated as "Xanthine dehydrogenase/oxidase"

in the case of Solyc01g088210 and Solyc01g088230, and as “Aldehyde oxidase” in the case of Solyc01g088200. The structure of the genes was the same: they had 10 exons with same start and end. The only thing that changed was the CDS positions that never overlapped with the other ones (Fig. 10).

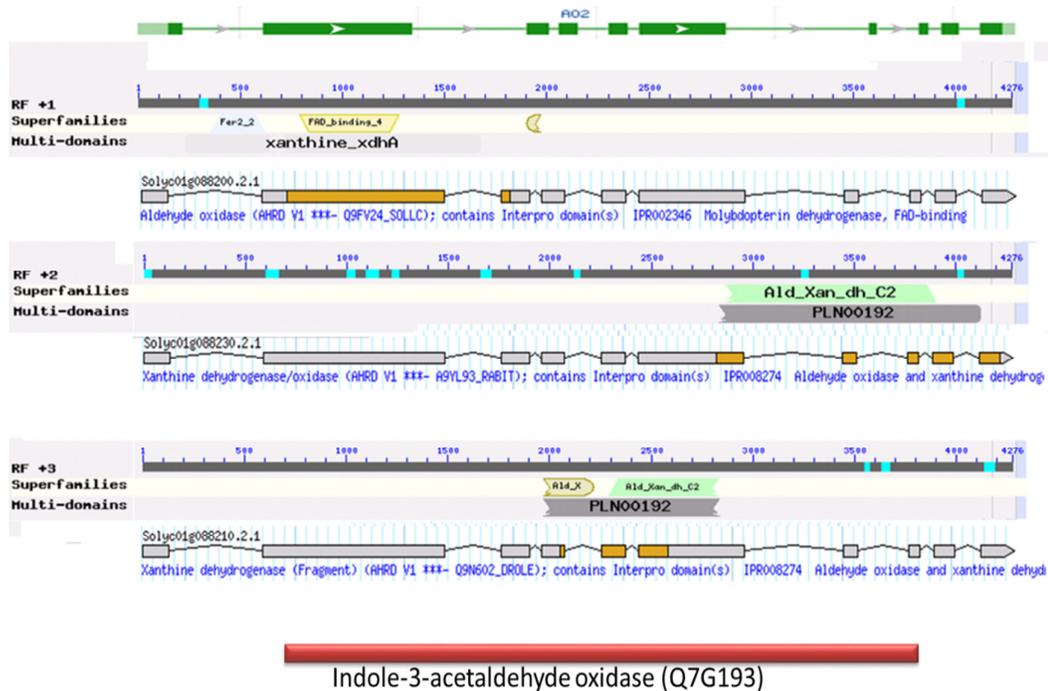


Figure 10 *Three exactly overlapping iTAG predicted genes codifying probably one protein*

In order to understand if the genes were alternative transcripts of the same locus or they were wrongly annotated as three instead of only one, their mRNA were first aligned versus the protein database in NCBI with a BLASTp. It resulted that the mRNA were aligned completely with one protein, which contains all the three domains, on different frame, annotated as belong to the different locus. This result suggested that the three genes

codified the same protein and they were probably annotated as three different genes instead of only one.

Still on chromosome 1, two genes, Solyc01g110700 and Solyc01g11180, resulted to be very long: 244094 nucleotides and 214622 nucleotides respectively (Fig. 11). These genes had a very long putative, probably wrongly predicted, 3' UTR that overlapped with other 53 genes.

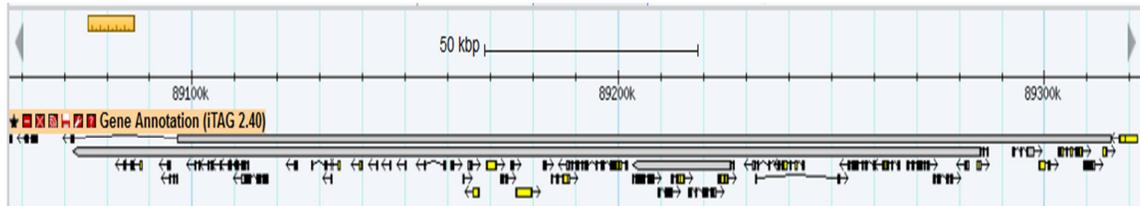


Figure 11 Snapshot of Genome Browser of predicted genes overlapping 53 other genes on chromosome 01

Overlapping loci are a problem for different expression analysis, i.e. RNA-seq analysis, since a read considered not specific for a locus is classified “ambiguous” and, in general, it is not counted at all. This became a big problem when a locus is completely included in another one, like the cases cited above, because all the reads of the locus will not be counted and the locus will be considered as not expressed at all. In the tomato genome the problems of overlapping genes regard especially UTR: in fact, if we count the overlapping CDS in the tomato annotation, only two CDS result to be overlapping each other, both cases on different strands.

3.4.2 Putative split genes

In order to check if there are other cases of genes annotated as two or more instead of one, we made a BLASTx query of the mRNA versus the

UNIPROT database and we took the consecutive annotated genes that matched the same protein (Fig. 12).

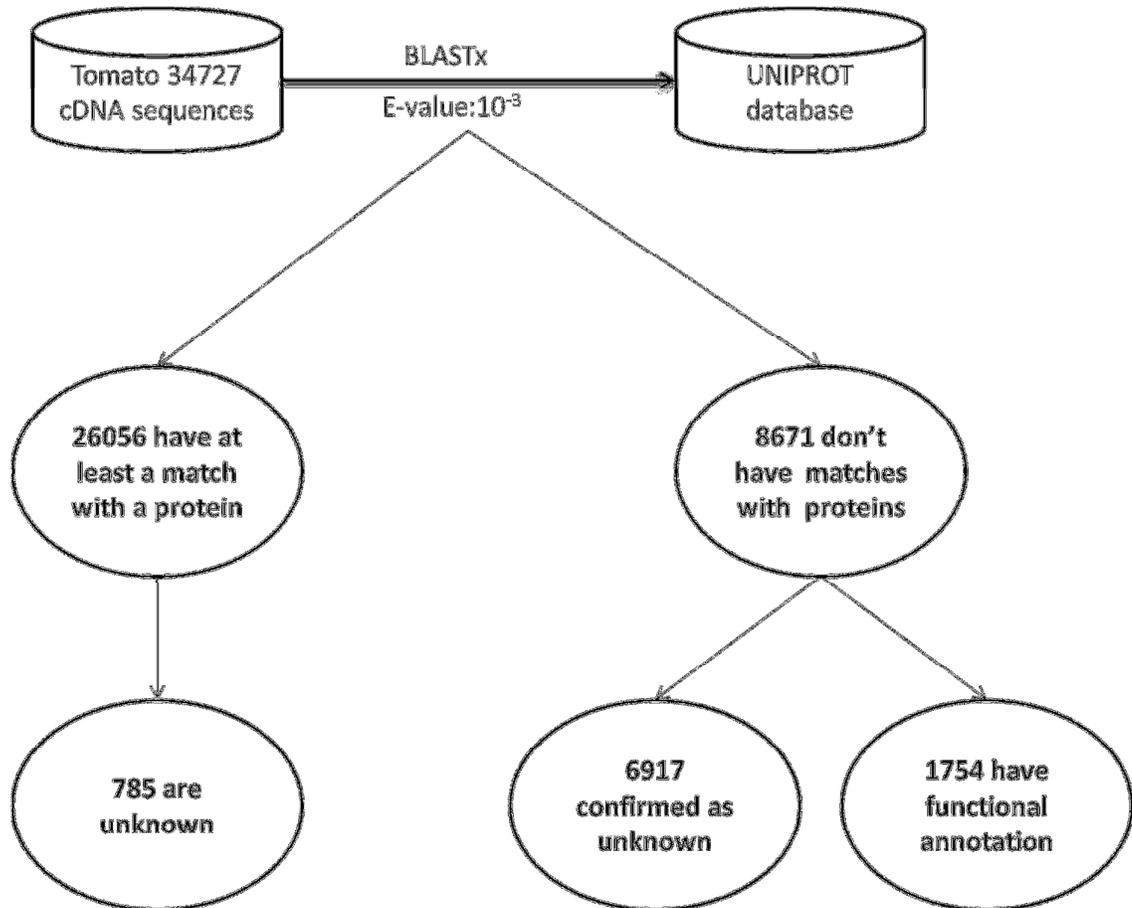


Figure 12 *BLASTx results of mRNA versus UNIPROT database*

As result of the alignment, 8671 mRNA didn't found match with proteins: among these, 2873 had a functional annotation. Since they not belong to a specific gene family manually checked, it is an open question on how their function was predicted. On the other hand, out of the 34727 mRNA aligned, 26056 had a match with at least one protein. Among these, 785 were still unknown genes for the current annotation, but their putative function could be upgraded.

From the 26056 mRNA that found match with at least a protein, we extracted consecutive locus that matched the same protein checking if they aligned the same or different position of the protein. We found out that 1878 genes matched the same protein of their consecutive gene but in different position, i.e. the first gene was aligned in the first part of the protein and the second gene was aligned in the last part of the protein, indicated that probably the genes were wrongly annotated as two instead of one.

However, not only two consecutive genes were found matching the same protein, but also more genes, up to 13 consecutive genes, that matched all the same protein in different position (Tab. 7). For example, 95 groups were formed by 3 consecutive genes and 40 groups were formed by 4 consecutive genes.

splitted genes						
# genes	# groups					
1878	766					
	2 genes	3 genes	4 genes	5 genes	6 genes	7 - 13genes
	595	95	40	16	8	12

Table 7 *Number of putative split genes and number of groups with a certain number of consecutive genes matching the same protein*

In order to confirm that the genes annotated separately could be annotated as one, we merged the mRNA of a group of four consecutive genes: Solyc11g067110, Solyc11g067120, Solyc11g067130 and Solyc11g067140, and we made a BLASTx versus the protein database, in NCBI.

Solyc11g067110, Solyc11g067120, Solyc11g067130 and Solyc11g067140 were located on chromosome 11, covering the chromosome region from 49933152 to 49970526. They are long genes with complex structures:

Solyc11g067110 was 6733 nucleotide long with 12 exons, Solyc11g067120 was 2977 nucleotide long with 5 exons, Solyc11g067130 was 7623 nucleotide long with 16 exons and finally Solyc11g067140 was 10796 nucleotide long with 20 exons. The four genes were predicted codifying for a DNA polymerase.

The best results of the BLAST was a “DNA polymerase epsilon catalytic subunit A” (A.N. F4HW04, 2161 aa) that was covered by the merged mRNA from the 4th until the 2156th amino acid.

The four mRNA merged were aligned versus the NCBI nucleotide database with a BLASTn, and as best result we found a tomato mRNA transcribed from the locus LOC101253967, annotated with the RefSeq method, which putative function is “DNA polymerase epsilon catalytic subunit A-like”. LOC101253967 was annotated on chromosome 11, had 49 exons, and it was located from the nucleotide 49933064 to 49971160 on the chromosome, overlapping completely the four genes examined, with an extra portion on 5' and 3' that covered the regions lacking on the protein found with the BLASTp, from the previous analysis. To try to understand the real possible genomic structure of “DNA polymerase epsilon catalytic subunit A”, we searched the gene codifying for it in the model organism *A. thaliana*, by BLASTn. The gene found in Arabidopsis is AT1G08260, annotated on chromosome 1, 15949 nucleotide long with 49 exons (Fig. 13).

This results showed how the putative structure of the gene that codify for the “DNA polymerase epsilon catalytic subunit A” is not done by four different genes but probably by only one long gene.

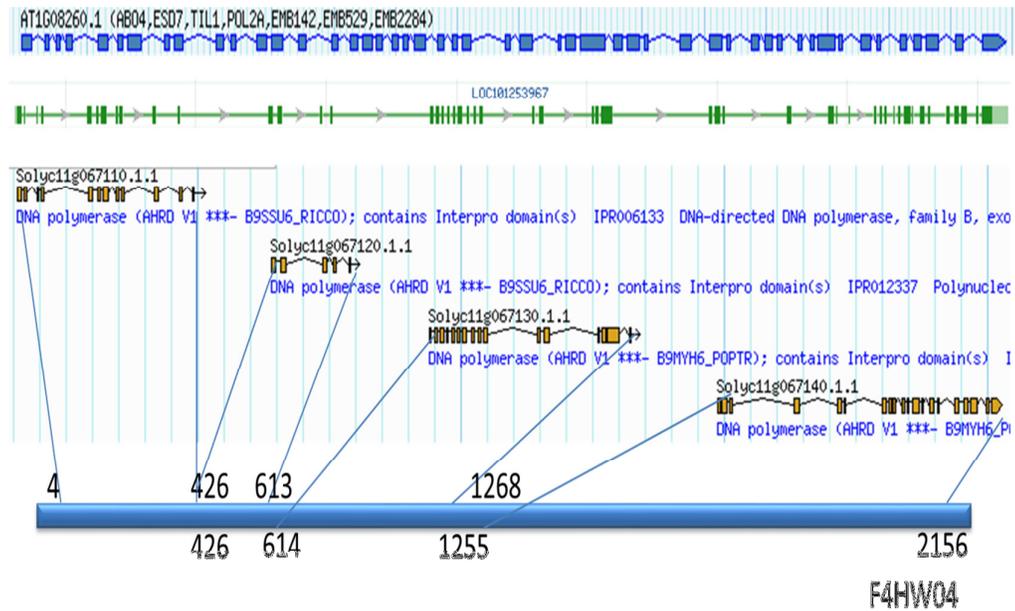


Figure 13 Snapshot of the four putative split genes on Genome Browser. Details of the alignment position of the genes on the protein F4HW04 are specified. Moreover, the structure of AT1G08260 gene, from *Arabidopsis thaliana*, and LOC101253967 gene, from *RefSeq* of tomato, is also shown

3.4.3 Curation of some tomato gene families

At the light of the resulted putative split genes described previously, we reviewed some gene families that were manually curated after the release of the tomato genome (Tomato Genome Consortium 2012) (Tab. 8).

In Table 13, are reported some of the tomato gene families manually curated based on the iTAG 2.3 annotation. Comparing the family members with the results of the putative split genes, the gene number of these families changed. In particular, the number of genes in the families curated during the release of the genome differs significantly. For example, the number of tomato Transcriptional Factor in iTAG 2.3 was 2459, after the comparison with the putative split genes, the new gene number was 2499, 40 genes more. This is

due to the fact that 95 genes were resulted as putative split, and 88 genes that were not annotated as TF resulted, after the BLASTx, having a match with a TF.

Table 8 *List of annotated gene family in tomato. Reference, gene reference number, number of putative split genes, number of added genes and new putative reference number is reported*

FAMILY	Reference	Reference number	Number of gene considered as one	Number of gene added	New Reference number
Cycline	Zhang et al. 2013	52	4	1	51
R-Genes	Andolfo et al. 2013	52	10	8	55
S1MLO	Chen et al. 2014	17	2	0	16
lePT1	Chen et al. 2014	9	0	1	10
S1HAK	Hyun et al. 2014	19	0	1	20
ARF	Zouine et al. 2014	24	4	1	23
NB-LRR	Andolfo et al. 2014	221	40	35	233
C3H	Xu 2014	80	10	9	85
GST	Csiszár et al. 2014	81	0	2	83
SIMAKKK	Wu et al. 2014	89	3	6	94
Cell Wall	Tomato Genome Consortium 2012	718	52	25	715
TF	Tomato Genome Consortium 2012	2459	95	88	2499
Cytp 450	Tomato Genome Consortium 2012	464	0	1	465
Cytp 450	Suresh et al. 2014	263	0	0	263
TF	Suresh et al. 2014	2458	103	66	2416
R-genes	Suresh et al. 2014	512	55	47	523
HSP	Suresh et al. 2014	153	0	1	154
KINASE	Suresh et al. 2014	1780	127	46	1759
TRANSPORTERS	Suresh et al. 2014	752	99	31	724
Ripening	Suresh et al. 2014	129	6	9	135

3.5 Remapping the tomato genome

The problems of the tomato genome annotation were also underlined through a remapping of the tomato mRNAs (iTAG vers. 2.3) on the tomato genome (SL2.40).

