BIOINFORMATICS STRATEGIES FOR GENOMICS: EXAMPLES AND APPROACHES FOR TOMATO

HAMED BOSTAN

University of Naples "Federico II"

PhD in Computational Biology and Bioinformatics

(Cycle XXVII)

Supervisor: Dr. Maria Luisa Chiusano

Coordinator: Prof. Sergio Cocozza

October 2015

ABSTRACT

My PhD is funded by the *Solanaceae Pollen thermotolerance – Initial Training Network (SPOT-ITN)* in the frame of the *European Marie Curie Actions*.

The consortium aims to investigate fundamental and applied aspects contributing to the protection of pollen at increased environmental temperatures, deciphering the underlying of pollen development and its response to heat stress, starting from analyses on Tomato. Obviously, the findings are supposed to be a guideline, and the procedures to be applicable to other plants in the future.

In the light of the *SPOT-ITN* project objectives, and to provide a comprehensive bioinformatics infrastructure to support extensive genomics analyses in tomato, we collected, processed and integrated different resources; and organized them into dedicated databases with appropriate query user interfaces. This bioinformatics effort required the design of the most adequate software to reconcile the manifold resources from different cell information levels *(genomics, transcriptomics, epigenomics)*. This is fundamental for data integration and analysis.

The development of appropriate tools to mine the data from the *"omics"* approaches employed to trace the pollen development and the heat stress response has also been necessary to the project.

In this thesis, the main efforts undertaken and the analyses conducted on the basis of such resources with the strategies and approaches developed are reported in details.

I would like to dedicate my thesis to my father and Mother.

Without their support, I could never come that far.

ACKNOWLEDGEMENT

I am grateful to my supervisor Dr. Maria Luisa Chiusano, whose expertise, understanding, precious guidance and support made it possible for me to proceed in this field of research. It was and still is a pleasure working with her.

I would like also to acknowledge Dr. Klaus-Dieter Scharf from Goethe University, Frankfurt, Germany for his precious advises and supports.

I am grateful of Prof. Luigi Frusciante for his kind supports during the course of my PhD.

I would like to express my gratitude to Dr. Valentino Ruggieri for supporting me to better understand the biological aspects in the field of genomics.

I would like to also acknowledge Drs. Luca Ambrosino, Chiara Colantuono and Alfonso Esposito for their kind supports in the field.

I would like to acknowledge Dr. Yuanyuan Chen for his contribution in the work of "The Role of TE-Derived Small Interfering RNAs in Tomato Pollen Development" in the frame of the SPOT-ITN project. I am also grateful of all the SPOT-ITN members for their precious collaborations and useful interactions.

Hereby, I acknowledge *GenXPro* Company (Frankfurt, Germany) as one of the partners in the SPOT-ITN project for providing the *MACE*, *MethSeq* and *Small-RNA* NGS collections for pollen developmental stages that were used for some of the analyses discussed in this work.

My PhD was funded by the Solanaceae Pollen Thermotolerance – Initial Training Network (SPOT-ITN) in the frame of Marie Curie Action European funding (Grant agreement 289220).

I would like also to acknowledge all the colleagues and friends in the Genopom building at the University of Naples "Federico II" for all the support and making this period for me delightful and joyful.

Contents

1 Introduction

The development of Bioinformatics has been tightly linked to international collaborations in genome sequencing projects and to efforts of the pharmaceutical industry in its drive for drug discovery and development.

Bioinformatics strongly evolved also thanks to several outstanding steps forward in the "omics" methodologies. With the advent of new sequencing and high-throughput technologies in the last years, large-scale genome projects have significantly changed the face of biology enhancing the role of structural and functional genomics research [1, 2]. The sequencing of whole genomes in short time together with detailed definition of molecular information acquired from the genome functionality caused a revolution in biological sciences for their contribution to the study of molecular processes and of the mechanism underlying the context of cellular systems [3]. Having deep information from data describing genome organization and its modifications, the transcripts expression, from the protein coding and non-protein coding context, and the different substances within a test sample provides novel, unexpected overviews of molecular aspects of Systems Biology [3, 4].

Powerful tools were organized and are still necessary to organize the data and, for example, to study genome structure and regulation covering aspects such as definition and analysis of genomic sequences, gene structure prediction, modeling of transcriptional and translational control and large scale comparative analyses [5].

1.1 Aims and scope

My PhD is funded by the *Solanaceae Pollen thermotolerance – Initial Training Network (SPOT-ITN)* in the frame of the *European Marie Curie Actions*. The project initiated in 2012, it includes 9 partner institutions in which 3 from the private sector. Five peers from 4 European member countries and one non-European partner are involved. The consortium aims to investigate fundamental and applied aspects contributing to the protection of pollen at increased environmental temperatures, deciphering the underlying of pollen development and its response to heat stress, starting from analyses on Tomato. Obviously, the findings are supposed to be a guideline, and the procedures to be applicable to other plants in the future.

This bioinformatics effort required data exchange and the design of the most adequate software to reconcile the manifold resources from different cell information levels *(transcriptomics, proteomics, metabolomics, epigenomics)*. This is fundamental for data integration and analysis. The development of appropriate tools to mine the data from the *"omics"* approaches employed to trace the pollen development and the heat stress response has also been necessary to the project (Figure 1).

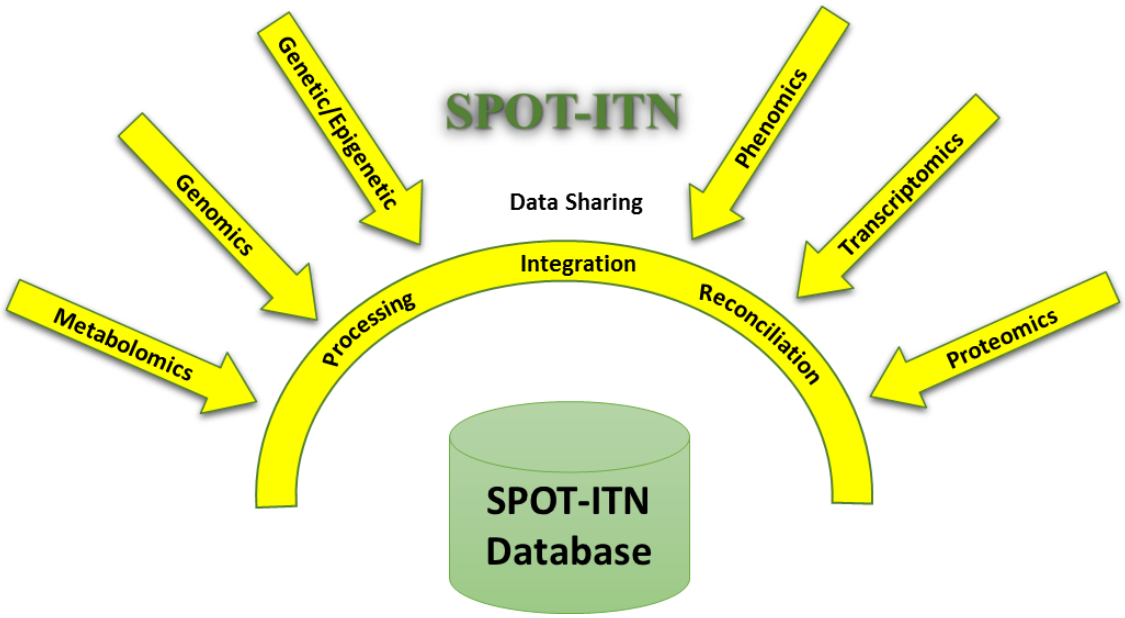


Figure 1: The SPOT-ITN Bioinformatics Platform Schema

My activities, in this thesis work, was the set-up of the bioinformatics platform for data sharing with a specific focusing to the organization and management of data from genomics and transcriptomics. I also contributed with the implementation of tools accompanying or integrating the already existing ones for improving data quality and supporting data analysis and supporting specific biological investigations useful to understand structure organization and functionalities of the tomato genome in the framework of the *SPOT-ITN* project.

1.2 Bioinformatics and "omics" Collections

1.2.1 Genome References

A genome reference is a sequence of DNA nucleotides (bases) assembled from the sequencing of DNA from a model species. With the advent of Next Generation Sequencing Technologies on 2005 [6] a new window to deliver fast, inexpensive and accurate genome information was opened [7]. Moreover, with

the advancement of such technologies and the number of genome sequencing projects working on different organisms and species, an increasing number of genome sequences for model and non-model organism were made available. As an example in the plant sciences, 18 genome sequences from Algae, one from *Bryophytes*, 59 for *Eudicots* and 19 for *Monocots* were made available (https://en.wikipedia.org/wiki/List_of_sequenced_plant_genomes).

1.2.2 Gene/Genome Annotation

Sequencing new genomes also involved the definition of their gene content. In fact, gene annotation is one of the main and routine steps in the genome analysis when the genome sequence becomes available. It is normally carried out before the genome sequence is deposited in the *GenBank* [8]. Several specific or general bioinformatics gene annotation pipelines also exist in the field [9-12].

Several pipelines (e.g.: *EuGene-PP*, *SEGMA* etc.) are used to predict the genes during the genome sequencing project while some others (such as RefSeq) are meant to collect the annotated genes later with some curations included.

NCBI's reference sequence (RefSeq) database [13] is a curated non-redundant collection which stores, organizes and provides access to the public sequences representing genomes, transcripts and proteins. On 13 July 2015, *RefSeq* database included 52,494,032 proteins, 11,803,354 transcripts and 55,267 organisms (*RefSeq* release 71). As a hub, *RefSeq* offers the integrated information from different resources and represent a current description of the sequence and its features if available [13]. *RefSeq* offers a reviewed collection in which the input from expert users and the other accessory details from the relevant scientific communities were combined. *GenBank RefSeq* collection is one of the main reference collections used in the research community.

1.2.3 Transcriptome Sequencing

The possibility of fast sequencing consistent transcriptome collections strongly contributed to gene annotation and to the understanding of differential expression in different biological context (tissues, stages, stress and pathologies) for several different species.

Several worldwide available resources collect transcriptome data in the form of sequences such as *dbEST* [14] and the Sequence Read Archive (*SRA*) [15]

dbEST is a division of *GenBank* [8], established on 1992 by the National Center for Biotechnology Information (*NCBI*), which is meant to collect raw reads and does not accept assembled sequences. In October 2015, the *dbEST* includes over 74 million Expressed Sequence Tag (EST) sequences from 2473 organisms. SRA is also another repository, established at *NCBI* on 2007, which includes DNA sequencing data from public collections especially in the form of short reads (normally less than 1,000 bp in length) produced by high throughput-sequencing (e.g.: *RNAseq*, *ChipSeq*, *MethSeq* etc.). As of October 2015, the SRA included over four quadrillion bases in its database.

1.2.4 Expressed Sequence Tags (ESTs)

Expressed Sequence Tags (ESTs) are small and error-prone RNA sequence pieces (normally ranging from 200 to 500 nucleotides) [16]. ESTs are derivative fragments produced by single sequencing pair sequencing [17] which are either generated by sequencing of one or both ends of an RNA molecule of all expressed genes. They are performed on randomly selected clones from c*DNA* libraries. ESTs are Small fragments of the mRNA that represent genes expressed in certain cells, tissues, or organs from different organisms, fishing of a gene out of a portion of genomic DNA is done by the "tags" matching base pairs (Figure 2) [18].

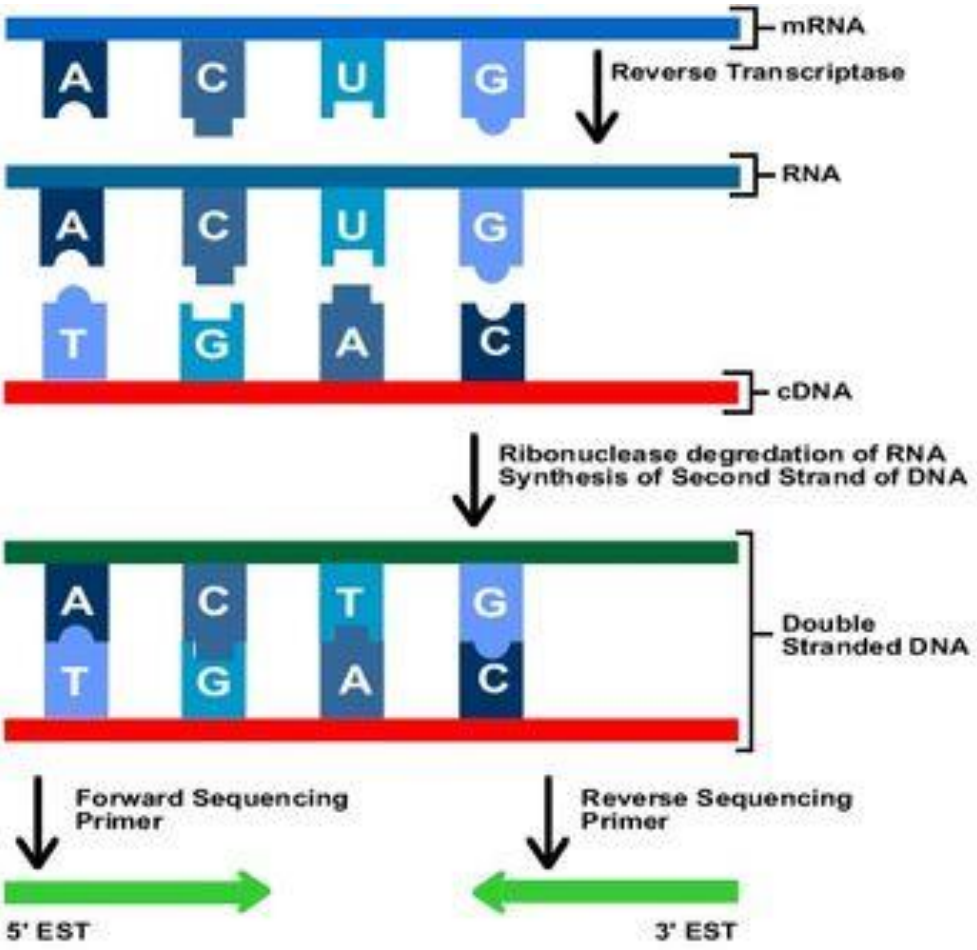


Figure 2: EST sample preparation for the sequencing (picture from [18])

Therefore, ESTs provide experimentally based important resources for comparative and functional genomic studies and represent reliable information for the annotation of genomic sequences [16].

1.2.5 Next Generation Sequencing Data

According to the materials used in this thesis, *RNAseq* and *MACE* techniques for the expression data, and *MethSeq* technique for the genome modification purposes are presented as follow:

RNAseq Data

RNAseq is a recently developed deep-sequencing technology exploiting Next generation sequencing technologies for parallel transcriptome profiling. It offers a significant level of precision comparing to the other methods in quantification of the produced transcripts and their isoforms [19].

Generally, an RNA population is converted to a cDNA fragments library. Depending on the protocol or approach selected, adaptor sequences are attached to one or both ends of the fragmented cDNAs (Figure 3). In most cases, an amplification process is subjected to the whole population. Depending on the sequencing technique, one end (*single-end*) or both ends (*pair-end*) is conducted. The reads typically range from 30 to 400 bases. *Illumina* (http://www.illumina.com/), Applied *Biosystems SOLiD* (http://www.appliedbiosystems.com/absite/us/en/home.html) and *Roche 454* Life Science systems (http://www.454.com/) are the example of such sequencing techniques applied for this purpose [19].

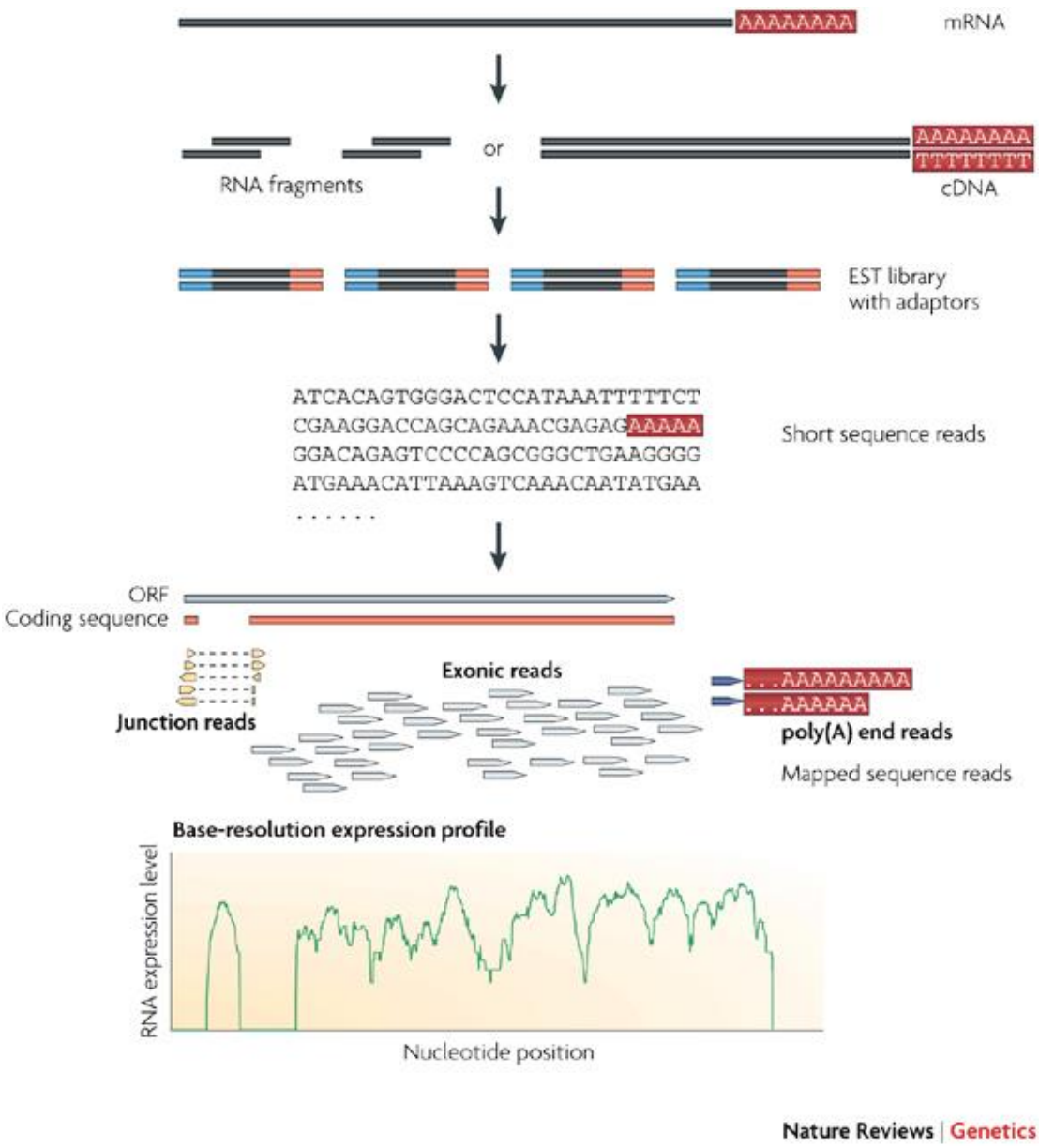


Figure 3: A typical RNAseq experiment (picture from [19])

Respectively, the resulted reads are mapped on the reference genome for the downstream analyses. When the genome reference is not available, assembly de novo of the transcript fragments is done to produce a genomic map of the sequenced species (This approach is also valid for the EST sequences). In both cases, several downstream analyses can be conducted such as expression quantification, structural and functional investigation etc.

MACE Data

Massive Analysis of cDNA Ends (*MACE*) [20] is a digital gene expression profiling technique and one of the latest advancement of tag-based gene expression analysis methods recently introduced by *GenXPro* Company in Germany [20]. It is also based on Next generation sequencing technologies. In *MACE* technique, a cDNA population is first linked to a streptavidin matrix via *3'-biotin*. The cDNA sequences are then fragmented into 50 to 500 bp pieces. One of the key point is that all the unbound fragments are discarded from the consequent analyses. A high-throughput sequencing is done on the bounded fragments starting from the bounding site.