The result of the alignments showed that out of 34727 mRNA, 27968 mapped only once in their predicted position (Tab. 9).

Table 9 *Results summary of the mapping of tomato mRNA (iTAG v. 2.3) on the tomato genome (SL2.40)*

Total number of transcripts	Number of transcript mapped only one time	Number of transcript mapped more than one time	Number of transcript not mapped
34727	30046	4593	88

Confirming prediction	Not confirming prediction	Confirming prediction	Not confirming prediction
27968	2078	4165	428

The other 6759 mRNA had different behaviors:

- 2078 were mapped only once on the genome but not in the correct predicted region;
- 4165 were mapped on their predicted region but also in other regions on the genome;
- 428 were not mapped on their predicted regions but are mapped multiple time somewhere else;
- 88 were not mapped.

For this latter category, a BLASTn was performed to check if these mRNA were not mapped due to a software limit. The result of the BLASTn showed that out of 88 mRNA not mapped with Genome Threader, 62 were exactly found in their predicted region, and meanwhile 24 mRNA were only overlapping with the locus of their predicted region. Two mRNA were not aligned also with BLASTn. These latter are the two long mRNA codified by the very two long mis-annotated genes cited before (Solyc01g110700 and Solyc01g11180).

3.5.1 mRNA not mapped in their predicted region

2506 mRNA that had only one match or multiple matches on the tomato genome were not mapped in the predicted annotation (Tab. 10). In some cases, the mRNA were mapped in the same region of their predicted position but with a different start (832 mRNA) or different end (557 mRNA), different start and end but still overlapping the locus (17 mRNA). In 1087 cases the start and the end of the remapping was the same, but the structure of the exons was different. These categories can be explained by the combination of the parameters of the tool that is biased by the minimum and maximum length of the introns given in input and by the repeated regions in this gene area. However, in 9 cases the mRNA were mapped on the same chromosome of the predicted annotation but not overlapping it, meanwhile in 4 cases the mRNA were mapped in completely different chromosomes compared to the predicted region.

Interesting is the fact that in 112 cases even if the mRNA was found in its predicted region they were mapped on the other strand.

Table 10 Number of mRNA that were not mapped in their predicted locus

mRNA not confirming predicted annotation					
different exon position	different start	different end	Different start and end		different chromosome
			overlapping	different location	
1087	832	557	17	9	4

3.5.2 mRNA mapped more than one time

Out of all the mRNA remapped with a percentage of coverage ≥ 80 and a percentage of identity ≥ 90 , we focused only of the 2256 mRNA remapped more than one time which had coverage and identity $\geq 95\%$ (Fig. 14).

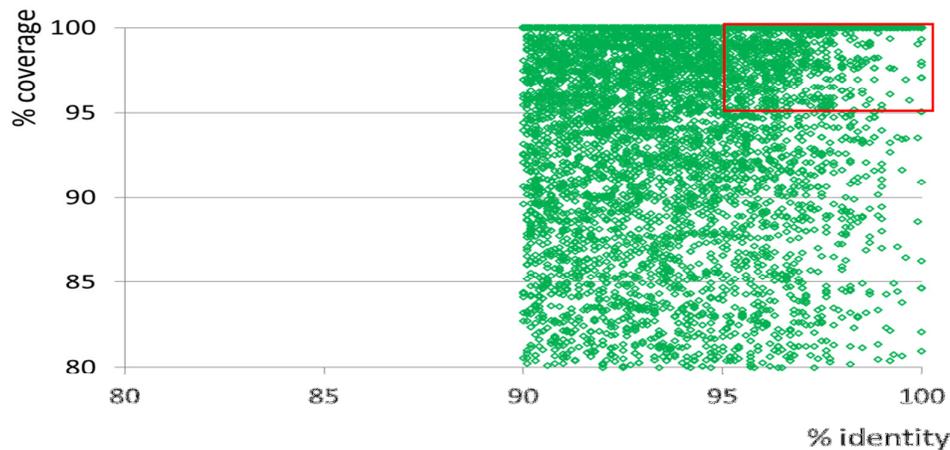


Figure 14 Plot of the remapped genes based on % of identity and coverage. In the red box are highlighted mRNA remapped with an identity and coverage $\geq 95\%$.

One of the mRNA that mapped more than one time was codified by Solyc00g005070 gene and alone had 287 duplications among the genome. This gene was predicted on chromosome 0, had 2 exons and its functional

annotation was unknown. The gene length was 244 nucleotides and its mRNA consisted only of 81 nucleotides, highly repeated (Fig. 15), suggesting that probably this sequence was wrongly annotated as gene.

Solyc00g005070.1.1 (Unknown Protein; 2 exons) has 287 duplications; mRNA sequence:
 ATGCTTCTAGCTTGGACTGGATCTTCTTCTTCAAGTCTTGATGCCTTGAAGTCCGGCATGGACTAGCTTCTT
 ATGTTTAG

Figure 15 mRNA repeated sequence of Solyc00g005070

The other 2255 mRNA had different number of duplications, from 1 to 93 duplications per mRNA, with different percentage of coverage and identity, with 8070 duplications in total (Fig. 16).

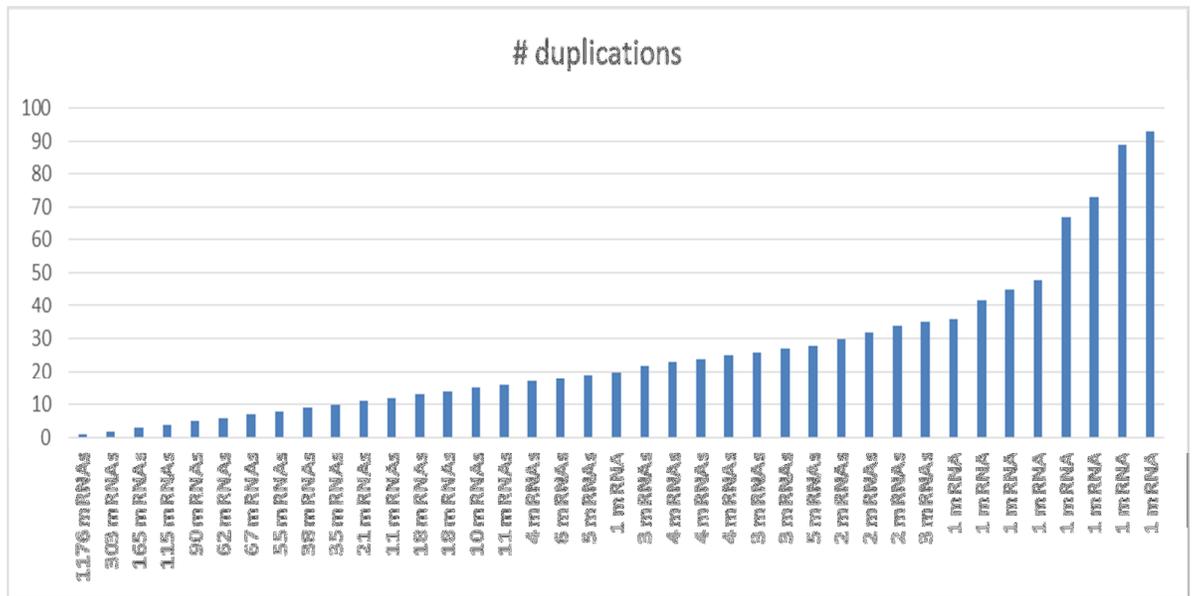


Figure 16 Number of duplications per mRNA

The remapping procedure also revealed that 228 mRNA of chromosome 0 mapped with high identity and coverage on other chromosomes. As it is

shown in the Fig. 17, in some cases the mapping was with 100% identity and 100% coverage. These latter were 18, and out of them 8 were mapped overlapping other predicted genes meanwhile 10 were remapped in area without predicted genes, indicated or the possible real position of that genes or the presence of a still non-annotated genes.

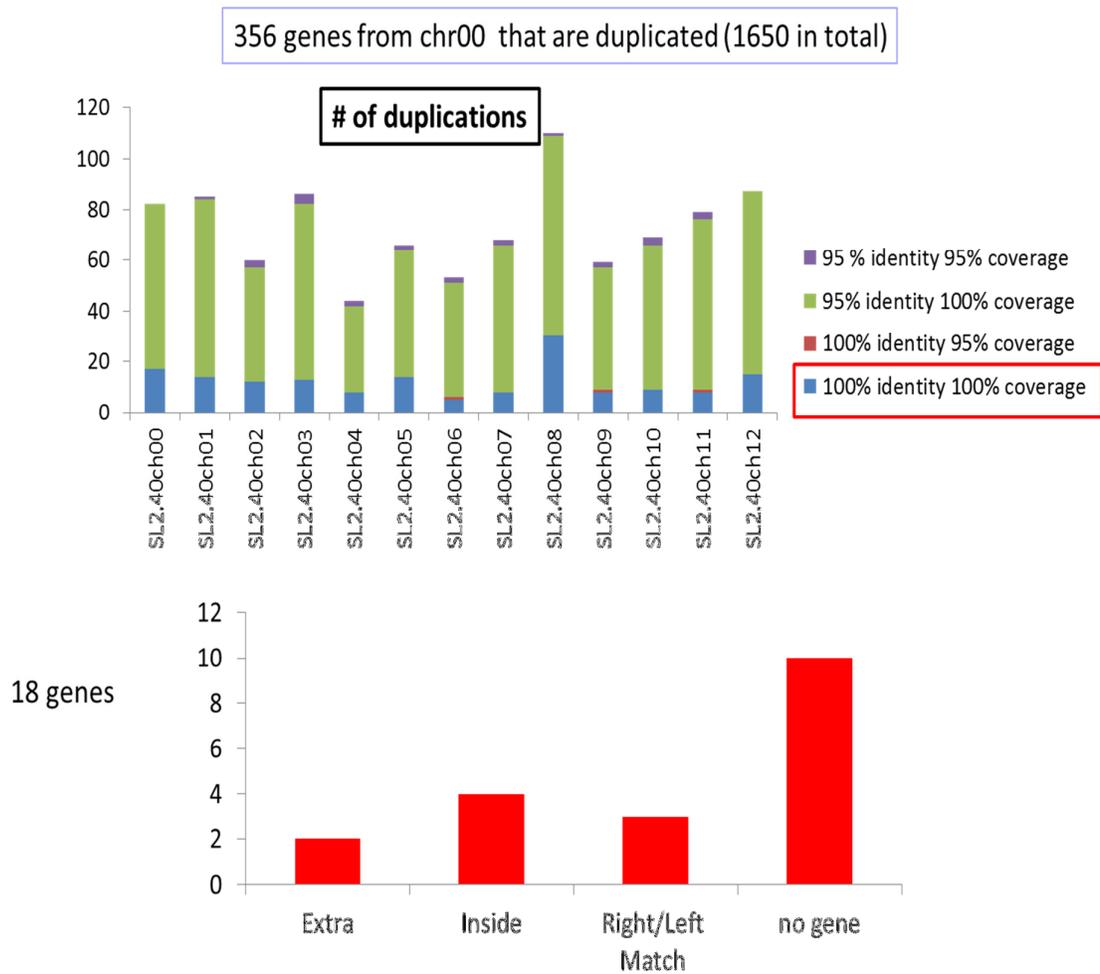


Figure 17 The detailed information of genes from chromosome zero mapping on other chromosomes with high identity and coverage with the type of their remapping overlapping on the predicted genes

3.5.3 Identical genes

Focusing on the mRNA that were remapped 100% of coverage and identity on the genome, we noticed that in some cases the remapped position of this mRNA was exactly the same of another predicted gene. Going into details, we compared the sequence of mRNA of this latter genes with the one remapped in the same position and it resulted that they have the same sequences.

Moving to the gene levels, we compared first the gene structure and then the gene locus sequences, including the intron, if presents, and 50 nucleotides before and after the gene locus. The results showed that the genes were identical in sequences and in structure (Tab. 11).

Checking the locus that were the same, we notice that two consecutive genes on chromosome 1 were identical to two consecutive genes on chromosome 9. Therefore, we performed a dotplot between the chromosome pieces of chromosome 1 and 9 that included the consecutive identical genes (Fig. 18).

The result of the dotplot highlighted that the two sequences of the different chromosomes are perfectly identical, suggesting a misassembly of the genome in those regions.

Table 11 List of identical genes. Per each genes is specified: length, exons number, identical region 50 nt after and before, strand and alignment coverage

	Sequences length (nt)	# exons	equal (included 50nt before and after the gene area)	strand	coverage
Solyc00g011550.1.1\ Solyc03g042510.1.1	694	2	YES	plus/plus	100%
Solyc00g047200.1.1\ Solyc11g056490.1.1	234	1	YES	plus/plus	100%
Solyc00g058890.1.1\ Solyc12g010550.1.1	411	1	YES	plus/plus	100%
Solyc01g007440.1.1\ Solyc09g064400.1.1	495	2	YES(2 mismatches)	plus/plus	100%
Solyc01g007450.1.1\ Solyc09g064410.1.1	201	1	YES(1 mismatch)	plus/plus	100%
Solyc01g106220.2.1\ Solyc01g106240.2.1	8922\8923	8	YES(3 gaps)	plus/plus	100%
Solyc03g091030.1.1\ Solyc03g091040.1.1	330	1	YES	plus/plus	100%
Solyc03g116300.1.1\ Solyc03g116310.1.1	303	1	YES	plus/plus	100%
Solyc03g093100.1.1\ Solyc08g016290.1.1	2491	4	YES(2 mismatches)	plus/plus	100%
Solyc03g120400.1.1\ Solyc05g012960.1.1\ Solyc09g014290.1.1	174	1	YES(2 mismatches)	plus/plus	100%
Solyc08g079210.1.1\ Solyc08g079220.1.1	360	1	YES	plus/plus	100%
Solyc10g008370.2.1\ Solyc10g008380.2.1	722	2	YES	plus/plus	100%
Solyc10g012380.1.1\ Solyc10g012390.1.1	450	1	YES	plus/plus	94%(517/550)
Solyc12g009730.1.1\ Solyc12g009750.1.1	2761	2	YES	plus/plus	100%
Solyc12g010370.1.1\ Solyc12g010760.1.1	1366	3	YES(1 mismatch)	plus/plus	100%

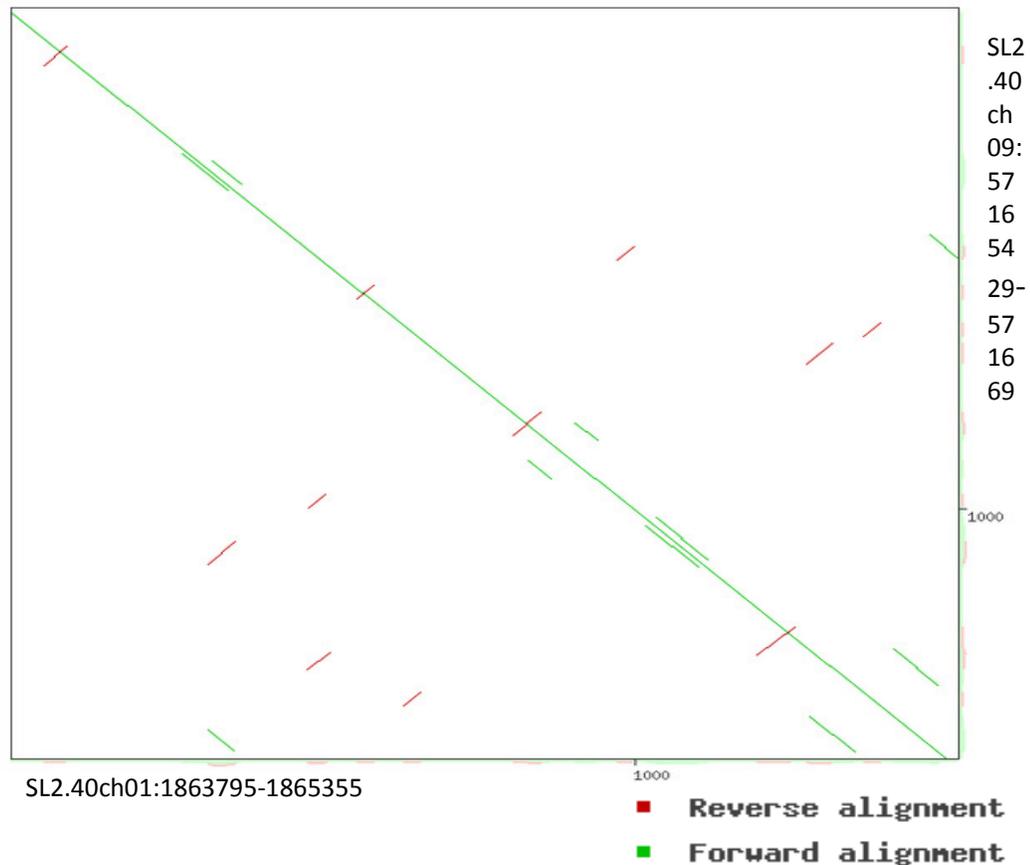


Figure 18 Dotplot of the region on chromosome 1 from 1863795 to 1865355 and of the region on chromosome 9 from 57165429 to 5716989.

3.6 iTAG annotation versus RefSeq annotation

The iTAG annotation is not available on the NCBI website, where the tomato genes can be exploited only with the RefSeq annotation. Despite the fact that the iTAG annotation is considered the official one and it is the most used into the scientific community, RefSeq annotation is as well exploited and it is a reference for tomato. For this reason we compared the two annotation in order to have a more comprehensive view of the tomato genes.

The total number of annotated gene in RefSeq was less than the iTAG one: 26628 genes, and also in this case alternative transcripts were not predicted.

Out of the total number of genes, only 1058 RefSeq annotated genes were identical to the iTAG ones though 22784 RefSeq genes overlapped at least one iTAG gene, meanwhile 2786 RefSeq genes not overlapped an iTAG locus. Analyzing the results of the comparison from the iTAG point of view, beyond the 1058 gene identical to RefSeq, 25049 genes overlapped at least one RefSeq locus and 8620 genes were not overlap any RefSeq locus (Fig. 19).

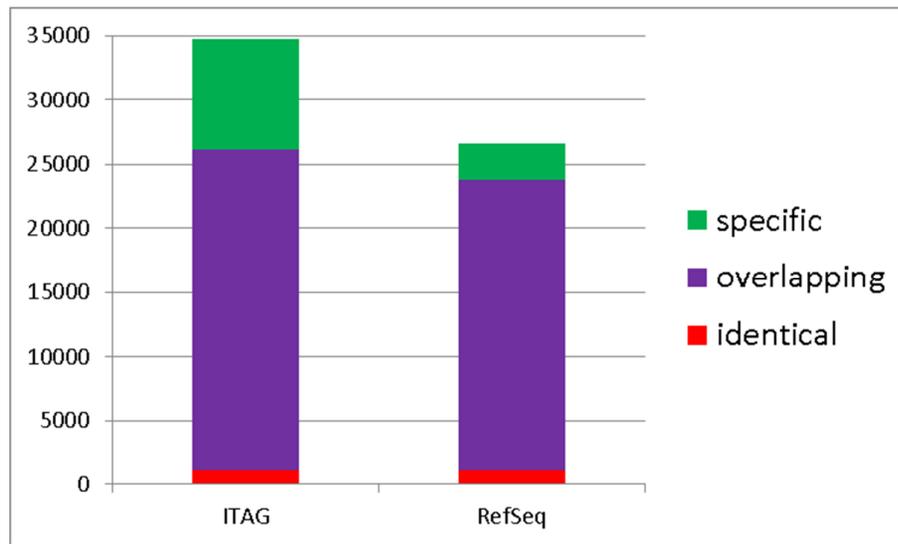


Figure 19 Total number of genes per iTAG and RefSeq, reporting the number of identical genes in the two annotations (red), the number of overlapping genes between the two annotations (purple) and the number of gene annotation specific (green)

When iTAG and RefSeq genes were in the same locus, it could happen that: i) to one iTAG locus corresponded one RefSeq locus (Fig. 20.A), ii) to two iTAG loci corresponded only one RefSeq locus (Fig. 20.B), iii) to one iTAG locus corresponded two RefSeq loci.

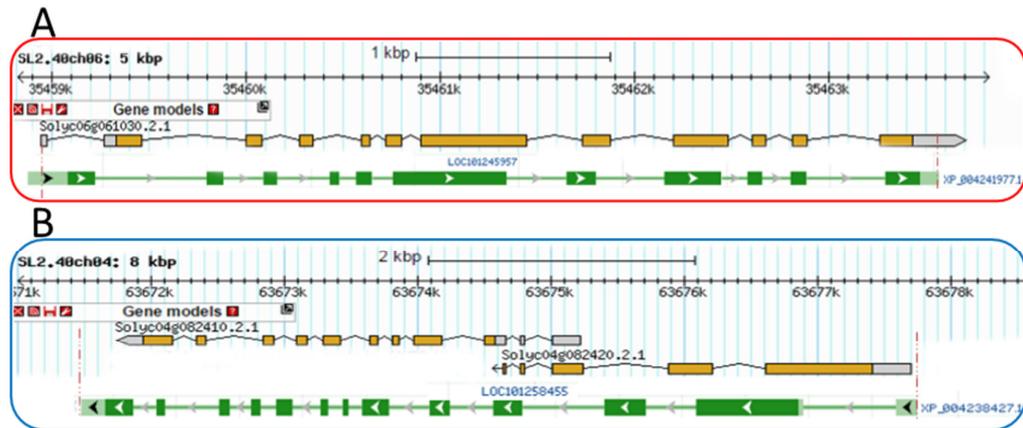


Figure 20 Example of overlapping locus between iTAG (orange) and RefSeq (green) annotations. In A is reported the overlapping between one iTAG gene and one RefSeq gene; in B is reported the overlapping among two iTAG genes and one RefSeq gene

3.7 “Guide” to the tomato annotation

In order to alert users to all the problems that are in the current annotations available for the tomato genome and make easier its exploitation, we set up a “guide” to how read them.

The guide can be exploited from the iTAG or RefSeq point of view and it give all the information resulted from the analyses cited above.

In the first part of the iTAG preferred annotation, after the general information given by the canonic annotation, information were added about the obsolete genes in 2.3 and 2.4 versions and the overlapping with other predicted genes (Fig. 21). In the overlapping field (OV), it is described the number of total overlapping and the very long genes that overlap more than 48 genes were also highlighted.

Afterwards, information about the remapping mRNA were provided (Fig. 22). In this field, three column can be exploited:

- 1) Remapping flag (REM):
 - a. COM: Confirmed One Match. mRNA that remapped only in the predicted locus,
 - b. NCOM: Not Confirmed One Match. mRNA that remapped only one time but not in the exact predicted locus;
 - c. CMM: Confirmed with More Matches. mRNA that remapped in the annotated locus and in other regions, number of remapping is also specified;
 - d. NCC: Not Confirmed with More Matches. mRNA that remapped not in the exact predicted locus but only in other regions, number of remapping is also specified;
 - e. NF: Not Found. mRNA not mapped on the genome.
- 2) Overlap mRNA (OM) and 3) Overlap Locus (OL), where there are underlined the mRNA that are mapped on predicted mRNA or locus:
 - a. ISS: Identical Same Strand. mRNA identical with 100% coverage and identity with other mRNA/locus, on same strand;
 - b. SSS: Similar Same Strand. mRNA with other mRNA/locus, on same strand;
 - c. ICS: Identical Complement Strand. mRNA identical with 100% coverage and identity with other mRNA/locus, on complementary strand;
 - d. SCS: Similar Complement Strand. mRNA similar with other mRNA/locus, on complementary strand;
 - e. InISS: Included Identical Same Strand. mRNA included with coverage 100% identity 100% in other mRNA/locus, on same strand;
 - f. InSSS: Included Similar Same Strand. mRNA included in other mRNA/locus, on same strand;

- g. InICS: Included Identical Complement Strand. mRNA included with 100% coverage and identity in other mRNA/locus, on complementary strand;
- h. InSCS: Included Similar Complement Strand. mRNA included in other mRNA/locus, on complementary strand.

After the remapping information, in the guideline there were information about the encoding Protein Validation (PV) (Fig. 23), in which three classes are shown:

- 1) AC: Annotation Confirmed. Annotated genes that have match with proteins or unknown genes that don't have match with a protein;
- 2) PAA: Protein Annotation Added. Unknown genes that have a match with a protein and a functional annotation can be added;
- 3) PAQ: Protein Annotation Questioned. Annotated genes that don't have a match with a protein.

iTAG ANNOTATION vers. 2.3								
GENE ID	CHR	START	END	Function	STRAND	# Exons	OBSOLETE (OB)	OVERLAPPING iTAG (OV)
Solyd01.g091060.1.1	SL2.40ch01	76546392	76549258	Pectinesterase (AHRD v	+	4		OV_Ov1VLG:overlapping1 very long gene
Solyd01.g091070.2.1	SL2.40ch01	76550313	76559336	Methionine aminopept	+	19		OV_Ov1VLG:overlapping1 very long gene
Solyd01.g091080.2.1	SL2.40ch01	76559722	76569026	Tubulin-specific chaper	-	14		OV_Ov1+1VLG:overlapping1 iTAG + 1 very long gene
Solyd01.g091090.2.1	SL2.40ch01	76567701	76573808	Beta-tubulin cofactor-li	-	7		OV_Ov1+1VLG:overlapping1 iTAG + 1 very long gene
Solyd01.g091100.2.1	SL2.40ch01	76575646	76576533	Pectinesterase (AHRD v	-	1		OV_Ov1VLG:overlapping1 very long gene
Solyd01.g091110.2.1	SL2.40ch01	76583261	76584010	Pectinesterase/pectine	-	1		OV_Ov1VLG:overlapping1 very long gene
Solyd01.g091120.2.1	SL2.40ch01	76585783	76587387	Os03g0825600 protein (I	-	1		OV_Ov1VLG:overlapping1 very long gene
Solyd01.g091130.2.1	SL2.40ch01	76596420	76600448	Nitroreductase (AHRD v	-	4		OV_Ov1+1VLG:overlapping1 iTAG + 1 very long gene
Solyd01.g091140.2.1	SL2.40ch01	76596747	76604991	Nitroreductase (AHRD v	-	6		OV_Ov1+1VLG:overlapping1 iTAG + 1 very long gene
Solyd01.g091160.2.1	SL2.40ch01	76612669	76617028	Agmatinase (AHRD V1 *	+	7		
Solyd01.g091170.2.1	SL2.40ch01	76623898	76626116	Agmatinase (AHRD V1 *	+	7		

Figure 21 Information about iTAG version 2.3 stored in the annotations guide. Gene ID, Chromosome, Start position, End position, Functional annotation, Strand, Number of exons, Genes obsolete in 2.4 and 2.3 versions, Genes overlapping other genes are reported

REMAPPING		
REMAPPING (REM)	overlap mRNA (OM) (coverage 100%; identity >=98%)	overlap locus (OL) (coverage 100%; identity >=98%)
REM_COM: Confirmed One Match		
REM_COM: Confirmed One Match		
REM_COM: Confirmed One Match		
REM_CMM2: Confirmed with More Matches	OM_SCS: Similar Complement Strand with Solyc02g064910.1.1	
REM_CMM2: Confirmed with More Matches		
REM_CMM2: Confirmed with More Matches		
REM_CMM2: Confirmed with More Matches	OM_InSCS: Included Similar Complement Strand with Solyc02g064920.1.1	OL_InSCS: Included Similar Complement Strand with Solyc02g064920.1.1
REM_CMM2: Confirmed with More Matches		OL_SCS: Similar Complement Strand with Solyc02g064910.1.1
REM_CMM2: Confirmed with More Matches	OM_SCS: Similar Complement Strand with Solyc02g064860.1.1	OL_SCS: Similar Complement Strand with Solyc02g064900.1.1
REM_CMM2: Confirmed with More Matches		
REM_COM: Confirmed One Match		

Figure 22 Information about Remapping of the mRNA on the tomato genome stored in the annotations guide. In the first column is reported the flag of the remapping and, where presents, the number of duplications. In the second and third column is reported if the gene overlap (with coverage = 100% and identity \geq 98%) another mRNA or a locus, respectively

Protein validation	
encoding Protein Validation (code: PV): tBLASTn e10-3	Putative To Be Joined (PTBJ)
PV_AC: Annotation Confirmed	
PV_AC: Annotation Confirmed	PTBJ_Putative To Be Joined with Solyc06g006020.1.1
PV_AC: Annotation Confirmed	PTBJ_Putative To Be Joined with Solyc06g006030.1.1 and Solyc06g006010.1.1
PV_AC: Annotation Confirmed	PTBJ_Putative To Be Joined with Solyc06g006040.1.1 and Solyc06g006020.1.1
PV_AC: Annotation Confirmed	PTBJ_Putative To Be Joined with Solyc06g006050.1.1 and Solyc06g006030.1.1
PV_AC: Annotation Confirmed	PTBJ_Putative To Be Joined with Solyc06g006040.1.1
PV_AC: Annotation Confirmed	
PV_AC: Annotation Confirmed	
PV_PAA: Protein Annotation Added	
PV_AC: Annotation Confirmed	

Figure 23 Information about Protein validation is shown. In the first column is specified if the gene has a match with a protein, in the second column it is reported if the gene is a putative split

Finally, the results about the comparison between iTAG and RefSeq loci were reported (RefCom) (Fig. 24). In this field, six classes are exploited:

- 1) IS: Identical Structure. Locus and exons starts and ends are identical in iTAG and RefSeq;
- 2) DS: Different Structure. iTAG and RefSeq locus are identical but exons starts and ends are not the same;
- 3) PO: partial overlapping. iTAG and RefSeq locus are different but overlapping;
- 4) OMR: Overlapping More Refseq. One iTAG gene overlaps more RefSeq genes;
- 5) OSR: Overlapping Same Refseq. Two or more iTAG genes overlap the same RefSeq;
- 6) NR: No RefSeq. iTAG gene is not overlap any RefSeq gene.

In all the classes listed below, the RefSeq identical or overlap with the iTAG gene is also reported together with them functional annotation.

The guide set up can be read also to exploit the tomato annotation from the RefSeq point of view, with the information about the comparison with iTAG locus (ItagCom). Also in this case, the classes reported are six and all the iTAG genes identical or overlapping with the RefSeq ones were specified (Fig. 25).