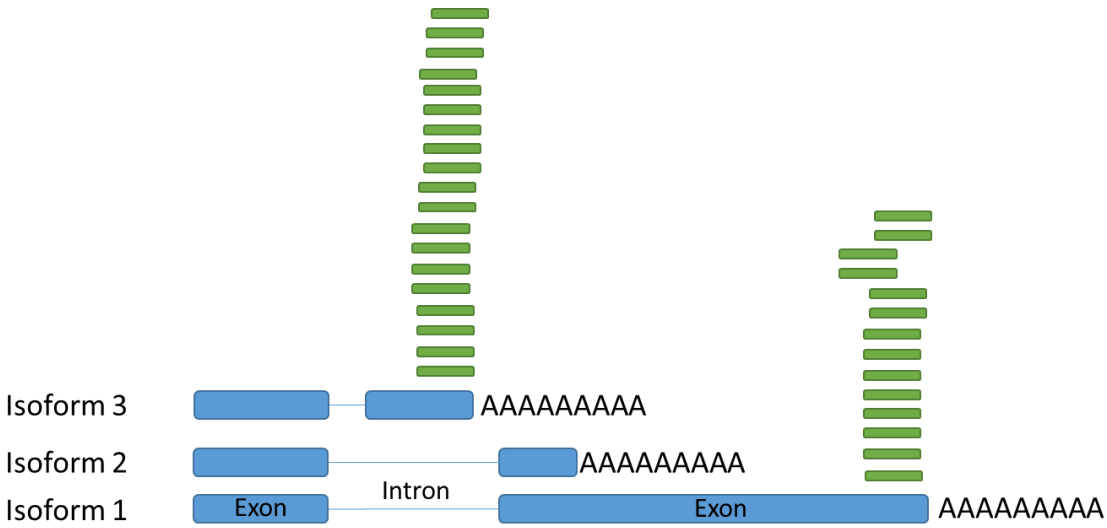


Figure 4: MACE reads alignment mapping on the genome representing different alternative transcripts and isoforms

Since *MACE* technique sequence the end of the transcripts (attached to the *poly A tail*), it can be a good way to detect the alternative transcripts and isoforms in the genomic loci (Figure 4). However the method is not able to provide/suggest the structure of the transcript.

Non-coding RNAs (Small- and Micro-RNAs)

The non-coding *RNAs* are referred to the class of RNAs that do not encode for any protein. They also represent a relevant component of the transcriptome level with relevant roles that are recently going to be more understood in molecular biology [21-25]. This class of RNAs are contributing (though still not well discovered and characterized) in various biological processes and complex cell control activities. *Small-RNA*s, including the silencing through homologous sequence interactions, can be named as short interfering (si)*RNAs* [26], small temporal (st)RNAs [27], heterochromatic siRNAs [28], tiny noncoding *RNAs* [29] and microRNAs (miRNAs) [30, 31]. Epigenetic modifications of the specific genomics regions, transposon silencing, RNA stability or translation are of those processes controlled by these classes of non-coding RNAs. Identification of the *Small-RNAs* and evolutionarily conserved RNA-mediated silencing pathways opened a new window to the understanding of the genomic processes in the field [32, 33].

MethSeq Data

Methylation-sensitive restriction enzyme assisted DNA methylation deep sequencing [34] (so called *MethSeq*) is one of the epigenomics approaches for detection of methylated and not methylated DNA sites. It is able to detect genome-wide CG methylation along the genome sequence. In this approach, *HpaII* is used as the methylation-sensitive enzyme, recognizing non-CpG-methylated CCGG sites. After this digestion process by *HpaII*, the size selected DNA fragments are subjected to the sequencing (Figure 5).

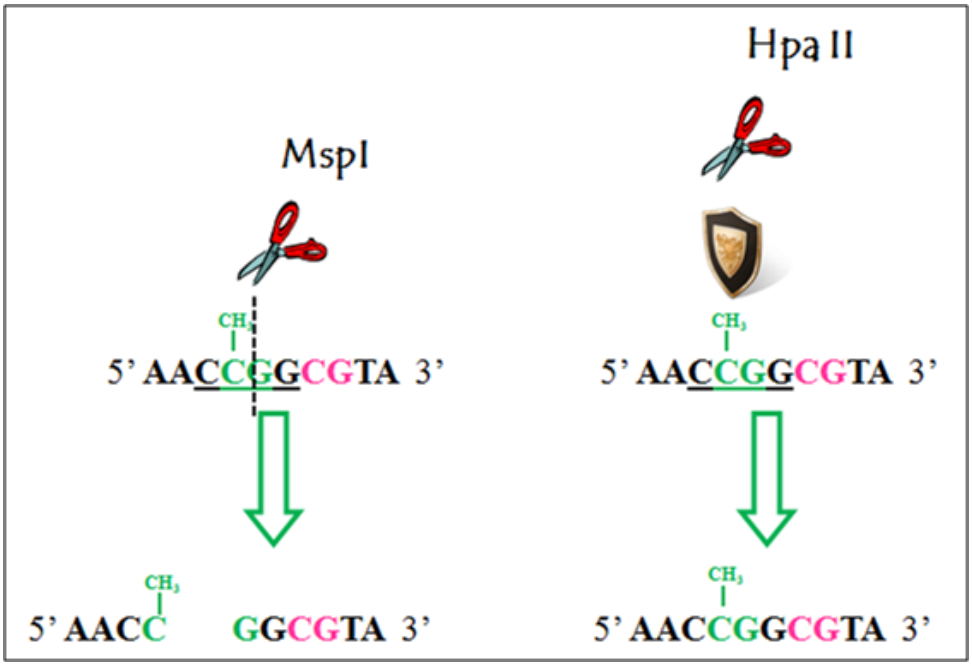


Figure 5: CG site cutting by enzyme in MethSeq methodology (picture taken from [35])

DNA methylation can be involved in the regulation of gene expression [36-38], protein function [39, 40] and mRNA processing [41, 42]. DNA methylation is also associated with the silencing of repeated regions, known also to cause genome instability in the plants and animals [43]. Hence, the detection of the DNA modification events associated to methylation is an important step to understand the role they can play in gene expression.

Data processing and assembly

The quality of data and the way it is processed have great impacts on the outcomes [1]. Most analytical tools assume that the input data has an accepted level of reliability, while for the sequencing data, both in genomics and in transcriptomics, due to technical or biological issues, that goes from machine biases, contaminations or several other aspects, the data should be quality assessed and pre-processed before any further analytical steps [2]. Besides of the data cleaning and trimming from usual factors (e.g.: additional sequences

used for sequencing purposes, vector sequences, low quality or missing bases), relying on the frequency of an evidence is also a common approach to enhance data reliability. To this end, the assembly of identical and overlapping sequences from the same reference as well as the definition of consensus from specific overlapping cut-offs are of those common methodological approaches to obtain high quality and reliable sequences. Sequence assembly and the clustering of sequences sharing identical or highly similar regions are also methodologies leading to the definition of variances such as those due to the *Single Nucleotide Polymorphisms* (*SNPs*) or splicing.

1.2.5.1.1 Sequence Cleaning and Trimming

In almost all the sequencing approaches, a piece of an additional sequence due to the technique (eq. vector sequences for the ESTs or barcode and adapter for the *RNAseq* data) contaminate the resulting target sequence(s) of interest. Based on the technique and the protocol used, the type of this sequence can vary. The need to remove the added sequences should be removed from the target sequence fragment is fundamental. Based on the type of data and the specific sequencing technology, several tools are developed to clean and trim such sequences from the raw sequence data. As an example for the EST sequences, *LUCY2* [44] and *SeqTrim* [4] are some of the tools to detect and remove the vector sequences from the EST. Sometimes the guided tools such as *RepeatMasker* [45] and the use of vector databases as the masking collections can be alternative approaches.

1.2.5.1.2 Quality Assessment

The trimming can result in very short sequences which should be discarded from the consequent analyses according to user defined specific cutoffs.

The length and the quality of the resulting sequences is a relevant aspect to be considered for several reasons. Due to technological or chemical issues and depending on the sequencing machine, the quality of the sequenced nucleotides may drop after a specific number of nucleotides. For the same reason, some sequences can also contain errors which should be considered depending on the proportion and position of the error occurrence. On the other hand, if the trimming of the sequence from adapters, vectors, or because of low quality nucleotides defines too short sequences, this will reduce their specificity. Indeed, the probability of similarity of short sequences compared to the longer ones, is higher, introducing bias to sequence assignment and to the definition of its role, such as in assembled sequences or in the detection of the correct reference genomic region when the sequence itself will be mapped. However this has a direct effect on the computational costs in terms of time complexity. Several quality assessment and correction tools [46-48] were developed to allow the quality assessment and low quality sequences removal by setting different options and cut-offs. Independently from these tools, several analytical pipelines also offer options to set length thresholds before further exploiting the resulting collection [49-52].

1.2.5.1.3 Tentative Consensus (TCs)

A Tentative Consensus (TC) sequence is the result of multiple sequences alignments. A consensus sequence has higher reliability because it is confirmed by several fragments that can also elongate the resulting product. The lack of a unique consensus can contribute information about variants or alternative splices.

CAP3 [50] is one of the most commonly used sequence assembly programs (offering different options such as overlapping cut-off, quality filtering and expansion thresholds etc.) used for the assembly from the EST collections and for TC definition.

The collections of all the tentative consensus from a transcriptome sequencing can be usually referred as unigene collections since a unigene should represent the unique reference for a transcript, thanks to the assembly process.

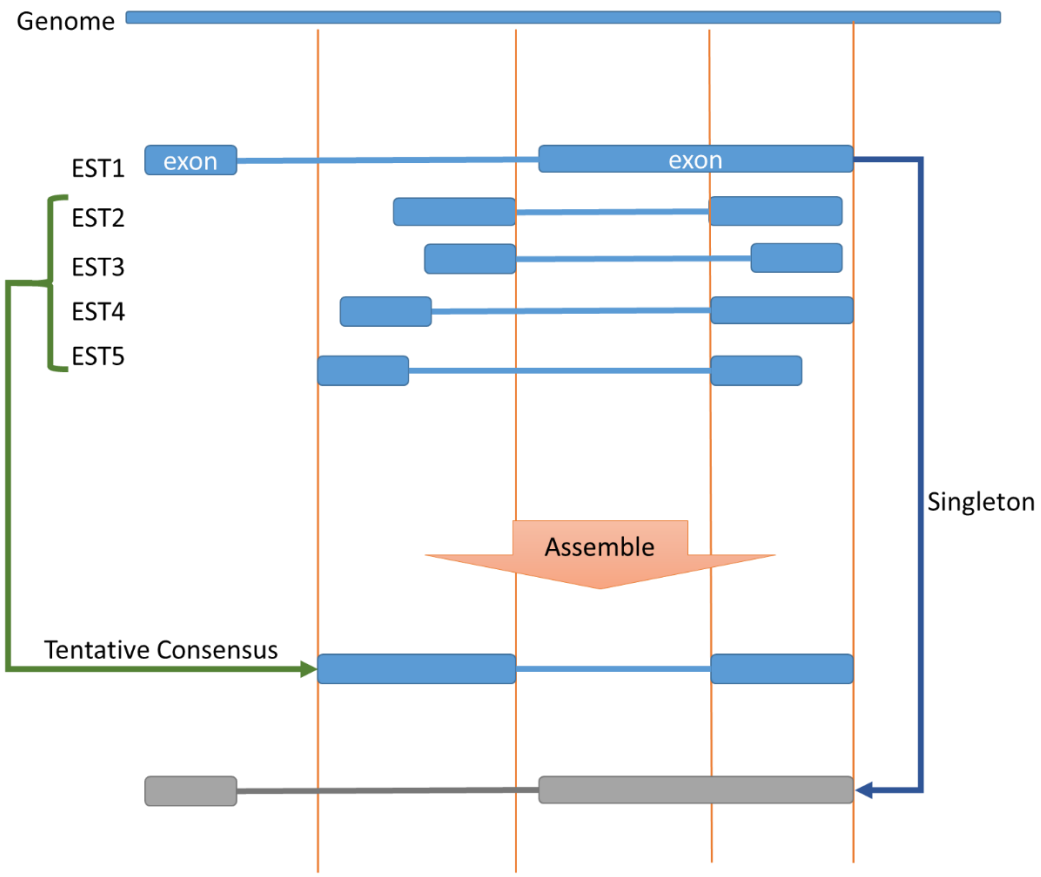


Figure 6: Tentative Consensus Assembly from EST sequences. Those not confirming a similar structure are left as singletons.

Based on the definition a TC is always defined at least by two sequences. The sequences from a library not contributing in any TC assembly are often referred as singletons (Figure 6). Some resources such as *ISOL@* [53], *NexGenEx-Tom* [54] offer the TC collections separating the singletons to flag the most reliable sequences since they can be independently investigated from their EST collections, but some others such as *SGN* [55], *PlantGDB* [56] and *DFCI* [57] provide the TC collections including the singletons together.

Considering the data from NGS, similar strategies are applied. Indeed the main difference is mainly due to the need of manipulating a higher number sequence fragments per analysis. In the field of transcriptome assembly, several tools and pipelines (e.g. *Denovo Trinity* [51], *Velvet* [58], *SOAPdenovo* [59], cufflinks [60], *SSAKE* [61] etc.) try to create the overlapping and consensus sequences from the short reads to assemble the entire transcript. Other tools (such as *BedTools* [62]) are also developed to create consensus sequences on the bases of reads overlapping on the genomic regions.

Quantification

In the NGS data quantification, an important summary statistic is the number of reads in a class (genomic feature such as gene, mRNA, exon etc.). The read count has indeed a linear function of the target abundance that is being measured: in *RNAseq* it can measure transcript abundance, or, in *MethSeq*, it can indicate absence of methyl group for sites that can be potentially methylated (eg. CpG dinucleotides).

There are two main approaches to follow to quantify a specific read amount from NGS approaches. In the genome reference based analyses, the reads are first mapped (aligned) on the genome sequence [60]. After the mapping, reads can be counted on the base of their occurrence on genomic features if available (e.g.: gene, mRNA, exon etc.). When the annotation is not available, an annotation free analyses based on different strategies may support the creation of reference genomic regions.

When a genome reference is not available, often a de novo assembly is done to have reference sequences as models [63]. In this case, the quantification of reads can be done by calculating the abundance of reads mapped on a specific model [51, 64].

In both case, the number of fragments (reads) counted is a quantity that may refer to that target abundance. This is valid for protein coding or non-coding transcripts. However, in terms of non-coding *RNAs*, often the genomic region or the class the reads will be assigned to, are not already available from the genome annotation. These regions, however, can be defined by identifying overlapping fragments mapped on the genome reference sequence [65]. In this approaches, often an offset of neighboring (e.g. a 100 window) for creating the reference feature is also considered. Furthermore, these clusters may be also intersected with other genomic features (e.g. coding regions, Transposon Elements etc.) for further downstream analyses [44, 66, 67] to localize the specific read amount on the genomic feature.

1.2.5.1.4 Raw Reads

The raw reads count is the simplest measure of quantitation for the high-throughput sequencing data. It counts up the reads within or overlapping a specific genomic region or a probe. There are several tools and packages which allow the fast and customized summarizing of the reads count for genomic features. Among all, *HTSeq-count* [68] and *featurecounts* [69] are the most common and user-friendly packages.

HTSeq-count is a Python script available in the *HTSeq* package developed to work with the NGS short reads (*fastq*). It is a fast and efficient software to summarize (assign) the reads mapped on a genomic reference to an overlapping genomic feature or class allowing several specific settings. *HTSeq-count* is of those summarizing tools which prefers to relay on the most certain evidences, and the discarding of ambiguous and multiple mapped reads is one of its principle procedures.

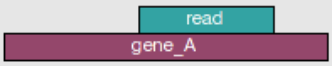
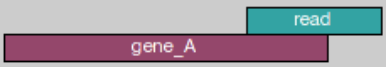
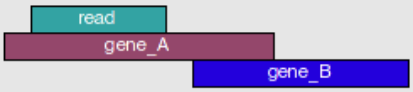
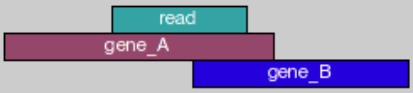
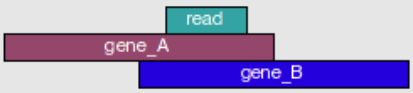
	union	intersection _strict	intersection _nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

Figure 7: Reads Counting options in HTSeq-count software (figure from http://www-huber.embl.de/users/anders/HTSeq/doc/count.html)

Figure 7 visually described the read assignment of the *HTSeq-count* package to the reference genomic feature using different parameters of *union*, *intersection_strict* and *intersection_nonempty*. As it can be observed, when the read "A" is overlapping with both "gene A" and "gene B", it is reported as ambiguous in all the cases.

featurecounts is also another tool developed in C environment for the short reads summarization. It is a highly efficient general-purpose software that allows the detailed counting of the mapped reads for various genomic features.

Against *HTSeq*, *featurecounts* can also deal with the multiple mapped reads and ambiguous features in which, the read will account for all the overlapping genomic elements. It also can deal with the genomic bins and chromosomal locations. *featurecounts* is available in the form of *SourceForge Subread* package or the *Bioconductor Rsubread* package [69].

Using any read to feature assignment tool for the quantification purposes, the summary statistic is normally calculated for all the class members per each replicate, creating a quantification matrix. As an example for the expression data of twenty genes of tomato in six different conditions (one replicate per each), a summary expression matrix of 20 × 6, excluding the headers and row names, will be produced.

ID	Cond1	Cond2	Cond3	Cond4	Cond5	Cond6
Solyc00g005000.2.1	0	2	1	1	5	3
Solyc00g005040.2.1	19	6	3	0	5	5
Solyc00g005050.2.1	2385	1725	1758	1509	1300	1450
Solyc00g005840.2.1	7564	5312	4826	5102	5400	5701
Solyc00g006470.1.1	35297	4119	1434	792	1802	2352
Solyc00g006490.2.1	2183	1330	1254	1187	662	926
Solyc00g006800.2.1	3606	2803	3610	2359	1705	1673
Solyc00g006810.2.1	3211	3089	3812	4021	3547	3321
Solyc00g006820.2.1	2292	2641	2265	1759	1350	1657
Solyc00g006830.2.1	1886	1915	1473	1274	975	1131
Solyc00g007010.2.1	820	738	712	618	404	531
Solyc00g007060.2.1	1259	1123	1079	801	538	752
Solyc00g007070.2.1	4166	3014	2869	2208	1533	1769
Solyc00g007080.2.1	34	65	47	47	29	8
Solyc00g007090.2.1	480	474	463	434	374	312
Solyc00g007100.2.1	1245	899	808	746	513	547
Solyc00g007110.2.1	284	268	230	143	171	151
Solyc00g007120.2.1	649	526	491	354	224	275
Solyc00g007130.2.1	122	70	60	50	53	76

Figure 8: Snapshot of an expression matrix for 20 example genes in 6 conditions

Figure 8 shows an example of a summary statistic table for 20 genes in 6 different conditions in the form of a gene expression matrix. As it can be observed, the name of the genes (here the genes of tomato) are listed as the row names while the conditions are the columns. The expression values corresponding to each gene in each condition is reported respectively. This summary statistic table (matrix) is then used for the gene expression analyses and profiling. However, as mentioned at the beginning of this topic, correction of the nonlinear effects that might be introduced due to the experimental conditions should be put into consideration for the quality purposes [70].

Normalization

Although it was claimed that RNA sequencing technology has a significant reduction of variability in comparison to microarrays, it is demonstrated that the unwanted and obscuring variability similar to what was first observed in microarrays can be also observed in the *RNAseq* data. [71]. In addition, the current limits in the sequencing technologies introduce a variety of biases to the data [72-75] suggest that normalization approaches are necessary to make the samples comparable. The aim of normalization approaches are to remove systematic technical effects with in the data, and minimize the impact of technical biases on the results [76]. Here, some of the most commonly used and popular normalization approaches used for the NGS data analyses are presented.

Reads Per Kilobase per Million (*RPKM*) mapped reads is a normalization allowing the comparison of the genes within a sample or between different samples by re-scaling the gene counts corrected for differences in both library sizes and gene length [77]. Although it has been shown that the correction of differences in gene length can introduce a bias in the differential analysis

especially for the lowly expressed genes [73], *RPKM* is still a popular and commonly used normalization approach in many practical applications.

$$RPKM = \frac{n * 1000\ bp * 10\text{^}6}{(L * N)}$$

Where *n* is the number of mapping reads, *L* is the length of transcript and *N* is the number of total reads in the sample collection. This method is a library size concept normalization approach.

Transcript Per Million (*TPM*) is the analogous approach of normalization to RPKM to correct the library size when the length of the transcript is not put into consideration.

$$TPM = \frac{n * 1000\ bp * 10\text{^}6}{N}$$

Where *n* is the number of mapping reads and *N* is the number of total reads in the sample collection.

This method is a library size concept normalization approach.

The reads assigned to a class or genomic feature (eg. gene or exon etc.) are divided by the total number of reads mapped on the genome (library size) for that specific lane (replicate/sample). The result is then multiplied by the total count mean across all the replicate/samples of the dataset.