REFSEQ	
RefSeq Comparison (RefCom)	
PO: Partial Overlapping with 101266519: alternative NAD(P)H dehydrogenase 2, mitochondrial-like	
PO: Partial Overlapping with 101266827: uncharacterized LOC101266827	
PO: Partial Overlapping with 543875: asparagine synthetase	
PO: Partial Overlapping with 101267711: probable protein phosphatase 2C 2-like	
PO: Partial Overlapping with 101267411: proliferation-associated protein 2G4-like	
OSR: Overlapping Same RefSeq with 101267998: 60S ribosomal protein L41-like	
OSR: Overlapping Same RefSeq with 101267998: 60S ribosomal protein L41-like	
NR: No RefSeq	
PO: Partial Overlapping with 101243790: BAG family molecular chaperone regulator 4-like	
PO: Partial Overlapping with 101244083: LRR repeats and ubiquitin-like domain-containing protein At2g30105-like	
PO: Partial Overlapping with 101244378: uncharacterized LOC101244378	
OSR: Overlapping Same RefSeq with 101244588: putative pentatricopeptide repeat-containing protein At1g12700, mitochondrial-like	
OSR: Overlapping Same RefSeq with 101244588: putative pentatricopeptide repeat-containing protein At1g12700, mitochondrial-like	
NR: No RefSeq	
PO: Partial Overlapping with 101244875: putative pentatricopeptide repeat-containing protein At1g12700, mitochondrial-like	
PO: Partial Overlapping with 101245163: putative deoxyribonuclease TATDN1-like	
PO: Partial Overlapping with 101245462: ubiquitin-activating enzyme E11-like	
PO: Partial Overlapping with 101246059: GILT-like protein F37H8.5-like	
PO: Partial Overlapping with 101245763: GILT-like protein F37H8.5-like	
PO: Partial Overlapping with 101246535: probable protein phosphatase 2C 26-like	
PO: Partial Overlapping with 101246828: 40S ribosomal protein S4-like	
PO: Partial Overlapping with 101247128: uncharacterized LOC101247128	
PO: Partial Overlapping with 101260538: uncharacterized LOC101260538	
PO: Partial Overlapping with 101260823: LOB domain-containing protein 13-like	
NR: No RefSeq	
NR: No RefSeq	
PO: Partial Overlapping with 101247430: CBL-interacting protein kinase 2-like	
PO: Partial Overlapping with 101247723: CBL-interacting serine/threonine-protein kinase 11-like	
NR: No RefSeq	
PO: Partial Overlapping with 101248013: EPIDERMAL PATTERNING FACTOR-like protein 6-like	

Figure 24 Information about comparison with *iTAG* versus RefSeq annotations

iTAG	
iTAG comparison (ItagCom)	
ItagCom_PO:Partial Overlapping with: Solyc01g005000.2.1:Glutamate decarboxylase (AHRD V1***- Q11D8_CIT3I);3B contains Interpro domain(s) IPR010107 Glutamate decarboxylase ;Ontology_term	
ItagCom_PO:Partial Overlapping with: Solyc01g005010.2.1:Protein XRI1 (AHRD V1***- XRI1_ARATH);Parent	
ItagCom_PO:Partial Overlapping with: Solyc01g005020.2.1:Sodium/hydrogen exchanger Na+H+ antiporter (AHRD V1***- Q8IET0_PLAF7);3B contains Interpro domain(s) IPR018418 Na+/H+ exchanger;2C	
ItagCom_OSI:Overlapping Same iTAG with: Solyc01g005030.2.1:Serine/threonine-protein kinase 36 (AHRD V1***- STK36_HUMAN);3B contains Interpro domain(s) IPR004240 Nonaspanin (TM9SF);Ontolo	
ItagCom_OSI:Overlapping Same iTAG with: Solyc01g005030.2.1:Serine/threonine-protein kinase 36 (AHRD V1***- STK36_HUMAN);3B contains Interpro domain(s) IPR004240 Nonaspanin (TM9SF);Ontolo	
ItagCom_PO:Partial Overlapping with: Solyc01g005040.2.1:Unknown Protein (AHRD V1);3B contains Interpro domain(s) IPR006702 Uncharacterised protein family UPF0497;2C trans-membrane plant;Pare	
ItagCom_PO:Partial Overlapping with: Solyc01g005050.2.1:Unknown Protein (AHRD V1);Parent	
ItagCom_PO:Partial Overlapping with: Solyc01g005060.2.1:Zinc finger protein (AHRD V1***- Q764N7_MALDO);3B contains Interpro domain(s) IPR007087 Zinc finger;2C C2H2-type ;Ontology_term	
ItagCom_PO:Partial Overlapping with: Solyc01g005070.2.1:Clathrin assembly protein (AHRD V1***- B6TYB1_MAIZE);3B contains Interpro domain(s) IPR011417 ANTH;Ontology_term	
ItagCom_PO:Partial Overlapping with: Solyc01g005080.2.1:Microtubule-associated protein MAP65-1a (AHRD V1**** CQJZ7_ORYNI);3B contains Interpro domain(s) IPR007145 MAP65/ASE1;Ontology_term	
ItagCom_PO:Partial Overlapping with: Solyc01g005090.2.1:Inositol 14 5-trisphosphate 5-phosphatase-like protein (AHRD V1**** Q6H453_ORYSJ);3B contains Interpro domain(s) IPR000300 Inositol polyph	
ItagCom_PO:Partial Overlapping with: Solyc01g005100.2.1:Prolyl 3-hydroxylase 1 (AHRD V1***- P3H1_CHICK);3B contains Interpro domain(s) IPR006620 Prolyl 4-hydroxylase;2C alpha subunit ;Ontology_t	
ItagCom_PO:Partial Overlapping with: Solyc01g005110.2.1:cDNA clone J065210E11 full insert sequence (AHRD V1***- B7F922_ORYSJ);3B contains Interpro domain(s) IPR013174 Dolichol-phosphate manna	
ItagCom_PO:Partial Overlapping with: Solyc01g005120.2.1:Xyloglucan endotransglucosylase/hydrolase 13 (AHRD V1**** C0IRH2_ACTOE);3B contains Interpro domain(s) IPR016455 Xyloglucan endotransgl	
ItagCom_PO:Partial Overlapping with: Solyc01g005130.2.1:Zinc finger protein 7 (AHRD V1***- B6U8J3_MAIZE);3B contains Interpro domain(s) IPR007087 Zinc finger;2C C2H2-type ;Ontology_term	
ItagCom_PO:Partial Overlapping with: Solyc01g005140.2.1:Cytochrome b561 (AHRD V1***- Q3LGX5_CITLA);3B contains Interpro domain(s) IPR006593 Cytochrome b561/ferric reductase transmembrane ;C	
ItagCom_PO:Partial Overlapping with: Solyc01g005150.2.1:Cytochrome b561 (AHRD V1***- Q3LGX5_CITLA);3B contains Interpro domain(s) IPR004877 Cytochrome b561;2C eukaryote ;Ontology_term	
ItagCom_PO:Partial Overlapping with: Solyc01g005160.2.1:U-box domain-containing protein (AHRD V1***- D7MID4_ARALY);3B contains Interpro domain(s) IPR003613 U box domain;Ontology_term	
ItagCom_PO:Partial Overlapping with: Solyc01g005170.1.1:Clathrin assembly protein (AHRD V1***- B6TW95_MAIZE);3B contains Interpro domain(s) IPR011417 ANTH;Ontology_term	
ItagCom_ID:Identical Structure with: Solyc01g005190.1.1:Zinc finger family protein (AHRD V1***- D7M6E6_ARALY);3B contains Interpro domain(s) IPR007087 Zinc finger;2C C2H2-type ;Ontology_term	
ItagCom_ID:Identical Structure with: Solyc01g005200.1.1:Zinc finger family protein (AHRD V1***- D7M6E6_ARALY);3B contains Interpro domain(s) IPR007087 Zinc finger;2C C2H2-type ;Ontology_term	
ItagCom_PO:Partial Overlapping with: Solyc01g005210.2.1:Alpha alpha-trehalose-phosphate synthase (UDP-forming) (AHRD V1**** A3CXN9_METMJ);3B contains Interpro domain(s) IPR001830 Glycosyl tr	
ItagCom_PO:Partial Overlapping with: Solyc01g005220.2.1:Necrotic spotted lesions 1(Fragment) (AHRD V1***- B6ZA18_HELAN);3B contains Interpro domain(s) IPR001862 Membrane attack complex comp	
ItagCom_PO:Partial Overlapping with: Solyc01g005230.2.1:S-adenosyl-L-methionine salicylic acid carboxyl methyltransferase (AHRD V1**** A7XZE9_9MAGN);3B contains Interpro domain(s) IPR005299 S	
ItagCom_PO:Partial Overlapping with: Solyc01g005240.2.1:Aspartokinase (AHRD V1***- B9RGY9_RICCO);3B contains Interpro domain(s) IPR001341 Aspartate kinase region ;Ontology_term	
ItagCom_PO:Partial Overlapping with: Solyc01g005250.2.1:Aspartate-semialdehyde dehydrogenase (AHRD V1***- B6TWwL_MAIZE);3B contains Interpro domain(s) IPR005986 Aspartate-semialdehyde de	
ItagCom_PO:Partial Overlapping with: Solyc01g005260.2.1:SEC14 cytosolic factor family protein (AHRD V1***- D7M8H6_ARALY);3B contains Interpro domain(s) IPR001251 Cellular retinaldehyde-binding/trip	
ItagCom_PO:Partial Overlapping with: Solyc01g005270.2.1:SEC14 cytosolic factor family protein (AHRD V1***- D7M8H6_ARALY);3B contains Interpro domain(s) IPR001251 Cellular retinaldehyde-binding/trip	
ItagCom_PO:Partial Overlapping with: Solyc01g005280.2.1:SEC14 cytosolic factor (AHRD V1***- B2W8N2_PYRTR);3B contains Interpro domain(s) IPR001251 Cellular retinaldehyde-binding/triple function;:	
ItagCom_PO:Partial Overlapping with: Solyc01g005290.2.1:SEC14 cytosolic factor family protein (AHRD V1***- D7M8H6_ARALY);3B contains Interpro domain(s) IPR001251 Cellular retinaldehyde-binding/trip	
ItagCom_PO:Partial Overlapping with: Solyc01g005300.2.1:Flavin-binding kelch domain F box protein (AHRD V1***- D7KVSS_ARALY);3B contains Interpro domain(s) IPR015915 Kelch-type beta propeller ;C	
ItagCom_PO:Partial Overlapping with: Solyc01g005310.2.1:Dynammin like protein (AHRD V1***- D3BTL7_POLPA);3B contains Interpro domain(s) IPR001401 Dynammin;2C GTPase region ;Ontology_term	

Figure 25 Information about comparison with RefSeq versus iTAG annotations

3.8 *Arabidopsis thaliana* microarray resources

Our survey on the available omics resources for plants was focused also on *Arabidopsis thaliana*, sequenced in 2000 (Arabidopsis Genome Initiative 2000) and considered the model organism for plants.

In this case, we made an overview of the results obtained for all the co-expression platforms of this species, summarizing the common features related to gene co-expression analysis, and specifically the correlation method, the normalization approach used and the dataset accessible (Tab. 12).

Table 12 List of the web based co-expression analysis databases offering resources including *Arabidopsis* related facilities. Release data, number of slides and normalization method are also shown.

Resource	Website	Release Data	Number of Slides	Normalization method
ATCOECIS	http://bioinformatics.psb.ugent.be/ATCOECIS	2009	322	RMA
ATTED II	http://atted.jp/	2007	11171	RMA
BAR	http://bar.utoronto.ca/welcome.htm	2005	405	MAS 5.0
COP	http://webs2.kazusa.or.jp/kagiana/cop0911/	2010	5272	MAS 5.0
CORNET	https://cornet.psb.ugent.be/	2009	NOT DEFINED	RMA
CRESS EXPRESS	http://cressexpress.org	2008	1799	RMA/MAS 5.0/GCRMA
CSB.DB	http://csbdb.mpimp-golm.mpg.de/csbdb/dbcor/ath.html	2004	NOT DEFINED	GCOS
GENECAT	http://genecat.mpg.de/cgi-bin/Ainitiator.py	2008	351	RMA
GENEMANIA	http://www.genemania.org/	2008	NOT DEFINED	NOT DEFINED
GENENVESTIGATOR	https://www.geneinvestigator.com/gv/	2004	9211	RMA/MAS 5.0
PLANET	http://aranet.mpimp-golm.mpg.de/	2011	1074	NOT DEFINED

We used the CESA7 gene (AT5G17420) to query each platform, the analyses were performed using the default settings proposed. Each result was ranked according to the specific correlation value proposed by each platform. We collected the top 20 co-expressed genes resulting from each analysis on each database. It is known that CESA7 is co-expressed with CESA4 (AT5G44030) and CESA8 (AT4G18780) in physiological conditions since this has been confirmed experimentally. The three genes code for single elements of a complex involved in the cell wall synthesis (Eckardt 2003). Another gene considered for this analysis is AT5G06680, implied in the gamma-tubulin complex. We collected the top 20 co-expressed genes resulting from each analysis on each database, using default parameters (Tab. 13.A, B and Tab. 14.A, B).

Using CESA7, despite the relevant differences in the dataset size, correlation and normalization methods proposed by each database, Genevestigator, Atted, Cop, Genecat, Bar and CressExpress share with all the other websites, about ~50% of their genes in the results and often, this value reaches or overcomes the 70% when considering couple comparisons, as it happens between Cop, Bar and Genecat, Csb.DB, Planet, Cornet and Genemania outputs instead, have less than 65% of elements shared with the results proposed by all the other databases. This can be explained by the fact that Genemania and Planet are not offering a specific ranking to list the co-expressed genes, but they are more focused on defining co-expressed gene modules. From a quality viewpoint, the presence of CESA 4 (AT5G44030) and CESA8 (AT4G18780) in the results of the CESA7 (AT5g17420) queries (Eckardt 2003) underlines the prediction skill of each database. As shown in the table 15, only Csb.DB and Planet seem to have some problems in the query results, but we have to specify that the first one does not show CESA8 because the probe of this gene was not included in the dataset exploited for this analysis, while Planet does not show CESA4 (AT5G44030) in the top 20,

despite it belongs to the cluster shown in its website result, simply because no rank has been proposed. Beyond these two particular databases, although in different rank positions, all the platforms confirm the co-expression of the CESA4-7-8 complex, and in the cases of Cornet, Genevestigator, Genemania, Genecat, CressExpress (RMA and gcRMA) and Atted, where the rank positions of their co-expressed genes have been clearly defined by their p-value correlation methods, CESA 4 and CESA 8 are listed in the first three positions, underlining the efficiency of these specific databases. Interestingly, collecting the top 20 co-expressed genes from each platform using AT5G06680 as query, there is not a database output very similar to another one as it happens for CESA7, and moreover the average of shared genes among the platform outputs does not exceed the 10% (Tab. 15). So, although using the same datasets and parameters, the similarity among the databases change totally when using CESA7 or AT5G06680, and the decreasing in the number of shared co-expressed genes can be very huge, as it happens between COP and BAR, where this value moves from 16 to 1. This underlines that the results proposed by the platforms must be compared among them since the common parameters developed to extract co-expressed gene lists can produce very different information.

So, one single answer from only one platform is not enough, since the co-expression profile of some genes may be very inflected by the conditions of the experiments used for the dataset building, as seen for AT5G06680, while this not happens for gene like CESA7, where the co-expression network shown in the queries is less variable, and probably depending from less conditions. In fact, despite some huge differences in the datasets size and experiments composition (i.e. passing from 11171 slides in Genevestigator to 351 of Genecat), CESA7 co-expression network remains confirmed among the platforms, while for AT5G06680 the co-expression profile may be harder to establish, due to a high modulating expression, or simply due to some

limits in the microarray signal detection. Beyond the dataset composition, normalization has a strong influence on the results too, as seen for AT5G06680 in CressExpress database using the dataset version 3.1, normalized with GCRMA, and the dataset version 3.2, normalized MAS5.0, where, despite the lacking of only 1 experiment out of 115 between the two versions during the analyses, there is only one gene shared by the two co-expression lists.