Upper Quantile (*UQ*) in principle is very similar to Total Count (*TC*) where the total read counts are replaced by the upper quartile of non-zero counts in the normalization factors calculation [78].

Median (*Med*) is also similar to *TC* where the total read counts are replaced by the non-zero median counts in the normalization factors calculation.

DESeq method is a normalization method included in the *R Bioconductor* package (version 1.6.0) [79]. As for many other normalization approaches, the method is based on the hypothesis that most genes are not differentially expressed. This assumption leads to this that the non-DE genes should possess similar number of reads across all the samples, with the approximate ratio of 1. Hence, to estimate the correction factor to be applied to all read counts of a lane to support this hypotheses, the median of the ratio for each of the genomic features or classes of its reads counts over its geometric mean across all the lanes are calculated as the scaling factor. *"sizefactors()"* and *"estimateSizeFactors"* are the functions calculating this factor in the *DESeq* package. Eventually, the final gene expression is calculated by dividing of the raw counts by this factor for each genomic feature or class in the corresponding lane.

Trimmed Mean of M-values (*TMM*) method is a normalization approach also implemented in the *edgeR Bioconductor* package (version 2.4.0) [76]. This method like *DESeq* is based on the assumption that most genes are not differentially expressed. In the TMM normalization one lane is considered as reference sample and all the others as the test samples. After excluding most of the expressed genes and the genes with the largest log ratios, the weighted mean of log ratios between the test and the reference sample is calculated. This factor should be marginal to 1, otherwise a correction factor will be estimated as the library sizes. This scaling size factor is calculated by *"calcNormFactors()"* function is the *edgeR Bioconductor* package. To obtain the normalized counts per each genomic feature or class, raw reads counts are divided by the normalization factors re-scaled by the mean of the normalized library sizes.

Post-Processing and Interpretation

In system biology, understanding the interactions within a living cell can lead to the characterization of molecular components and common functionalities [80, 81]. As a matter of fact, a major challenge in biology is to decipher the dynamics observed in complex intracellular networks of interactions which lead to the structure organization and function of living cells [80]. Correlation based approaches and cluster analyses the two major strategies used for the construction of such interaction networks and profiling. Here we provide a brief description of such approached applied in the field of biology.

1.2.5.1.5 Correlation Analyses

Correlation networks are used widespread in bioinformatics in which describes the correlation patterns among components (genes, proteins etc.) across the different samples conditions, levels etc. [82]. Nowadays, it is hardly found a bioinformatics tool or package developed for the co-expression or co-regulation analyses in which, the correlation analyses was not deployed in it e.g.[82-85].

1.2.5.1.6 Cluster Analyses

As well as the "Correlation Analyses", clustering approaches have proven to be useful to identify the molecular components with similar behavior [86]. In statistics, clustering is a process in which the data is divided into similar or homologous groups, objects or categories minimizing the variance within each cluster [87]. It is an active field of research mainly used in for the pattern recognition and machine learning. Due to the specificity and sensitivity of statistical models and the type of data subjected to the cluster analyses, various clustering algorithms has recently emerged in the field. Hierarchical Clustering (a connectivity based model), *K-menas* algorithm (a centroid model), clique

algorithm (a graph based model) are of those commonly used approaches in the field of life sciences. Providing a comprehensive review on the clustering approaches available is a major effort which is out of the topic of this work, but here we suffice to the introduction of few commonly clustering algorithms used in the "omics" data analyses.

Hierarchical clustering is a very popular and commonly used clustering approach in which, a similarity or dissimilarity measure (Euclidean distance, Squared Euclidean distance, Manhattan distance etc.) is used to link the objects together in a greedy way [88]. In other words, a dendrogram is formed representing the similarity or dissimilarity of objects where the distance between the linkages is the measure of their dissimilarity. Based on the way the data is traversed to be grouped together, the Hierarchical Clustering can be divided into Agglomerative (bottom up) or Divisive (top down) categories. The Hierarchical Clustering is available in the stats package implemented in R.

Although there are some packages such as *pvclust* [89] in which a cluster number can be assigned to partition the dendrogram into different clusters based on different parameters (such as p-value for the *pvclust* package), the principle Hierarchical Clustering approach is independent from the number of clusters to be specified.

k-means clustering [33] is a hard clustering technique (each object falls into one cluster only) in which given a specific number of k, where k represents the number of clusters, the objects will be divided into k disjoint groups with maximum similarity and dissimilarity within the cluster and between the clusters respectively [90].

In *k-means* clustering, each cluster is represented by a centroid (c_i) which is the mean or weighted average of its data points.

$$E = \sum_{i=1}^{k} \sum_{O \in c_i} |O - \mu_i|^2$$

Where O is a data object in cluster C_i, μ_i is the centroid (mean of objects) of C_i, and E is the objective function to be minimized.

The objective function E tries to minimize the sum of the squared distances of objects from their cluster centers.

As observed above, *k-means* clustering only works on the numeric data but not categorical. In general, the decrease in the number of clusters results to lose some fine information but simplify the procedure to a great extend [87].

K-means clustering is one of the most used approaches in terms of gene expression profiling and pattern partitioning.

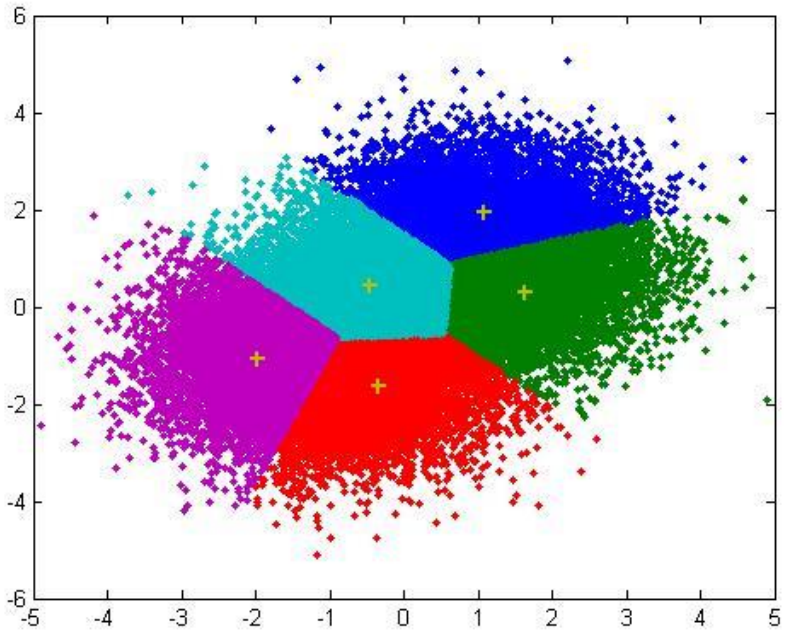


Figure 9: Demonstration of k-means clustering as a hard clustering technique (picture from [91]

As presented in Figure 9, the data points are categorized into different groups, in which each data point belongs to a specific cluster. Each cluster has a centroid that is the mean of all the data points within that cluster. Depending on the cases, different distance metric are used as a measure of variance for each cluster.

C-means clustering is a fuzzy clustering approach which was first presented on 2001 for text and image segmentation [12]. Like other fuzzy clustering methods, c-means is a soft clustering, in which each data object can fall into multiple clusters possessing a degree of membership. Obviously, a hard clustering or grouping of objects using a specific number of cluster can be done on the soft clustering considering the membership level of each element in each cluster.

Fuzzy C-means clustering also is implemented in package "e1071" of R environment.

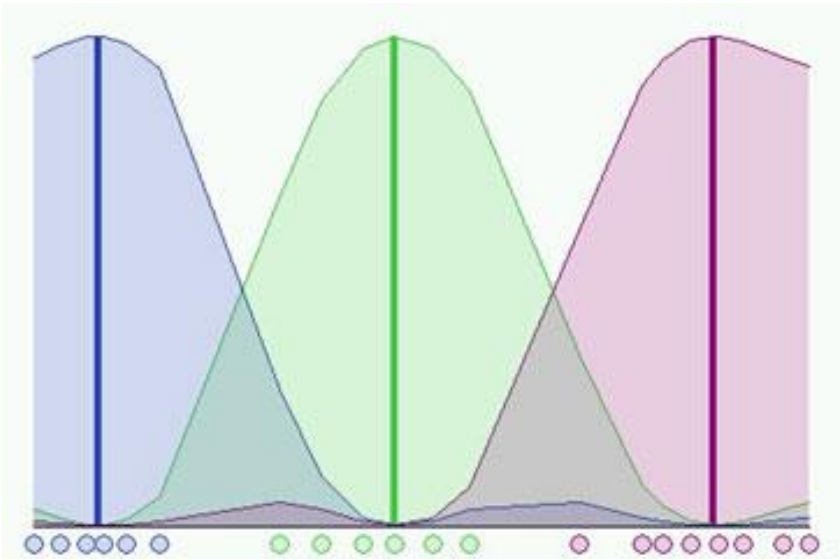


Figure 10: Demonstration of c-means clustering as a soft clustering technique (picture from [92])

As presented in Figure 10, the clustering topology in c-means clustering is presented. As it can be observed, each data point can belong to different clusters

possessing different degrees of membership depending on its distance to each cluster center.

Self-Organizing Map (SOM) approach was first introduced by *Teuvo Kohenen* in 1998. It is an effective visualization tool for relevant mapping of a high-dimensional distribution onto a low-dimensional grid. Hence, it is able to convert complex, nonlinear statistical relationships among high-dimensional data objects into simple geometric relationships on a low-dimensional display [93]. SOM is an excellent tool in exploratory phase of data mining [94] which is vastly applied to the omics data analyses [72, 74, 75, 80-82, 95].

Besides of the original toolkit developed by *Teuvo Kohenen* [93] for the SOM, several other packages are also implemented in R packages such as *kohonen*, *som* etc.