Table 13.A Complete list of the CESA7 query results as offered by each database. Statistical parameters exploited to describe the results are also reported

PLANET		ATTED		BAR		COP				CORNET			CRESS RMA					
AGI	Corr. Value	AGI	MR	AGI	r-value	AGI	VF	%ile	CC	AGI	r-value	p-value	AGI	p-value	slope	T	DOF	r2
AT4G18640	Not available	AT5G15630	1	AT5G15630	0.994	AT1G27380	0.98	97.80	0.860	AT4G18780	0.920	2:1.11E-36	AT4G18780	2.15E-290	0.45	44.74	1717	0.538
AT4G18780		AT5G44030	1.04	AT5G03170	0.989	AT4G27435	0.98	97.80	0.940	AT5G44030	0.920	2:2.05E-36	AT5G44030	2.74E-284	0.53	44.06	1717	0.531
AT1G63520		AT4G18780	2.05	AT3G50220	0.988	AT2G41610	0.95	97.00	0.900	AT5G60020	0.910	2:7.17E-35	AT5G54690	2.99E-281	0.51	43.72	1717	0.527
AT3G08490		AT5G54690	2.08	AT5G54690	0.988	AT3G16920	0.95	97.00	0.930	AT5G54690	0.890	2:8.19E-30	AT5G15630	3.75E-277	0.56	43.26	1717	0.522
AT3G27200		AT3G16920	3.02	AT3G18660	0.985	AT3G50220	0.95	97.00	0.910	AT5G60720	0.860	2:4.57E-26	AT3G16920	3.93E-275	0.39	43.04	1717	0.519
AT3G45870		AT3G18660	3.07	AT3G16920	0.980	AT4G28500	0.95	97.00	0.840	AT5G03170	0.860	2:2.01E-25	AT5G60020	1.25E-274	0.50	42.98	1717	0.518
AT1G12260		AT2G37090	4	AT1G27380	0.979	AT5G15630	0.95	97.00	0.970	AT5G01360	0.850	2:1.82E-23	AT5G60720	5.56E-274	0.77	42.91	1717	0.518
AT1G05310		AT2G38080	4.02	AT4G18780	0.977	AT5G03170	0.93	96.40	0.940	AT3G62020	0.830	2:3.62E-21	AT5G03170	7.80E-269	0.45	42.34	1717	0.511
AT1G24030		AT5G03170	4.05	AT5G44030	0.977	AT5G44030	0.93	96.40	0.940	AT5G15630	0.820	2:4.72E-20	AT1G27440	2.76E-265	0.76	41.94	1717	0.506
AT1G58070		AT1G27440	5.05	AT3G15050	0.973	AT1G22480	0.91	95.60	0.850	AT1G62990	0.810	2:3.16E-19	AT2G38080	4.33E-262	0.33	41.59	1717	0.502
AT3G52900		AT5G60020	5.07	AT1G07120	0.970	AT5G67210	0.91	95.60	0.850	AT1G54790	0.790	2:1.21E-16	AT4G27435	1.63E-255	0.54	40.86	1717	0.493
AT2G38080		AT5G60720	6	AT4G27435	0.969	AT3G15050	0.89	94.60	0.890	AT1G73640	0.790	2:1.92E-16	AT5G01360	3.99E-254	0.48	40.70	1717	0.491
AT5G45970		AT5G01360	6	AT2G38080	0.968	AT2G29130	0.88	94.00	0.840	AT3G50220	0.790	2:1.92E-16	AT1G32100	4.34E-254	0.52	40.70	1717	0.491
AT3G59690		AT1G79620	6.02	AT2G29130	0.966	AT1G32770	0.88	94.00	0.830	AT5G03260	0.780	2:1.44E-15	AT5G16600	1.28E-251	0.74	40.42	1717	0.488
AT1G33800		AT4G18640	6.08	AT1G63910	0.966	AT2G38080	0.88	94.00	0.930	AT5G47530	0.770	2:9.83E-15	AT2G37090	7.86E-251	0.48	40.34	1717	0.487
AT1G09440		AT3G62020	8.01	AT5G67210	0.965	AT5G54690	0.87	93.50	0.920	AT1G32100	0.770	2:1.21E-14	AT5G03260	1.26E-249	0.66	40.20	1717	0.485
AT5G03260		AT5G60490	8.02	AT1G22480	0.964	AT3G18660	0.86	93.10	0.880	AT2G03200	0.760	2:9.15E-14	AT2G29130	2.51E-248	0.63	40.06	1717	0.483
AT3G16920		AT4G28500	8.05	AT4G28500	0.962	AT1G09610	0.84	91.90	0.860	AT4G08160	0.760	2:2.43E-13	AT3G50220	6.21E-248	0.50	40.02	1717	0.483
AT5G51890		AT3G59690	9.07	AT1G08340	0.962	AT1G27440	0.78	88.60	0.890	AT1G58070	0.740	2:3.3E-12	AT3G62020	2.68E-246	0.45	39.83	1717	0.480
AT5G40020		AT4G27435	9.08	AT3G62020	0.961	AT4G18780	0.73	85.5	0.870	AT2G27740	0.740	2:4.73E-12	AT1G24030	4.63E-245	0.27	39.70	1717	0.479

Table 13.B Complete list of the CESA7 query results as offered by each database. Statistical parameters exploited to describe the results are also reported

CRESS GCRMA						CRESS MAS						CSB			GENE CAT		GENE MANIA		GENEVE STIGATOR	
AGI	p-value	slope	T	DOF	r2	AGI	p-value	slope	T	DOF	r2	AGI	spearman	p-value	AGI	r-value	AGI	weight	AGI	r-value
AT4G18780	0	0.95	62.04	1228	0.758	AT5G15630	0	0.82	74.95	1613	0.777	AT5G44030	0.908	0	AT5G15630	0.950	AT4G18780	0.102	AT5G44030	0.900
AT3G16920	0	0.74	57.84	1228	0.732	AT5G54690	0	0.85	74.69	1613	0.776	AT2G38080	0.901	0	AT5G44030	0.934	AT5G44030	0.100	AT5G15630	0.880
AT5G44030	6.66E-306	0.84	51.05	1228	0.680	AT5G44030	0	0.78	69.80	1613	0.751	AT5G15630	0.897	0	AT4G18780	0.933	AT5G03170	0.074	AT4G18780	0.880
AT2G38080	2.00E-285	0.63	48.20	1228	0.654	AT1G27440	0	1.18	54.91	1613	0.652	AT2G28760	0.876	0	AT5G54690	0.922	AT2G25540	0.073	AT5G54690	0.870
AT5G60020	2.24E-262	0.70	45.05	1228	0.623	AT2G37090	0	0.69	52.10	1613	0.627	AT5G03170	0.872	0	AT5G03170	0.918	AT3G16920	0.071	AT5G60020	0.830
AT3G62020	2.14E-242	0.94	42.36	1228	0.594	AT5G60720	0	0.80	51.87	1613	0.625	AT5G60720	0.837	0	AT3G16920	0.899	AT2G32540	0.069	AT2G37090	0.830
AT2G37090	8.93E-229	0.92	40.54	1228	0.573	AT4G27435	0	0.77	50.68	1613	0.614	AT5G54690	0.836	0	AT1G27440	0.895	AT2G32530	0.069	AT5G01360	0.820
AT5G03260	1.28E-224	0.80	39.99	1228	0.566	AT4G18780	1.25E-298	0.48	46.32	1613	0.571	AT1G47410	0.827	2.22E-16	AT5G60720	0.894	AT4G24010	0.069	AT5G60720	0.820
AT2G28760	6.49E-212	0.78	38.31	1228	0.545	AT5G03170	7.21E-285	0.66	44.73	1613	0.554	AT4G18640	0.817	4.44E-16	AT3G18660	0.891	AT2G32610	0.069	AT2G38080	0.810
AT5G54690	1.52E-205	0.88	37.47	1228	0.534	AT3G62020	1.11E-258	0.61	41.72	1613	0.519	AT1G32100	0.811	6.66E-16	AT2G38080	0.890	AT2G33100	0.069	AT3G16920	0.810
AT5G40020	9.58E-200	0.66	36.71	1228	0.523	AT5G60490	4.20E-258	0.70	41.66	1613	0.518	AT1G33800	0.801	3.11E-15	AT3G62020	0.877	AT1G32180	0.069	AT5G03170	0.790
AT4G08160	5.59E-192	0.76	35.69	1228	0.509	AT3G50220	8.85E-256	0.41	41.39	1613	0.515	AT5G59290	0.786	2.29E-14	AT4G28500	0.875	AT4G15290	0.069	AT1G27440	0.780
AT5G01360	1.33E-190	0.64	35.51	1228	0.507	AT5G01360	6.57E-254	0.63	41.18	1613	0.513	AT5G03260	0.778	6.68E-14	AT2G37090	0.872	AT4G15320	0.069	AT5G03260	0.770
AT1G32100	5.78E-184	0.55	34.64	1228	0.494	AT2G38080	1.02E-242	0.44	39.89	1613	0.497	AT1G27440	0.775	8.73E-14	AT2G41610	0.865	AT4G38190	0.069	AT3G50220	0.760
AT2G27740	5.26E-183	0.79	34.51	1228	0.493	AT5G47530	2.70E-231	0.65	38.58	1613	0.480	AT5G60490	0.762	4.11E-13	AT4G27435	0.863	AT4G23990	0.069	AT3G18660	0.750
AT5G15630	1.10E-178	1.10	33.94	1228	0.484	AT2G41610	6.66E-227	0.72	38.08	1613	0.474	AT5G67210	0.755	8.61E-13	AT5G60020	0.863	AT4G24000	0.069	AT4G08160	0.730
AT5G18970	4.33E-178	0.88	33.87	1228	0.483	AT5G16490	3.63E-224	0.76	37.77	1613	0.469	AT4G27435	0.742	3.42E-12	AT3G50220	0.858	AT5G60720	0.068	AT1G79620	0.720
AT3G18660	3.30E-177	1.19	33.75	1228	0.481	AT3G18660	1.20E-222	0.54	37.59	1613	0.467	AT5G14510	0.727	1.53E-11	AT1G09610	0.836	AT5G54690	0.062	AT5G40020	0.720
AT4G35350	2.10E-170	0.56	32.86	1228	0.468	AT1G73640	4.26E-218	0.64	37.07	1613	0.460	AT2G38320	0.668	2.30E-09	AT3G15050	0.831	AT2G37090	0.060	AT1G132100	0.710
AT4G27435	1.85E-168	0.73	32.60	1228	0.464	AT1G08340	1.82E-211	0.63	36.31	1613	0.450	AT1G20850	0.666	2.65E-09	AT1G27380	0.823	AT2G32620	0.005	AT1G08340	0.710

Results

Table 14.A Complete list of the AT5G06680 query results as offered by each database. Statistical parameters exploited to describe the results are also reported

PLANET	ATTED		BAR		COP				CORNET			CRESS RMA						CRESS GCRMA					
	AGI	MR	AGI	r-value	AGI	VF	%ile	CC	AGI	r-value	p-value	AGI	p-value	slope	T	DOF	r2	AGI	p-value	slope	T	DOF	r2
AT3G4369	AT1G62020	3	AT1G09820	0.801	AT1G55325	0.61	76.70	0.95	AT2G01210	0.75	2:2.94E-13	AT3G18524	2.15E-290	0.45	44.74	1717	0.538	AT4G11450	2.351E-261	0.88	44.91	1228	0.622
AT5G1330	AT2G21390	3	AT1G09290	0.8	AT1G12930	0.56	73.00	0.92	AT1G64450	0.73	2:3.82E-11	AT4G14970	2.74E-284	0.53	44.06	1717	0.531	AT2G35530	2.318E-252	0.91	43.69	1228	0.609
AT3G1800	AT4G20740	8.9	AT2G29190	0.788	AT2G25760	0.55	70.60	0.92	AT3G57830	0.7	2:2.46E-9	AT1G04050	2.99E-281	0.51	43.72	1717	0.527	AT3G21100	1.93E-233	0.84	41.16	1228	0.580
AT3G6185	AT4G09980	15.4	AT1G73820	0.784	AT5G58100	0.54	69.50	0.92	AT3G57860	0.7	2:2.86E-9	AT4G11450	3.75E-277	0.56	43.26	1717	0.522	AT1G55540	8.93E-218	0.99	39.09	1228	0.555
AT5G0564	AT2G38770	16.9	AT2G38770	0.784	AT2G35110	0.54	69.50	0.92	AT2G33560	0.7	2:2.86E-9	AT5G63960	3.93E-275	0.39	43.04	1717	0.519	AT1G14850	8.24E-209	1.00	37.90	1228	0.539
AT5G2460	AT1G65380	21.6	AT5G45790	0.78	AT3G06340	0.54	69.50	0.92	AT3G54080	0.7	2:4.49E-9	AT3G09730	1.25E-274	0.50	42.98	1717	0.518	AT3G23780	3.60E-205	0.83	37.42	1228	0.533
AT3G0718	AT1G55325	31	AT5G02850	0.779	AT1G27595	0.53	68.60	0.92	AT5G43020	0.69	2:8.11E-9	AT1G26370	5.56E-274	0.77	42.91	1717	0.518	AT5G12440	2.01E-203	0.69	37.19	1228	0.530
AT1G3406	AT5G18960	35.7	AT5G55040	0.777	AT5G51340	0.53	68.60	0.93	AT5G67200	0.69	2:1.45E-8	AT5G63950	7.80E-269	0.45	42.34	1717	0.511	AT3G20010	2.06E-200	0.68	36.80	1228	0.525
AT3G4931	AT4G02070	44.8	AT5G55660	0.773	AT5G38880	0.52	67.40	0.92	AT3G63290	0.69	2:2.23E-8	AT3G10390	2.76E-265	0.76	41.94	1717	0.506	AT2G23700	7.89E-197	0.76	36.33	1228	0.518
AT4G3522	AT1G26370	45.3	AT2G33500	0.772	AT3G45190	0.52	67.40	0.92	AT5G26850	0.69	2:2.23E-8	AT1G23380	4.33E-262	0.33	41.59	1717	0.502	AT3G19120	3.64E-196	0.82	36.24	1228	0.517
AT4G3912	AT3G06340	48.7	AT3G19120	0.768	AT5G15680	0.51	66.30	0.92	AT1G68640	0.69	2:2.58E-8	AT2G21800	1.63E-255	0.54	40.86	1717	0.493	AT5G40740	1.27E-193	0.82	35.91	1228	0.512
AT5G6434	AT4G24490	51.8	AT3G27520	0.765	AT1G63700	0.51	66.30	0.92	AT5G67270	0.68	2:4.53E-8	AT1G14850	3.99E-254	0.48	40.70	1717	0.491	AT1G73590	1.74E-192	0.49	35.76	1228	0.510
AT5G0981	AT3G63290	53.4	AT3G06340	0.752	AT3G43700	0.49	63.50	0.93	AT4G38660	0.68	2:4.53E-8	AT2G20300	4.34E-254	0.52	40.70	1717	0.491	AT2G39090	4.90E-192	0.83	35.70	1228	0.509
AT5G1896	AT5G66770	55.1	AT3G55320	0.746	AT1G26170	0.49	63.50	0.93	AT5G10020	0.68	2:5.2E-8	AT2G43990	1.28E-251	0.74	40.42	1717	0.488	AT1G06590	3.52E-190	0.75	35.45	1228	0.506
AT3G1731	AT1G55350	60.9	AT2G33610	0.746	AT1G04950	0.48	62.50	0.92	AT5G57590	0.68	2:7.88E-8	AT1G77720	7.86E-251	0.48	40.34	1717	0.487	AT3G63290	1.68E-189	1.03	35.36	1228	0.505
AT2G4659	AT3G18524	61.4	AT5G17410	0.744	AT1G27850	0.47	61.20	0.91	AT3G61250	0.67	2:1.56E-7	AT2G40070	1.26E-249	0.66	40.20	1717	0.485	AT5G63950	2.23E-187	0.67	35.09	1228	0.501
AT5G1305	AT3G20010	68.3	AT4G22140	0.742	AT1G34320	0.47	61.20	0.92	AT1G54180	0.67	2:1.78E-7	AT3G20020	2.51E-248	0.63	40.06	1717	0.483	AT1G48270	8.94E-187	1.24	35.01	1228	0.500
AT1G2389	AT5G10020	70.7	AT1G08610	0.741	AT4G32620	0.47	61.20	0.92	AT5G63920	0.67	2:2.66E-7	AT4G14290	6.21E-248	0.50	40.02	1717	0.483	AT2G40640	9.93E-185	0.83	34.74	1228	0.496
AT5G0411	AT4G20910	74.8	AT1G30460	0.741	AT5G27970	0.47	61.20	0.92	AT5G63960	0.67	2:3.47E-7	AT1G21740	2.68E-246	0.45	39.83	1717	0.480	AT2G47020	4.43E-184	1.01	34.65	1228	0.495
AT1G6757	AT3G01380	79.1	AT1G28420	0.74	AT3G15120	0.46	59.8	0.92	AT5G63950	0.66	2:5.86E-7	AT1G73590	4.63E-245	0.27	39.70	1717	0.479	AT2G25420	5.40E-183	0.78	34.51	1228	0.493

Table 14.B Complete list of the AT5G06680 query results as offered by each database. Statistical parameters exploited to describe the results are also reported

CRESS MAS						CSB		GENE CAT		GENE MANIA		GENEVE STIGATOR	
AGI	p-value	slope	T	DOF	r2	AGI	spearman	AGI	r-value	AGI	weight	AGI	r-value
AT1G55325	3.78E-141	0.68	28.03	1613	0.328	AT1G47670	0.647	AT1G26170	NoT AvAilAbe	AT5G17410	2.752	AT3G22780	0.55
AT4G33200	5.965E-141	0.61	28.00	1613	0.327	AT3G51050	0.628	AT5G05560		AT3G61650	1.297	AT1G73590	0.55
AT5G10020	2.565E-138	0.43	27.68	1613	0.322	AT1G30970	0.622	AT3G21100		AT5G05620	1.255	AT2G02560	0.55
AT2G40070	1.678E-137	0.71	27.58	1613	0.321	AT1G68550	0.617	AT5G13300		AT5G37830	0.530	AT5G17410	0.55
AT3G61240	4.4E-136	0.55	27.41	1613	0.318	AT1G69295	0.612	AT3G16620		AT1G20570	0.400	AT1G14850	0.53
AT1G73590	3.746E-129	0.29	26.55	1613	0.304	AT1G52150	0.611	AT2G16880		AT1G80260	0.400	AT5G10020	0.53
AT5G13300	9.417E-129	0.49	26.50	1613	0.303	AT1G80530	0.606	AT3G20020		AT3G43610	0.400	AT5G60690	0.53
AT3G12590	1.22E-128	0.70	26.48	1613	0.303	AT1G73590	0.581	AT5G18960		AT3G11520	0.378	AT5G15680	0.52
AT5G65700	1.02E-126	0.46	26.24	1613	0.299	AT5G22740	0.581	AT3G20010		AT2G13650	0.120	AT1G55350	0.52
AT3G58580	1.66E-117	0.71	25.08	1613	0.281	AT4G33210	0.577	AT1G14850		AT2G22425	0.096	AT3G18524	0.52
AT5G23550	2.49E-115	0.70	24.80	1613	0.276	AT2G25970	0.574	AT3G10390		AT1G79280	0.092	AT4G36180	0.52
AT1G65380	5.84E-115	0.59	24.76	1613	0.275	AT1G09960	0.570	AT4G33200		AT4G40042	0.081	AT1G55325	0.52
AT4G31430	3.80E-114	0.42	24.65	1613	0.274	AT3G54080	0.570	AT3G15970		AT1G69295	0.081	AT2G05120	0.51
AT1G53380	4.27E-114	0.51	24.65	1613	0.274	AT5G65700	0.570	AT1G77720		AT3G22590	0.071	AT5G67100	0.51
AT2G38770	1.93E-113	0.50	24.56	1613	0.272	AT1G52310	0.565	AT1G72560		AT5G35430	0.052	AT2G27040	0.51
AT2G47900	2.49E-113	0.56	24.55	1613	0.272	AT5G64390	0.565	AT2G18850		AT5G14720	0.052	AT1G61010	0.51
AT3G19540	6.34E-110	0.52	24.11	1613	0.265	AT5G44670	0.565	AT1G47230		AT3G27325	0.048	AT3G63130	0.51
AT1G07705	1.23E-108	0.74	23.94	1613	0.262	AT5G67630	0.565	AT1G10490		AT1G77720	0.047	AT3G23780	0.51
AT5G22740	1.68E-107	0.34	23.79	1613	0.260	AT3G58040	0.563	AT1G65380		AT5G18960	0.040	AT2G28380	0.51
AT1G79650	3.66E-106	0.50	23.62	1613	0.257	AT4G15900	0.563	AT1G16190		AT3G53760	0.016	AT2G44830	0.51

Table 15 Summarizing comparison of the databases, querying CESA7 gene (light green boxes) and AT5G06680 gene (light orange boxes), checking top 20 co-expressed genes. Average of shared genes among the databases, excluding the self-matching value (yellow boxes) is specified.

	ATTED	BAR	COP	CORNET	CRESS RMA	CRESS GCRMA	CRESS MAS	CSB.DB	GENECAT	GENEMANIA	GENEVESTIGATOR	PLANET	average of shared genes among the databases	AT5G06680
ATTED	20	2	2	2	2	2	4	0	3	1	4	0	2	
BAR	11	20	1	0	0	1	1	0	0	1	1	0	0,64	
COP	11	16	20	0	0	0	1	0	1	0	2	0	0,64	
CORNET	9	7	6	20	2	2	1	1	0	0	1	0	0,82	
CRESS RMA	14	11	11	12	20	4	2	1	4	1	3	0	1,73	
CRESS GCRMA	12	9	8	11	13	20	1	1	3	0	3	0	1,55	
CRESS MAS	14	11	11	11	13	10	20	3	3	0	3	1	1,82	
CSB.DB	10	7	8	7	10	8	9	20	0	1	1	0	0,73	
GENECAT	15	14	16	9	14	11	14	8	20	2	1	2	1,73	
GENEMANIA	4	2	2	2	4	3	3	4	4	20	1	1	0,73	
GENEVESTIGATOR	14	10	10	11	14	13	13	8	13	4	20	0	1,82	
PLANET	3	3	3	3	5	5	2	3	3	1	5	20	0,36	
average of shared genes among the databases	11,7	9,80	9,9	8,5	11,6	9,8	10,9	7,9	11,8	3,20	11	3,27		
CESA 7 (AT5G17420)														

NORMALIZATION: ■ RMA ■ MAS 5.0 ■ GCRMA ■ GCOS

3.8.1 Consequences of mutant inclusion

In order to assess the effect of the presence of heterogeneous samples in a dataset, we evaluated the possible consequences of mutant inclusion (mut+)/exclusion (mut-) from a dataset of experiments in physiological conditions. A dataset obtained collecting 63 experiments in physiological condition from Nascarrays was exploited to calculate the Pearson's correlations among each gene-pair. Similarly, 16 mutants involving experiments were added to the dataset organizing a collection of 79 experiments. Hence, we purposely selected two pools of genes as case examples extracted from the two described datasets, including the genes most and least affected by co-expression instability, respectively. For each gene pool, 200 genes were tested for pair wise co-expression (39800 gene pairs) as measured by Pearson's correlation coefficient (r) and associated P-value, based on the two sets of samples (mut+) or (mut-). In a preliminary analysis on both stable and unstable gene pools, we extensively tested the existence of a relationship between the frequency of gene co-expression and presence/absence of mutants in the dataset. A Chi-square testing for independence of gene co-expression and mutant inclusion/exclusion showed a clear pattern of interdependence between the two variables. Significant differences among observed and expected occurrences of co-expressed and not co-expressed gene pairs, with or without mutants, were observed not only for all data pooled but also for most of the 200 tested single genes both for stable and unstable gene pools. In particular, in the case of unstable genes the Chi-square test resulted highly significant ($P < 0.001$) in 114 cases, significant ($0.001 < P < 0.05$) in 30 cases, and not significant ($P > 0.05$) in 31 cases. In 25 cases the tested gene was not co-expressed in neither dataset (i.e. occurrences of co-expression equal to zero both for mutant inclusion and exclusion), hence the Chi-squared did not apply. Chi-square results for stable genes were similar to unstable pool, with 143 highly significant, 32

significant, and 19 not significant values. Six genes were always co-expressed in both datasets (i.e. occurrences of non-co-expression equal to zero for both mutant inclusion and exclusion), hence the Chi-squared did not apply.

In the case of unstable gene pools, considering the significance of mutant-related effects, 14% of the tested gene pairs (5300 out of 39800) were significantly affected by mutant inclusion in the dataset. All single types of significant effects were relevant, being observed with significant occurrence among the tested genes. However, important differences among the types of effect were recorded. Inhibition of co-expression highly prevailed, with 1622 and 784 total cases of positive and negative correlations (i.e. 30.6% and 14.8% of all the significant observed effects) becoming not significant after exclusion of mutants from the dataset. Significant changes of magnitude in gene co-expression were also frequently observed, mostly in the case of positive correlations (1434 cases, corresponding to 27.1% of all the significant effects), but not for negative and not significant correlations (140 and 236 cases, respectively, corresponding to 2.6% and 4.5% of all the significant effects). Induction of gene co-expression after mutants exclusion, i.e. non-significant correlations turning into significant values, either positive or negative, were relatively frequently observed (512 and 492 cases, corresponding to 2.3% and 2.2% of all the significant effects for positive and negative correlations). Co-expression inversion after mutant exclusion, i.e. positive correlation turning into negative correlation and vice versa, was also recorded, although very rarely, with 70 (1.3% of all the significant effects) and 10 (0.2%) cases, respectively. In the case of stable genes only 0.04% of the tested gene pairs (14 out of 39800) were significantly affected by mutant inclusion in the dataset. Among the single types of significant effects, co-expression inhibition, induction, and inversion did not occur. Changes of magnitude in gene co-expression were the only significant effects observed,

with a small, not significant occurrence, all corresponding to positive correlations in both mut+ and mut- datasets.

Finally, in order to verify the perturbation of the mutants in a dataset, we used AT1G01290 and AT1G20580, classified as unstable and stable genes from previous analysis, to query each of the databases for gene co-expression in *Arabidopsis* (Tab. 16). Comparing the results, summarized in table 8, is evident that the co-expressed genes shared among all the databases used are few for both the genes in query, highlighting that despite AT1G20580 was considered a stable gene that is immune to the presence or not of mutants in the dataset, it suffers other kind of factors, such as the different datasets exploited by each resources and the normalization method.

Table 16 Summarizing comparison of the databases, querying AT1G01290 gene (grey boxes) and AT1G20580 gene (light blue boxes), checking top 20 co-expressed genes. In the table, is specified the average of shared genes among the databases, excluding the self matching value (yellow boxes).

	ATTED	BAR	COP	CORNET	CRESS RMA	CRESS GCRMA	CRESS MAS	CSB.DB	GENEMANIA	GENEVESTIGATOR	PLANET	average of shared genes among the databases	AT1G20580 (stable)
ATTED	20	2	3	2	4	4	6	1	2	2	1	2,7	
BAR	0	20	1	2	1	2	2	0	1	1	1	1,3	
COP	3	1	20	1	1	1	1	0	1	3	0	1,2	
CORNET	3	0	0	20	2	1	3	0	1	0	0	1,2	
CRESS RMA	5	0	1	0	20	4	5	1	1	1	1	2,1	
CRESS GCRMA	4	0	3	0	9	20	7	1	1	1	2	2,4	
CRESS MAS	2	1	2	0	6	8	20	1	1	1	1	2,8	
CSB.DB	1	2	2	0	3	2	2	20	0	1	1	0,6	
GENEMANIA	0	0	1	0	0	0	0	0	20	2	0	1	
GENEVESTIGATOR	5	0	3	1	3	3	2	1	0	20	2	1,4	
PLANET	2	0	1	2	0	0	0	0	1	1	20	0,9	
average of shared genes among the databases	2,5	0,40	1,7	0,6	2,7	2,9	2,3	1,3	0,20	1,9	0,7		
AT1G01290 (unstable)													

NORMALIZATION: ■ RMA ■ MAS 5.0 ■ GCRMA ■ GCOS

3.9 Rna-seq analysis in tomato leaves under drought stress

In collaboration with the lab of Dr. Grillo, CNR-IBBR Institute of Plant Genetics in Portici, and Dr. Bagnaresi, C.R.A. in Fiorenzuola, where I was hosted for one month, we defined appropriately the already available pipelines for RNA-seq technology, in order to identify genes differentially expressed (DEG) in tomato under drought stress.

The experiment carried out by Dr. Grillo's group consisted of a two drought stress cycles in tomato cv M82 (Fig. 26).

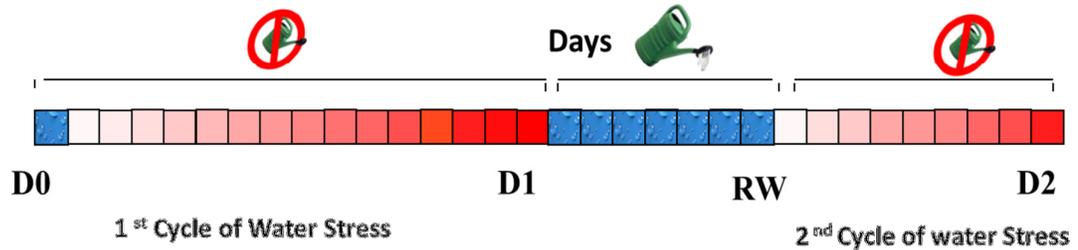


Figure 26 *Experimental Plan. D0= stop irrigation; D1= 16th day of drought (1st cycle of stress); RW= 7th day of rewatering; D2= 8th day of water deficit (2nd cycle of stress)*

RNA extractions from plant leaves were done at D1, RW and D2 steps, together with the control that was irrigated the whole time. All the stages had 3 technical replicates.

Statistical analysis were performed in order to found DEG taking in consideration different conditions pair. In particular: i) D1 and control (CNTRL); ii) re-watering (RW) and CNTRL; iii) D2 and CNTRL; iv) D1 and RW; v) D2 and RW and vi) D1 and D2.

From the analysis of all the comparison described, 966 genes showed differential expression in at least one of the analyzed conditions and were therefore considered as differentially expressed genes (DEGs) (Fig. 27).