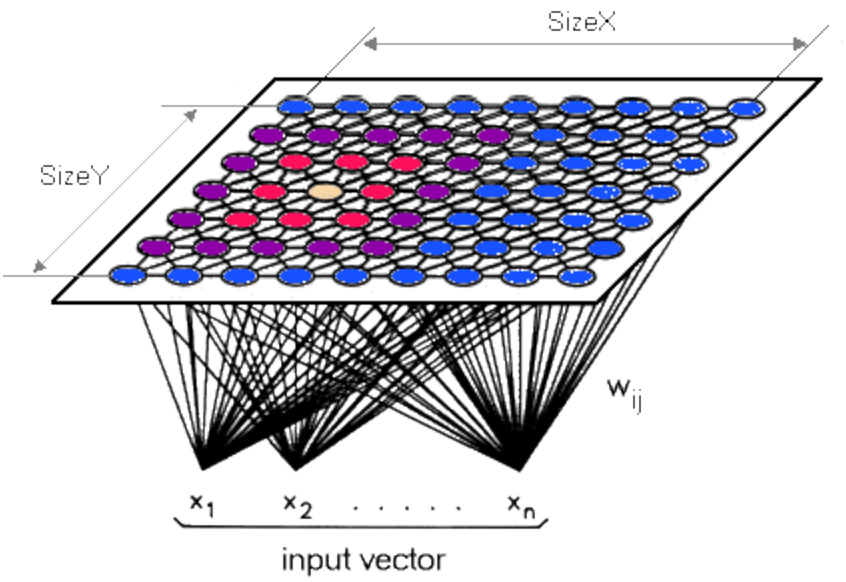


Figure 11: Self-organizing map clustering schema (picture from [96])

As depicted in Figure 11, the self-organizing map of n different inputs (here defined as vector) scattered on a two dimensional map is presented. In this methodology, similar data points converge while the dissimilar points diverge.

1.2.5.1.6.1 Optimal number of clusters

A simple method to determine the number of clusters for a sample population of objects is the "Rule of thumb" [83] which is formulated as follow:

$$k \approx \sqrt{\frac{n}{2}}$$

Where *k* is the number of clusters and *n* is the total number of objects to be clustered.

This method is not very precise or flexible but still is a popular and common method to be used for simple calculation of k at the first place to have an idea of the data behavior.

1.2.5.1.6.1.1 The Elbow method

The "Elbow method" is another approach for identifying the number of clusters for a group of objects subjected to the cluster analyses which can be traced back to 1997 [83]. By plotting the percentage of variance explained by each cluster against the number of cluster (analogous to F-Test), a curve will appear that in might signifies a drastic change (lowering) in the percentage if variance explained in the clusters (Figure 12). Obviously in this method the number of clusters have chosen is always with a level of ambiguity [84].

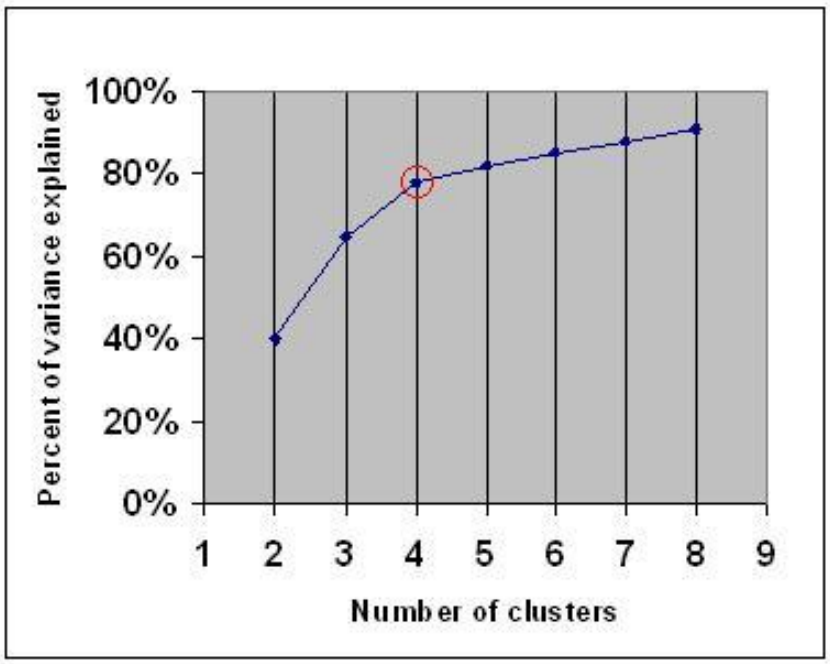


Figure 12: Elbow method for defining the number of clusters in a sample data (Picture from [97])

Here we presented some common clustering algorithms used in "omics" data analyses and two simple but popular approach for choosing the proper number of clusters in a population. A comprehensive survey on the clustering algorithms can be found at [89, 98].

1.2.5.1.7 Enrichment Analyses

GO Enrichment Analyses is the identification of the class of genes or proteins mainly over- but sometimes also under-presented in a set [99]. Functional analysis of large gene lists mostly resulted from high-throughput methodologies is a big challenge Gene annotation enrichment analysis is a promising approach in which the likelihood GOs associated to the resulted gene list is a measure of identifying the biological processes relevant to the study [100].

Since the inception of GO Annotation Project [101], the gene product function on the bases of Gene Anthology representing the relevant biological knowledge

was made available gradually. This data is now available at *AmiGO 2* Database (http://amigo.geneontology.org/amigo) providing access to these information in various ways (query, database service, web services etc.). *Ensembl BioMarts* [102] as another reference resource provides dedicated GO Anthology collections to the research communities.

Moreover to better exploit these resources and information, several Enrichment Analytical tools and packages were developed during the last years [99, 103-108].

Some tools such as *Blast2Go* [109] also provide some pipelines for the annotation of GO using the GO Anthology resources based on the gene product functional annotation by tuning several pipelines.

GO Terms Enrichment Analysis are based on some statistical approaches to detect the over- and sometimes under-presented GO terms. Here we present Fisher Exact Test as one of the most commonly used statistical methodologies for the GO Terms Enrichment Analyses.

1.2.5.1.7.1 Fisher Exact Test

Fisher test is a statistical test of significant on a contingency table (Table 1).

Table 1: an example of a contingency table

	Healthy	**Diseased**	**Total**
Male	10	23	33
Female	11	8	19
Total	21	31	52

A contingency table is a matrix form table in a in which the (multivariate) frequency distribution of the variables are summarized.

Table 2: variables representation for the contingency table elements to conduct fisher test

	Healthy	**Diseased**	**Total**
Male	a	b	a+b
Female	c	d	c+d
Total	a+c	b+d	n = a+b+c+d

With respect to the labeling reported in (Table 2), the formula to calculate the Fisher Exact Test significance level is as follow:

$$P = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{a!\,b!\,c!\,d!\,n!}$$

Due to the formulation of *Fisher Exact Test* and usage of factorial notations, however the formula is valid for any sample, the analyses can be directly done only on a small sets. To overcome this issue, many tools use the *Stirling* approximation to estimate the factorial for large numbers.

$$W = \log\frac{N!}{\sqrt{2\pi}} \cong \left(N + \frac{1}{2}\right)\log(N) - N\ Log\ (e)$$

$$W = A + B$$

$$N! = \sqrt{2\pi}\ \times 10^{A}\ \times 10^{B}$$

Or

$$N! = C * \ 10^{A}\ \ with\ C = \sqrt{2\pi}\ \times\ 10^{B}$$

Where *A* is the integer part and *B* is the decimal part of *W*. Hence, any *N* can be calculated as follow:

$$\frac{N_1!\ N_2!}{N_3!} = \frac{C_1 \times\ 10^{A1}\ \times\ C_2 \times\ 10^{A2}}{C_3 \times\ 10^{A3}} = \frac{C_1 \times\ C_2}{C_3} \times 10^{A1+A2+A3}$$

In this way, all the multiplication and division of factorials with the same base is possible in a simple but estimated way.

1.2.6 Orthology

Identification of ortholog/paralog genes is an important issue in molecular biology that supports structural, functional and evolutionary inferences [110-123]. Detection of ortholog/paralog genes has a wide range of applications in functional investigations and comparative genomics [113, 124]. As an example, a common procedure to characterize the genes in a newly sequenced genome is to investigate the orthology relationships for transferring functional information from the genes in the model organisms [125-127]. It can also highlight the species specific peculiarities. As another example, paralogy also allows to understand the expansion or reduction of some gene families or functionalities in the evolution process.

Due to the importance such approach can convey, several bioinformatics pipelines for detection of ortholog genes have been developed during the last years. *Inparanoid* [128] and *OrthoMCL* [129] are of those commonly used.

Consequently, several platforms also offer ortholog resources to the research community. Among all, PLAZA [130], *Phytozome* [58], Ensemble Plant *BioMart BioMart* [59],*Inparanoid* [61] and *OrthoMCL* [131] are the most comprehensive plant resources in the field.

1.3 Solanaceae and S. lycopersicum (Tomato)

The availability of the tomato genome and the gene annotation together with different "omic" resources such as collections from Expressed Sequence Tags (ESTs), Tentative Consensuses (TCs), Transcript Indices (unigenes) and the NGS data opened a new window to the research community to further and better investigate the genomic resources on the tomato genome space. However, the completion of a genome sequencing effort is never at an end, and the need of a reliable annotation is fundamental to fully exploit the acquired knowledge (section 2.3).

1.4 Resources for Tomato

Various biological databases and platforms such as *ISOLA* platform [53], *SolEST* database [109], *TomatEST* [132], *PlantGDB* [56] and *Dana-Farber* [57], *KafTOM* [133], *MiBase* [134] were providing relevant transcriptomics data (ESTs and TCs) collections for tomato to the community even before the release of the tomato genome. With the sequencing of the tomato genome, its annotation and other high-throughput data collections offering deep and comprehensive information for this plant species, the advancement of some existing or introduction of new collections providing manifold resources for this important crop was pushed forward prominently.

The *Solanaceae* Genomic Network (*SGN*) website [55] as a reference website is offering different resources and tools for the tomato genomics. Various resources such as the tomato genome sequence and its annotation (different versions), datasets for phenotyping, markers and maps, genes and pathways, and several other major collections are available on this reference website. It also includes some information and datasets from relevant plant species such as Potato, Eggplant, Tobacco and Arabidopsis to support the comparative

genomics. SGN also offers the combined results from *RNAseq* from unspecified collections in the form of short reads mapped onto the genome and accessible by a genome browser interface as coverage plots.

The Tomato Functional Genomics Database (*TFGD*) [135] is also a website specifically aimed to provide a representative resource for gene expression collections from tomato, including data from heterogeneous platforms (ESTs, microarrays, *RNAseq*).

To our knowledge, the Tomato Genomic Resources Database (*TGRD*) [136] is the other tomato related resource also providing the RNA-seq based expression of tomato genes in selected tissues (leaf, root, flower and fruit) from the *Heinz* reference collection [137] .

Major reference databases such as *NCBI, UniProt* and *RCSB* together with some tomato dedicated resources such as *SGN* [55] and *TFGD* [135] and *ProMEX* [138] offer proteomic data collections for this crop species.