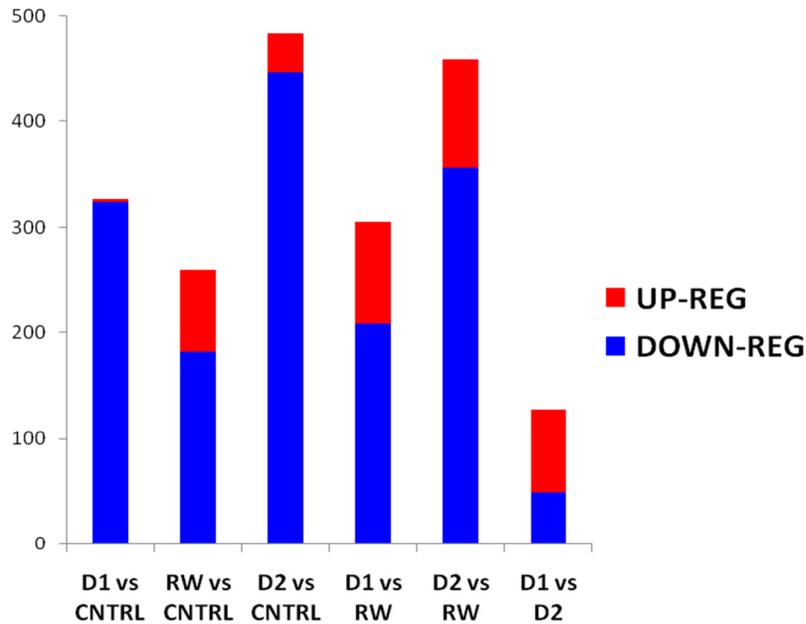


Figure 27 Total number of Differential Expressed Genes (DEGs) per comparison, specifying how many are UP or DOWN regulated

The comparison with the higher number of DEGs was D2 vs CNTRL, and in all the comparison the number of down regulated genes was much higher than the up regulated ones. This behavior is not confirmed only in D1 vs D2 comparison, where the up regulated genes were more.

After a general overview of the comparison, we focalized only on four comparison: D1 vs CNTRL, D2 vs CNTRL, D2 vs RW and D1 vs RW in order to find the key genes in drought response. We compared the DEGs among all the comparison selected and it is resulted that 119 genes were always differentially expressed. Meanwhile, 34 DEGs were D1 vs CNTRL specific, 93 specific for D2 vs CNTRL, 83 were present only in D2 vs RW and finally 69 were D1 vs RW specific (Fig. 28).

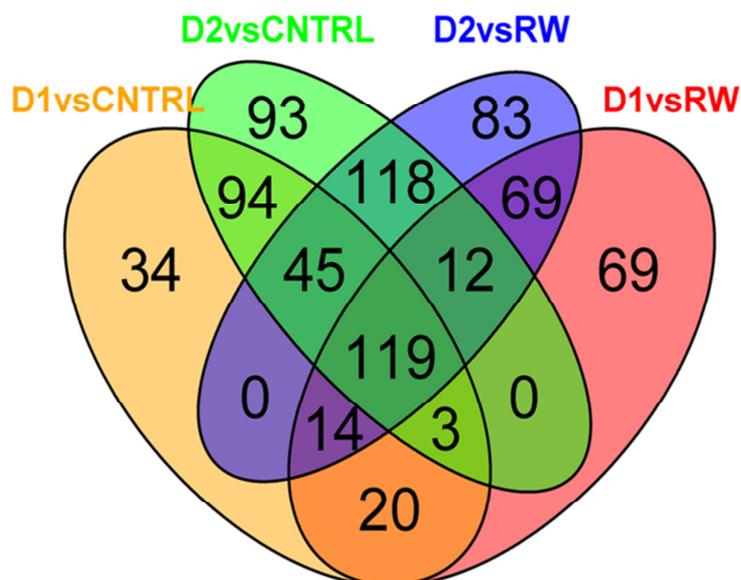


Figure 28 Venn diagram showing the common DEGs among four comparison: *D1vsCNTRL*, *D2vs CNTRL*, *D2vsRW* and *D1vsRW*

GO enrichment analysis was performed on the 119 genes in common among the comparison. The results was 11 GO enriched all involved in metabolic process or cell structure, confirming the changes of the functionality of the cell in response at the stress (Tab. 17).

Table 17 List of GO enriched from 119 DEGs

GOID	p value	# of DEGs	GOTERM	ONTOLOGY
GO:0006334	4,04E-07	6	nucleosome assembly	BP
GO:0009765	5,55E-06	4	photosynthesis, light harvesting	BP
GO:0005985	9,21E-06	11	sucrose metabolic process	BP
GO:0005982	9,94E-06	11	starch metabolic process	BP
GO:0042335	6,61E-05	4	cuticle development	BP
GO:0009505	1,36E-11	15	plant-type cell wall	CC
GO:0000786	9,98E-08	6	nucleosome	CC
GO:0048046	8,42E-06	10	apoplast	CC
GO:0045330	2,56E-05	4	aspartyl esterase activity	MF
GO:0030599	0,000108279	5	pectinesterase activity	MF
GO:0046982	0,000177526	5	protein heterodimerization activity	MF

By a cluster analysis of all 966 DEGs, seven clusters of DEGs with respect to their behavior similarity were selected for further investigation. Among them, 5 clusters showed higher expression level in control and re-watering conditions, while the remaining two clusters showed higher expression in D1 and D2 conditions (Fig. 29.A).

GO enrichment analyses were performed on the selected 5 clusters (1, 2, 3, 4 and 5) and 2 clusters (6 and 7) independently. The enrichment results highlighted that the genes related to photosynthetic light harvesting (such as Chlorophyll a/b binding protein-Solyc08g067320) and to modification of cell wall (i.e. Pectinesterase-Solyc09g075350) were found down regulated in D1 and D2 (Fig. 29.B). Interestingly, several genes encoding for Histone H3 and genes of sucrose and starch metabolic processes were found down regulated (Fig. 29.B). Our cluster analysis highlighted that genes up regulated during the cycle of drought stress are related to stress such as response to water stimulus (i.e. Dehydrin, Solyc01g109920.2) and water deprivation (such as 2 NAC domain protein IPR003441- Solyc12g013620.1/Solyc07g063410.2).

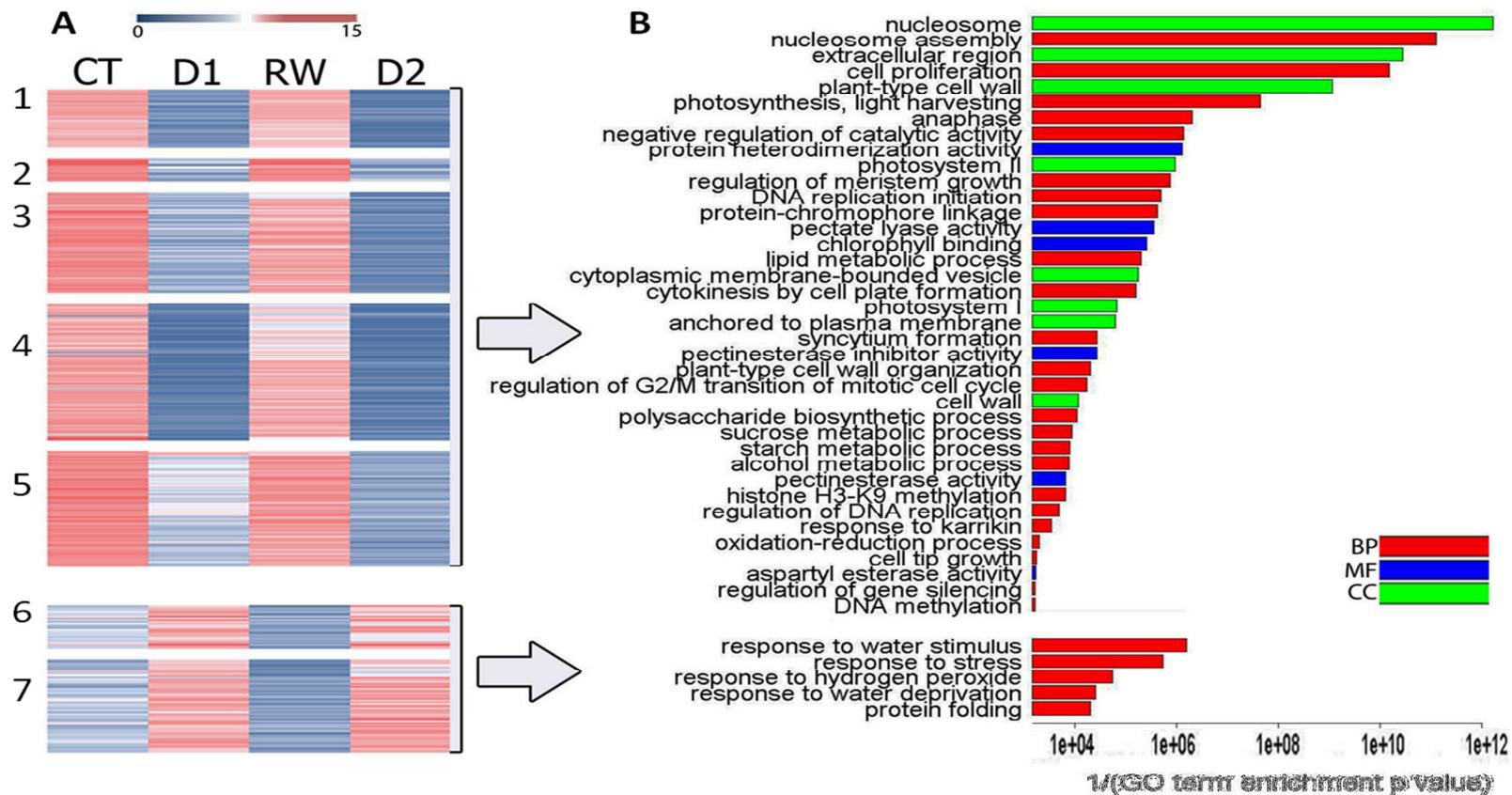


Figure 29 A) DEGs cluster selected for GO analyses. Clusters 1, 2, 3, 4 and 5 have opposite behavior compared to clusters 6 and 7
B) Barplot showing the results of GO analyses of clusters 1--5 and 6-7, respectively

Discussion

In this work, we exploited different plant genomic resources in terms of uniformity between reference collections. We put our attention on *Solanum lycopersicum* (tomato), the species studied in the laboratory where I carried out my PhD. Moreover, we extended our overview on other two species: *Arabidopsis thaliana*, model organism for plants, and *Solanum tuberosum* (potato), another Solanaceae recently sequenced.

The overview of all the genomics resources for this 3 species highlighted that for *Arabidopsis* and tomato the data available are quite homogenous, while for potato the resources available were heterogeneous and not updated.

Even though the genomic resources available for tomato exploit the most updated annotation versions (iTAG vers. 2.3 and 2.4), we highlighted how this two annotations definitely do not correspond to novel predictions, since the two genomes only differ in N nucleotides included to improve the genome assembly. Indeed, iTAG 2.4 only correspond to the translation of iTAG 2.3 on the new genome setting.

Going deeper into the tomato genome annotation, it was evident that the reference annotation are still lacking in many information, such as in some repeated genes (LSU and SSU). Moreover, many of the genes are wrongly annotated. Indeed, particular attention has been dedicated to the predicted genes that overlap other predicted genes and for the two long genes that overlap more than 50 other genes. Moreover, from our analysis it resulted that 1878 genes were probably wrongly defined as two or more genes instead of being one. Indeed, the presence of these putative split genes may affect the quality of many genomics analyses such as, as an example, RNA-seq analyses.

In order to check the overall quality of the tomato annotation, a remapping of the tomato mRNA on the tomato genome was performed. The results of this analysis emphasized not only the high content of repeated genes, 4593 genes mapped more than one time on the genomes, but also void regions that may contain gene not yet annotated.

Furthermore, in order to check the reliability of the available tomato annotations, we compared the official one (iTAG v.2.3) with the one available in RefSeq, on the NCBI website. The comparison highlighted the huge differences in the two annotations, which have only 1058 predicted genes with the same locus length and exon structure. Moreover, 8620 genes were predicted only in iTAG annotation while 2786 were predicted only in RefSeq. All the information obtained thanks to the analyses performed on the tomato annotation and to the comparison with RefSeq were collected into a “tomato annotation guide”, useful for the exploitation of an improved reference annotation.

In the last part of our work, we focused our attention on transcriptomics analyses, taking in consideration the three species considered: Arabidopsis, tomato and potato. In order to exploit co-expressed genes we went through microarray data. In this case the heterogeneity of platforms implemented for tomato and potato didn't allow us to go deeper on this aspect for Solanaceae. On the contrary, too many resources concerning gene expression collections from the same microarray platform were available for Arabidopsis. All this multitude of resources pushed us to compare them in order to understand which one was the most reliable. The co-expression platforms available were 11 and we exploited each providing a complete overview and also investigating the results from the same query. Indeed we investigated on the collection of co-expressed genes for CESA 7 and AT5G06680, which code for an element of a complex involved in the cell wall synthesis and for an element in the gamma-tubulin complex, respectively. The results highlighted

the huge differences in the platform outputs, due not only to different normalization methods exploited by each resources but especially to the difference in the collected datasets. In fact, further analysis confirmed that the heterogeneity of dataset from the same platform can affect the results. As an example, the inclusion or exclusion of mutants in a dataset affects the number of gene correlations and, consequently, the results from a co-expression analysis.

Finally, transcriptomics analysis were also performed on tomato, exploiting more advanced technology such as RNA-seq, in the light of setting up a pipeline for RNA-seq analysis but also to apply methodologies from gene co-expression to this type of data.

A suitable strategy to analyse RNA-seq data in tomato was set up in order to find differentially expressed genes (DEGs). This pipeline was test in a specific sample study for investigate response to drought stress. From the results of the analyses it was possible to clearly define a key role for a specific set of genes that was also confirmed by real time in collaboration with Doc. Grillo's group (data from real time are not shown here, since out of the scope of this thesis).

The basis of this analysis required a suitable gene annotation. As a consequence, we investigated on RPKM variations due to the exploitation of our revised annotation.

References

- Adams, M. D., S. E. Celniker, R. A. Holt, C. A. Evans, J. D. Gocayne, et al. (2000). "The genome sequence of *Drosophila melanogaster*." Science **287**(5461): 2185-2195.
- Anders, S. and W. Huber (2010). "Differential expression analysis for sequence count data." Genome Biol **11**(10): R106.
- Andolfo, G., F. Jupe, K. Witek, G. J. Etherington, M. R. Ercolano, et al. (2014). "Defining the full tomato NB-LRR resistance gene repertoire using genomic and cDNA RenSeq." BMC Plant Biology **14**(1): 120.
- Andolfo, G., W. Sanseverino, R. Aversano, L. Frusciante and M. R. Ercolano (2013). "Genome-wide identification and analysis of candidate genes for disease resistance in tomato." Molecular Breeding **33**(1): 227-233.
- Arabidopsis Genome Initiative, T. (2000). "Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*." Nature **408**(6814): 796-815.
- Bevan, M., K. Mayer, O. White, J. A. Eisen, D. Preuss, et al. (2001). "Sequence and analysis of the *Arabidopsis* genome." Curr Opin Plant Biol **4**(2): 105-110.
- Blanc, G., K. Hokamp and K. H. Wolfe (2003). "A Recent Polyploidy Superimposed on Older Large-Scale Duplications in the *Arabidopsis* Genome." Genome Res **13**(2): 137-144.
- Bombarely, A., N. Menda, I. Y. Teclé, R. M. Buels, S. Strickler, et al. (2011). "The Sol Genomics Network (solgenomics.net): growing tomatoes using Perl." Nucleic Acids Res **39**(Database issue): D1149-D1155.
- Bowers, J. E., B. A. Chapman, J. Rong and A. H. Paterson (2003). "Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events." Nature **422**(6930): 433-438.
- Brazma, A., P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, et al. (2001). "Minimum information about a microarray experiment (MIAME)-toward standards for microarray data." Nat Genet **29**(4): 365-371.