In terms of the metabolomics and pathway information, besides of the tomato dedicated reference websites such as *SGN* [55] and *TFGD* [135] offering dedicated resources to this plant species, general metabolomics and pathway databases such as *KEGG*: kyoto encyclopedia of genes and genomes [139] and *Plant Metabolic Network (PMN)* [140] offer precious information to the resource community.

Availability of different tools such as *Genome Browse (Gbrowse) [141]* and *Integrative Genomics Viewer (IGV)* [142] could also enable the fast and easy investigation of different transcriptomic levels on the bases of a viewer to browse the genome with its "omics" annotation and content with some specific query input formats.

1.5 Thesis organization

This chapter is followed by other three chapters representing the materials and methods (chapter 2), results and discussions (chapter 3), and conclusion (chapter 4). Chapter 2 includes all the materials in this work. It also includes all the methodologies and implementations to setup our required resources and conduct our analyses. Chapter 3 represents the result of our effort introducing the major bioinformatics tools and platforms designed, and the results produced during this work. Eventually, chapter 4 provides a conclusion on the discussed topics highlighting the major key points in this effort. The cited references are all listed in the "*Reference*" section at the end of the thesis.

1.6 Summary

An introduction to the field of bioinformatics, its application and the impact it has on the biology area (specifically in plant) was provided in this chapter. In addition, the research line I followed during my PhD in the frame of the *SPOT-ITN* project, including the objectives and responsibilities, were introduced. Eventually, different relevant technologies, data types, methods and approaches as the prerequisite to better understand the foundation of this research topic was described and presented. The thesis organization and chapters content was also presented in brief.

2 Materials and Methods

2.1 Introduction

The materials and methods used in this thesis work are here presented. During the course of this PhD and at the light of the *SPOT-ITN* project, various public data collections relevant to our work were collected. In addition, several private collections from the *SPOT-ITN* partners were also considered to be organized in dedicated platforms. Different methodologies and approaches to collect, reconcile, and integrate these collections were also designed and implemented. The platforms and tools to store and analyze these data also to support our analytical objectives were also developed and built. Here, the description of these resources, the tools, platforms and methods to deal with these data are presented in details.

2.2 Experimental Design in the *SPOT-ITN*

In the specific framework of the *SPOT-ITN*, a leader experiment was considered for a common study from which different collections from *transcriptomics*, *epigenomics*, *proteomics* and *metabolomics* analyses were made available to the whole consortium.

Tomato plants *(S. lycopersicum cv. Red Setter)* were grown under controlled conditions in a glasshouse *(Agrobios; Metapontum, Italy)* and pollen samples were collected at three development stages—*tetrads* (T) (Pollen mother cells), *post-meiotic* (PM) stage (*microspores*) and mature stage (M) (*binucleate* pollen)— harvested according to the length of anthers (T: 4-6 mm, PM: 6-10 mm, and M: >10 mm). Three independent experiments were performed during

three consecutive days. Samples derived from one day were treated as biological replica. For heat stress (HS) experiments, HS plants were transferred in a preheated growth chamber and exposed to 38°C for 1 hour. The temperature was decreased to 25°C gradually within 30 minutes and plants were allowed to recover for an additional hour at 25°C. Untreated plants (control) were kept in the growth chamber for the same time period at 25°C.

GenXPro Company (Frankfurt Germany) provided the *MACE, MethSeq* and *Small-RNA* sequence collections from the common experiment, which was conducted in Metapontum (*Bari, Italy*). The partners at the Vienna University are in charge of the proteomic data (LCMS and GCMS) from the same samples. The partners from the *Wageningen* University are in charge of metabolomics data production (LCMS and GCMS) from the same sample, and its assignment to the corresponding pathways of interest.

2.2.1 Collection

MACE libraries were prepared as described by a protocol established by *GenXPro GmbH (Frankfurt, Germany).* We received 18 libraries of *illumina HiSeq* sequences for each of *MACE,* Small-RNA and *MethSeq* sequences (3 biological replicates for each of the Tetrad, Post-meiotic and Mature pollen developmental stages, each stage for 2 conditions of physiological and heat stress) cleaned from the adapters and barcode sequences, and low quality bases. Details on the sample preparation are presented in ANNEX III.

2.3 Genome Reference for Tomato

With attention to the tomato plant which is the focus of the *SPOT-ITN* project, and to set up a genome centric infrastructure to allow genome based analyses, the reference sequences for tomato were collected and analyzed. The reference

sequences were then charged into our infrastructure and used for any relevant investigation and analyses.

2.3.1 Chromosomes and BAC Sequences

Both versions of the 2.40 [137] and 2.50 [143] of the 12 pseudomolecules sequences representing the reference tomato genome, *S. lycopersicum* cv *Heinz*, together with the chromosome zero, which includes all the contig sequences not assigned to any other chromosome yet, were downloaded from the *SGN* website [55].

Version 2.40 is the most commonly used version before the newly released version of the 2.50. To understand the differences and peculiarities of the two genomes, a sequence based comparison between the two versions was carried out. The result of the comparison is presented in the chapter 3.

One hundred and twelve (112) BAC sequences not anchored along the reference genome (pseudomolecules) as independent resources which could be investigated in it gene content based on the different annotation tracks were downloaded from *SGN* website [55]. In fact, 1227 tomato BACs (among the 1338 in total) were anchored along the pseudomolecules.

2.4 Annotations for Tomato

To understand the genomic content of the tomato genome sequence, different annotations available for this plant species were collected. In some cases such as the tomato gene annotation, some efforts to improve its quality were also undertaken. First, the collections and datasets collected for tomato genome annotation are presented. Then the effort for revision of the tomato gene annotation and the effort to improve its quality will be described in details.

The official gene annotations of tomato, *iTAG 2.3* and *iTAG 2.4*, each representing 34,727 and 34,725 gene loci respectively provided by the *International Tomato Genome Consortium* (*ITAG*), was downloaded from *SGN* website [55].

The *RefSeq* 2.3 gene annotation representing 25,946 gene loci for tomato was downloaded from *NCBI RefSeq* database [13]. The initial downloaded files from *RefSeq* database were in *GenBank* format that were parsed and converted into gff3 format using an in-house parser.

The *SGN* infernal (*insilico* predicted small-RNA regions, produced by *Infernal* [144] software) version 2.3 and 2.4 were downloaded from the *SGN* ftp resource [55].

The *iTAG 2.3* and *iTAG 2.4* repeat annotations (normal repeat and repeat aggressive) representing the repetitive regions of the tomato *S. lycopersicum* genome were downloaded from the SGN ftp resource [55].

One hundred and ten (110) identified and known micro-RNA sequences for the Tomato *S. lycopersicum* were downloaded (April 2015) from the *MirBase* [85] database for the downstream analyses.

2.4.1 Efforts on the improvement of Annotation (Guided/Revised gene Annotation)

The *iTAG 2.3* predicted loci were intersected with the same annotation to detect the overlapping genes. They were also intersected with the *iTAG 2.3* repeat aggressive annotation to identify the genes predicted in the repeated regions. All the *iTAG 2.3* transcripts were also remapped along the tomato genome using *GenomeThreader* [49] (version 1.6.5), a software tool to compute gene structure predictions using a similarity-based approach via spliced alignments. Each mapped sequence was then compared and labelled according to the reference gene annotation. A blast versus the *UNIPROT* reviewed database (downloaded

on February 2013) was also conducted on the same mRNA collection. Further analyses on the resulting data revealed ambiguous or miss-located loci that will be discussed extensively in chapter 3.

Validation and Confirmation

The *iTAG 2.3* predicted loci were intersected with all the tomato ESTs, TCs and unigenes collections available in the Tomato Genome Platform organized within this effort. *RNAseq* expression signaling of each *iTAG 2.3* predicted loci was checked on the bases of *Heinz* expression collection available at *NexGenEx-Tom* [54] platform.

All the predicted *iTAG 2.3* loci were also intersected with the *RefSeq* 2.3 gene annotation. Each locus was then flagged with its type (Figure 13) and coverage of overlapping.

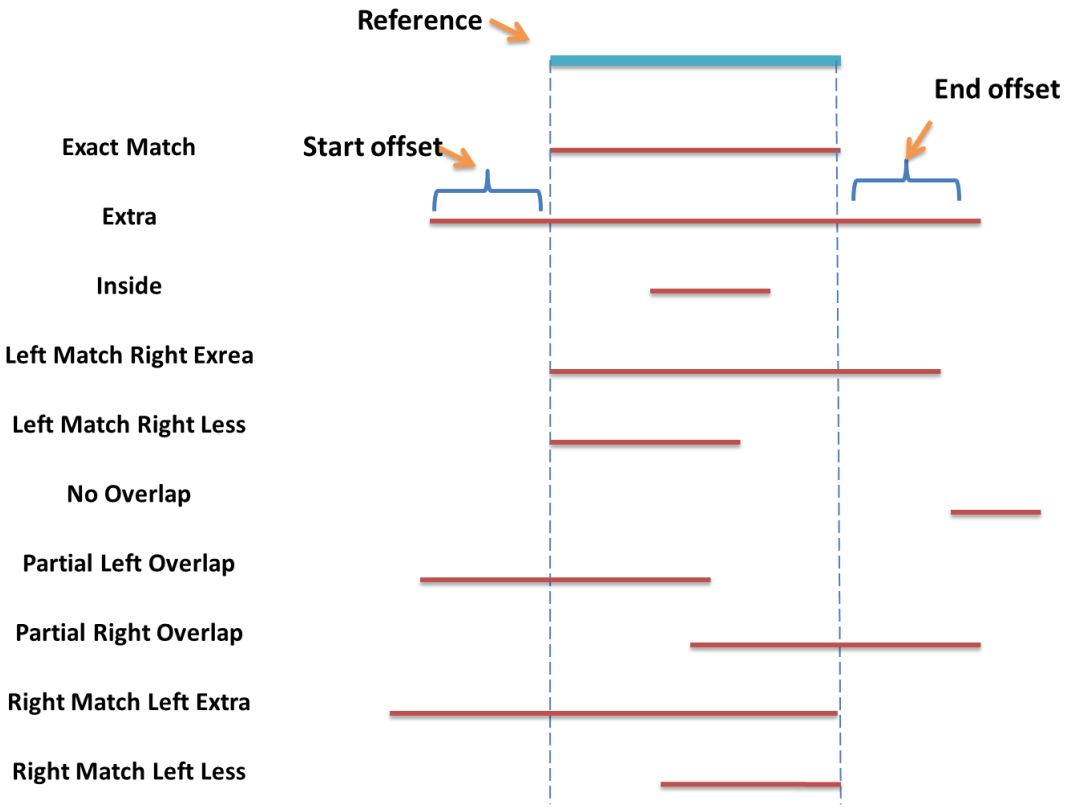


Figure 13: Different flags for labelling the different overlapping status of 2 genomic feature or transcript

As presented in Figure 13 all the possible overlapping types are defined by a flag. The percentage of coverage for the query with respect to the subject locus is calculated as follow:

$$c = \frac{q.l - so - eo}{s.l} \times 100$$

Where $q.l$ is the query length, so is the start offset, eo is the end offset and $s.l$ is the subject length.