- C. elegans Sequencing Consortium, T. (1998). "Genome sequence of the nematode *C. elegans*: a platform for investigating biology." Science (New York, N.Y.) **282**(5396): 2012-2018.
- Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, et al. (2009). "BLAST+: architecture and applications." BMC Bioinformatics **10**(1): 421.
- Chang, S. B., T. J. Yang, E. Datema, J. van Vugt, B. Vosman, et al. (2008). "FISH mapping and molecular organization of the major repetitive sequences of tomato." Chromosome Res **16**(7): 919-933.
- Chen, A., X. Chen, H. Wang, D. Liao, M. Gu, et al. (2014). "Genome-wide investigation and expression analysis suggest diverse roles and genetic redundancy of Pht1 family genes in response to Pi deficiency in tomato." BMC Plant Biology **14**(1): 61.
- Chen, Y., Y. Wang and H. Zhang (2014). "Genome-wide analysis of the mildew resistance locus o ('MLO') gene family in tomato ('*Solanum lycopersicum*L.)."
- Chu, Y. and D. R. Corey (2012). "RNA sequencing: platform selection, experimental design, and data interpretation." Nucleic acid therapeutics **22**(4): 271-274.
- Cokus, S. J., S. Feng, X. Zhang, Z. Chen, B. Merriman, et al. (2008). "Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning." Nature **452**(7184): 215-219.
- Conesa, A., S. Götz, J. M. García-Gómez, J. Terol, M. Talón, et al. (2005). "Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research." Bioinformatics **21**(18): 3674-3676.
- Conte, M. G., S. Gaillard, N. Lanau, M. Rouard and C. Périn (2008). "GreenPhylDB: a database for plant comparative genomics." Nucleic Acids Res **36**(Database issue): D991-D998.
- Craigon, D. J., N. James, J. Okyere, J. Higgins, J. Jotham, et al. (2004). "NASCArrays: a repository for microarray data generated by NASC's transcriptomics service." Nucleic Acids Res **32**(Database issue): D575-577.
- Csiszar, J., E. Horvath, Z. Vary, A. Galle, K. Bela, et al. (2014). "Glutathione transferase supergene family in tomato: Salt stress-regulated expression of

representative genes from distinct GST classes in plants primed with salicylic acid." Plant Physiol Biochem **78**: 15-26.

Cunningham, F., M. R. Amode, D. Barrell, K. Beal, K. Billis, et al. (2015). "Ensembl 2015." Nucleic Acids Res **43**(Database issue): D662-669.

De Bodt, S., D. Carvajal, J. Hollunder, J. Van den Cruyce, S. Movahedi, et al. (2010). "CORNET: a user-friendly tool for data mining and integration." Plant Physiol **152**(3): 1167-1179.

Di Filippo, M., A. Traini, N. D'Agostino, L. Frusciantè and M. L. Chiusano (2012). "Euchromatic and heterochromatic compositional properties emerging from the analysis of *Solanum lycopersicum* BAC sequences." Gene **499**(1): 176-181.

Duvick, J., A. Fu, U. Muppirala, M. Sabharwal, M. D. Wilkerson, et al. (2008). "PlantGDB: a resource for comparative plant genomics." Nucleic Acids Res **36**(Database issue): D959-D965.

Eckardt, N. A. (2003). "Cellulose Synthesis Takes the CesA Train." The Plant Cell **15**(8): 1685-1687.

Eisen, M. B., P. T. Spellman, P. O. Brown and D. Botstein (1998). "Cluster analysis and display of genome-wide expression patterns." Proceedings of the National Academy of Sciences **95**(25): 14863-14868.

Gene Ontology Consortium, T. (2004). "The Gene Ontology (GO) database and informatics resource." Nucleic Acids Res **32**(suppl 1): D258-D261.

Goodstein, D. M., S. Shu, R. Howson, R. Neupane, R. D. Hayes, et al. (2012). "Phytozome: a comparative platform for green plant genomics." Nucleic Acids Res **40**(Database issue): D1178-D1186.

Gremme, G., V. Brendel, M. E. Sparks and S. Kurtz (2005). "Engineering a software tool for gene structure prediction in higher organisms." Information and Software Technology **47**(15): 965-978.

Griffiths-Jones, S., A. Bateman, M. Marshall, A. Khanna and S. R. Eddy (2003). "Rfam: an RNA family database." Nucleic Acids Res **31**(1): 439-441.

References

- Hirsch, C. D., J. P. Hamilton, K. L. Childs, J. Cepela, E. Crisovan, et al. (2014). "Spud DB: A Resource for Mining Sequences, Genotypes, and Phenotypes to Accelerate Potato Breeding." Plant Gen. **7**(1): -.
- Hoheisel, J. D. (2006). "Microarray technology: beyond transcript profiling and genotype analysis." Nat Rev Genet **7**(3): 200-210.
- Hyun, T. K., Y. Rim, E. Kim and J.-S. Kim (2014). "Genome-wide and molecular evolution analyses of the KT/HAK/KUP family in tomato (*Solanum lycopersicum* L.)." Genes & Genomics **36**(3): 365-374.
- Kim, D., G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, et al. (2013). "TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions." Genome Biol **14**(4): R36.
- Koornneef, M. and D. Meinke (2010). "The development of Arabidopsis as a model plant." Plant J **61**(6): 909-921.
- Korbel, J. O., A. E. Urban, J. P. Affourtit, B. Godwin, F. Grubert, et al. (2007). "Paired-end mapping reveals extensive structural variation in the human genome." Science **318**(5849): 420-426.
- Kuo, W. P., T.-K. Jenssen, A. J. Butte, L. Ohno-Machado and I. S. Kohane (2002). "Analysis of matched mRNA measurements from two different microarray technologies." Bioinformatics **18**(3): 405-412.
- Langmead, B. and S. L. Salzberg (2012). "Fast gapped-read alignment with Bowtie 2." Nat Methods **9**(4): 357-359.
- Lapitan, N. L. V., M. W. Ganal and S. D. Tanksley (1991). "Organization of the 5S ribosomal RNA genes in the genome of tomato." Genome **34**(4): 509-514.
- Martin, M. (2011). "Cutadapt removes adapter sequences from high-throughput sequencing reads." 2011 **17**(1).
- McKusick, V. A. and F. H. Ruddle (1987). "A new discipline, a new name, a new journal [editorial]." Genomics **1**.

References

- Mostafavi, S., D. Ray, D. Warde-Farley, C. Grouios and Q. Morris (2008). "GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function." Genome Biol **9 Suppl 1**: S4.
- Mueller, L. A., R. K. Lankhorst, S. D. Tanksley, J. J. Giovannoni, R. White, et al. (2009). "A Snapshot of the Emerging Tomato Genome Sequence." Plant Gen. **2**(1): 78-92.
- Mutwil, M., S. Klie, T. Tohge, F. M. Giorgi, O. Wilkins, et al. (2011). "PlaNet: combined sequence and expression comparisons across plant networks derived from seven species." Plant Cell **23**(3): 895-910.
- Mutwil, M., J. Obro, W. G. Willats and S. Persson (2008). "GeneCAT--novel webtools that combine BLAST and co-expression analyses." Nucleic Acids Res **36**(Web Server issue): W320-326.
- Nicholson, J. K. and J. C. Lindon (2008). "Systems biology: Metabonomics." Nature **455**(7216): 1054-1056.
- Obayashi, T. and K. Kinoshita (2010). "Coexpression landscape in ATTED-II: usage of gene list and gene network for various types of pathways." J Plant Res **123**(3): 311-319.
- Ogata, Y., H. Suzuki, N. Sakurai and D. Shibata (2010). "CoP: a database for characterizing co-expressed gene modules with biological information in plants." Bioinformatics **26**(9): 1267-1268.
- Pepper, S., E. Saunders, L. Edwards, C. Wilson and C. Miller (2007). "The utility of MAS5 expression summary and detection call algorithms." BMC Bioinformatics **8**(1): 273.
- Peterson, D. G., W. R. Pearson and S. M. Stack (1998). "Characterization of the tomato (*Lycopersicon esculentum*) genome using in vitro and in situ DNA reassociation." Genome **41**(3): 346-356.
- Peterson, D. G., S. M. Stack, H. J. Price and J. S. Johnston (1996). "DNA content of heterochromatin and euchromatin in tomato (*Lycopersicon esculentum*) pachytene chromosomes." Genome **39**(1): 77-82.
- Potato Genome Sequencing Consortium, T. (2011). "Genome sequence and analysis of the tuber crop potato." Nature **475**(7355): 189-195.

Powell, S., D. Szklarczyk, K. Trachana, A. Roth, M. Kuhn, et al. (2011). "eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges." Nucleic Acids Res.

Proost, S., M. Van Bel, L. Sterck, K. Billiau, T. Van Parys, et al. (2009). "PLAZA: a comparative genomics resource to study gene and genome evolution in plants." Plant Cell **21**(12): 3718-3731.

Pruitt, K. D., T. Tatusova and D. R. Maglott (2007). "NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins." Nucleic Acids Res **35**(suppl 1): D61-D65.

Quackenbush, J. (2003). "Genomics. Microarrays--guilt by association." Science **302**(5643): 240-241.

Redman, J. C., B. J. Haas, G. Tanimoto and C. D. Town (2004). "Development and evaluation of an Arabidopsis whole genome Affymetrix probe array." Plant J **38**(3): 545-561.

Rhee, S. Y. (2003). "The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community." Nucleic Acids Res **31**(1): 224-228.

Robinson, M. D., D. J. McCarthy and G. K. Smyth (2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." Bioinformatics **26**(1): 139-140.

Sharma, S. K., D. Bolser, J. de Boer, M. Sønderkær, W. Amoros, et al. (2013). "Construction of Reference Chromosome-Scale Pseudomolecules for Potato: Integrating the Potato Genome with Genetic and Physical Maps." G3: Genes|Genomes|Genetics **3**(11): 2031-2047.

Shearer, L. A., L. K. Anderson, H. de Jong, S. Smit, J. L. Goicoechea, et al. (2014). "Fluorescence In Situ Hybridization and Optical Mapping To Correct Scaffold Arrangement in the Tomato Genome." G3: Genes|Genomes|Genetics.

Sherman, J. D. and S. M. Stack (1995). "Two-dimensional spreads of synaptonemal complexes from solanaceous plants. VI. High-resolution recombination nodule map for tomato (*Lycopersicon esculentum*)." Genetics **141**(2): 683-708.

- Slonim, D. K. and I. Yanai (2009). "Getting started in gene expression microarray analysis." PLoS computational biology **5**(10): e1000543.
- Spellman, P. T., G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, et al. (1998). "Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast *Saccharomyces cerevisiae* by Microarray Hybridization." Molecular Biology of the Cell **9**(12): 3273-3297.
- Srinivasasainagendra, V., G. P. Page, T. Mehta, I. Coulibaly and A. E. Loraine (2008). "CressExpress: a tool for large-scale mining of expression data from *Arabidopsis*." Plant Physiol **147**(3): 1004-1016.
- Steinhauser, D., B. Usadel, A. Luedemann, O. Thimm and J. Kopka (2004). "CSB.DB: a comprehensive systems-biology database." Bioinformatics **20**(18): 3647-3651.
- Sugarbaker, D. J., W. G. Richards, G. J. Gordon, L. Dong, A. De Rienzo, et al. (2008). "Transcriptome sequencing of malignant pleural mesothelioma tumors." Proc Natl Acad Sci U S A **105**(9): 3521-3526.
- Suresh, B. V., R. Roy, K. Sahu, G. Misra and D. Chattopadhyay (2014). "Tomato genomic resources database: an integrated repository of useful tomato genomic information for basic and applied research." PLoS One **9**(1): e86387.
- Tang, X., A. Hou, M. Babu, V. Nguyen, L. Hurtado, et al. (2008). "The *Arabidopsis* BRAHMA chromatin-remodeling ATPase is involved in repression of seed maturation genes in leaves." Plant Physiol **147**(3): 1143-1157.
- Tomato Genome Consortium, T. (2012). "The tomato genome sequence provides insights into fleshy fruit evolution." Nature **485**(7400): 635-641.
- Torres, T. T., M. Metta, B. Ottenwalder and C. Schlotterer (2008). "Gene expression profiling by massively parallel sequencing." Genome Res **18**(1): 172-177.
- Toufighi, K., S. M. Brady, R. Austin, E. Ly and N. J. Provart (2005). "The Botany Array Resource: e-Northerns, Expression Angling, and promoter analyses." Plant J **43**(1): 153-163.
- Trapnell, C., B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, et al. (2010). "Transcript assembly and quantification by RNA-Seq reveals unannotated

transcripts and isoform switching during cell differentiation." Nat Biotech **28**(5): 511-515.

UniProt Consortium, T. (2015). "UniProt: a hub for protein information." Nucleic Acids Res **43**(D1): D204-D212.

Vallejos, C. E., S. D. Tanksley and R. Bernatzky (1986). "Localization in the Tomato Genome of DNA Restriction Fragments Containing Sequences Homologous to the rRNA (45s), the Major Chlorophyll a/b Binding Polypeptide and the Ribulose Biphosphate Carboxylase Genes." Genetics **112**(1): 93-105.

Vandepoele, K., M. Quimbaya, T. Casneuf, L. De Veylder and Y. Van de Peer (2009). "Unraveling transcriptional control in Arabidopsis using cis-regulatory elements and coexpression networks." Plant Physiol **150**(2): 535-546.

Wang, Z., M. Gerstein and M. Snyder (2009). "RNA-Seq: a revolutionary tool for transcriptomics." Nature Reviews Genetics **10**(1): 57-63.

Weber, A. P., K. L. Weber, K. Carr, C. Wilkerson and J. B. Ohlrogge (2007). "Sampling the Arabidopsis transcriptome with massively parallel pyrosequencing." Plant Physiol **144**(1): 32-42.

Wikstrom, N., V. Savolainen and M. W. Chase (2001). "Evolution of the angiosperms: calibrating the family tree." Proc Biol Sci **268**(1482): 2211-2220.

Wilke, M., S. K. Holland, M. Altaye and C. Gaser (2008). "Template-O-Matic: a toolbox for creating customized pediatric templates." Neuroimage **41**(3): 903-913.

Winkler, H. (1920). Verbreitung und Ursache der Parthenogenesis im Pflanzen- und Tierreiche / von Dr. Hans Winkler. Jena :, G. Fischer.

Wu, J., J. Wang, C. Pan, X. Guan, Y. Wang, et al. (2014). "Genome-wide identification of MAPKK and MAPKKK gene families in tomato and transcriptional profiling analysis during development and stress response." PLoS One **9**(7): e103032.

Xu, J. and E. D. Earle (1996). "Direct FISH of 5S rDNA on tomato pachytene chromosomes places the gene at the heterochromatic knob immediately adjacent to the centromere of chromosome 1." Genome **39**(1): 216-221.

References

- Xu, R. (2014). "Genome-wide analysis and identification of stress-responsive genes of the CCCH zinc finger family in *Solanum lycopersicum*." Mol Genet Genomics **289**(5): 965-979.
- Young, M., M. Wakefield, G. Smyth and A. Oshlack (2010). "Gene ontology analysis for RNA-seq: accounting for selection bias." Genome Biol **11**(2): R14.
- Zhang, T., X. Wang, Y. Lu, X. Cai, Z. Ye, et al. (2014). "Genome-wide analysis of the cyclin gene family in tomato." Int J Mol Sci **15**(1): 120-140.
- Zimmermann, P., M. Hirsch-Hoffmann, L. Hennig and W. Gruissem (2004). "GENEVESTIGATOR. Arabidopsis microarray database and analysis toolbox." Plant Physiol **136**(1): 2621-2632.
- Zouine, M., Y. Fu, A. L. Chateigner-Boutin, I. Mila, P. Frasse, et al. (2014). "Characterization of the tomato ARF gene family uncovers a multi-levels post-transcriptional regulation including alternative splicing." PLoS One **9**(1): e84203.