Coding Process (mapping descriptor)

A combinational coding system was applied for summarizing the information into a short descriptor. The descriptor is a multi-part identifier explaining the features of a transcript, and the way it is mapped on the genome. As an example for the Gene ID *Solyc00g00500.1.1*, the ID "*Solyc00g00500.1.1_7M2*" or "*Solyc00g00500.1.1_7Mch02: 098555:095668*" can be produced. The methodology to assemble such tags is explained in Table 3, Table 4,Table 5,Table 6.

Table 3: transcript coding representation for the remapped sequences in multiple locations

Part1	Flag (Delimiter)	Part2
Number of total map versus the genome	"M" leter indicating mapping method and delimiter	Mapping time or position
7	M	2 or chr02:098555:095668

As presented in table 3 for the Gene ID "*Solyc00g00500.1.1_7M2*", the transcript *Solyc00g00500.1.1* has mapped seven times on the genome in which this locus (the one used as the example) is the 2nd locus of those mappings (because of the number 2 after character "M"). Using the tag including the chromosomal start and ends, the location of the transcript for this specific mapping position is provided.

Supplementary code:

To provide comprehensive information regarding the transcript, another combinational tag representing the general overview of the transcript is provided as follow:

Table 4: code letters used for the type of mapping (identity and coverage measure) for a transcript

Code Letter	Code Label	% Identity	% Coverage
A	Perfect Map	100	100
B	Good Map	95	95
C	Moderate Map	90	90
D	Map	90	80

As an example, *"Solyc00g00500.1.1_7M1_1A2B1C3D"* lists that this transcript with 7 mappings has 1 perfect map (A = 100% coverage and 100 % identity), 2 Good Maps (B = 95% coverage and 95% identity), 1 Moderate Map (C = 90% coverage and 90% identity) and 3 Maps (D = 90% coverage and 80% identity) on the genome. Extra flags can be added due to the combinational coverages and identities if required.

Binary Quality Identifier:

To specify the quality of each mapped transcript on the genome, a binary quality code was proposed to be placed in the quality column of the *GFF3* file. According to the coding presented before, a binary bit location based score is used to provide quality measures for all the transcripts. This coding helps to filter out those transcripts with specific quality and keep those of interest. The quality score can be also added to the identifier as another tag.

Table 5: binary representation of the quality for a transcript

Flag	A	B	C	D
Base	2^3	2^2	2^1	2^0
Value	8	4	2	1

As an example for the identifier "*Solyc00g00500_7M1_1A2B1C3D_8*", number 8 as the last flag is the indication of flag "A" which represents the 100% identity and coverage. This number will be stored in the quality column of the *GFF3* file to be used for the filtering purposes.

As the last flag to fulfil our ambitious goal for providing exhaustive information related to each transcript, the overlapping of the locus with other transcripts or being located in the void region (no overlapping with any other locus) is also added to the identifier.

As an example, for the identifier "*Solyc00g00500_7M1_1A2B1C3D_8_NO*", no overlapping with any other locus for this mapping is reported. The type of flags are presented as bellow:

Table 6: overlapping labels for a mapped transcript

Flag	Description	Example
NO	In void region (no Overlap)	_NO
OVP	Overlap with percentage (max percentage)	_OVP85

As presented in table 6, the overlapping status of a transcript when mapping on the genome can be presented. "*_NO*" represents that the transcript is not

overlapping any other locus while "*_OVP85*" represents that the query transcript has 85% of overlap with another locus.

We parsed these info into the *GFF3* file format to be easily used for the genomic tools and the associated analyses considering this multi-part ID as the locus identifier. The *GFF3* is the standard file format for representing the genomic features in text file (http://gmod.org/wiki/GFF3).

2.4.2 Joint annotation

Besides checking the remapping status of the tomato *iTAG* predicted genes to assess the quality of annotation, we also tried to take advantage of different resources for better flourishing of the tomato genome annotation. Here we present different methods for joining and complementing the two gene annotations available for the tomato genome (*iTAG* and *RefSeq*).

ITAG Preferred

With the aim of complementing *iTAG* annotation with the information available in *RefSeq*, the *RefSeq* loci not available in *iTAG* were extracted and added to the *iTAG* gene annotation. In the sense of those loci overlapping between the two annotations, the priority was given to the *iTAG* as the reference, and those of *RefSeq* overlapping any *iTAG* loci were discarded from the collection.

RefSeq Preferred

The same strategy was carried out to have a *RefSeq* reference based annotation complemented by *iTAG* predicted genes.

Joint (loose) annotation

Both annotations of *iTAG* and *RefSeq* were merged together in one unified GFF3 file.

2.5 Supportive Transcriptome Collections

To include comprehensive and exhaustive transcriptome resources for the tomato in our bioinformatics infrastructure, the following procedure was undertaken.

2.5.1 Expressed Sequence Tags

Twenty different EST collection (Table 7) were downloaded from the *GenBank* database [8]. Each collection was then subjected to the data processing defined in "*ESTs, TCs and* Unigenes *data processing*" part of this section.

Table 7: Expressed Sequence Tag (EST) collections stats

Species Name	Species Code	TaxonID	Starting # Sequences
Nicotiana tabacum	TOBAC	4097	334809
Solanum lycopersicum	SOLLC	4081	298370
Solanum tuberosum	SOLTU	4113	250127
Coffea arabica	COFAR	13443	174275
Capsicum annumm	CAPAN	4072	118651
Solanum melongena	SOLME	4111	98089
Coffea canephora	COFCA	49390	69066
Nicotiana benthamiana	NICBE	4100	56180
Petunia x hybrid	PETHY	4102	50705
Solanum torvum	SOLTO	119830	28743
Solanum habrochaites	SOLHA	62890	26019
Nicotiana langsdorffii x Nicotiana sanderae	NICLS	164110	12537
Solanum pennelli	SOLPN	28526	10946
Nicotiana sylvestris	NICSY	4096	8583
Solanum chacoense	SOLCH	4108	7752
Solanum phureja	SOLPH	172790	2099
Solanum lycopersicum X pimpellifolium	SOLLP	286530	1008
Capsicum chinense	CAPCH	80379	442
Nicotiana attenuata	NICAT	49451	355
Solanum peruvianum	SOLPE	4082	69

From here on, we refer to each species using the "*Species Code*" defined in the table above.

2.5.2 Tentative Consensus Collections

Twenty Tentative Consensus (TC) Collections resulting from the assembly of the EST collections described in "*ESTs, TCs and* unigenes *data processing*" of this section to create more reliable and extended sequences were also included in the platform. The singleton sequences not included in this collection, since they are available from the corresponding species EST collection.

2.5.3 Unigenes

Three *S. lycopersicum* collection of unigenes were downloaded from *SGN* [55], *Dana-farber* (*DFCI*) [57] and *PlantGDB* [56] websites each representing 42257, 52502 and 56845 transcript sequences respectively. Besides the sequence collections, the functional annotation of each collection was also downloaded to be charged into the platform. It is important to note that TCs and singletons are put together in these unigene collections.

2.5.4 ESTs, TCs and Unigenes data processing

Each of the twenty EST collections were cleaned from the vector sequences available at *NCBI's* Vector database (ftp://ftp.*NCBI*.nih.gov/blast/db/FASTA/vector.gz) downloaded on February 2013. They were also masked from the repeat sequences available in *RepBase* repeat database downloaded from http://www.girinst.org/ on February 2013. *RepeatMasker* software [145] was used for both vector trimming and repeat masking procedure.

To have the more reliable transcript collections, assembly of the EST sequences from each collection to create Tentative Consensus (TCs) was carried out using the *ParPEST* pipeline [16].

Each EST, TC and unigene collection was then mapped independently versus the *S. lycopersicum* 2.40 and 2.50 genome sequences and the un-mapped BACs (see 2.3.1) using *Genome Threader* [49] software (identity >= 0.90 and coverage >= 0.80). In term of the unigenes, the functional annotation provided by each data source downloaded from the reference website was used for the annotation of the loci.

Each of the ESTs, TCs and unigenes were then blasted versus the *SwissProt/UniprotKB* database (downloaded on February 2013) using 10E-3 and the first 10 best hits were collected for further functional investigations.

2.6 NGS data

To enrich our platforms with the NGS expression data, major available collections were collected, processed and charged into our infrastructure. Several private collections were also included in the platform to support specific analyses. Here we present the collection's properties and the procedure they underwent.

2.6.1 RNAseq

Different *RNAseq* data collections used in our analyses are presented in terms of species, SRA accession number, stages of plant, and the number of replicates.

Table 8: The RNAseq collection information representing the species, the number of replicates for each tissues/stages and the associated SRA accession number.

Species	SRA Accession No.	Stage	Number of Replicates
Tomato *Solanum lycopersicum* cv. *Heinz 1907*	SRP010775	Root	2
		Leaf	2
		Flower	2
		Flower bud	2
		1 cm fruit	2
		2cm fruit	2
		3 cm fruit	2
		Mature green fruit	2
		Breaker fruit	2
		Fruit after 10 days	2
Tomato *Solanum pimpinellifolium*	SRP010775	Leaf	2
		fruit at 5 days after the breaker stage	2
		Breaker fruit	2
		Immature fruit	2
Tomato *Ailsa Craig*	SRX098400	3' end sequencing	1
		Immature green fruit	4
		Mature green fruit	4
		Breaker fruit	4
		Fully ripe fruit	4
		5' end sequencing	3
Grapevine *Vitis vinifera*	SRP001320	Berry verison	3
		Berry ripening	2
		Berry post-fruit set	2

As presented in Table 8, the *Illumina HiSeq 2000* [146] RNA-seq data collection of *Solanum lycopersicum* cv. *Heinz* [137] including 20 libraries, each one representing one of two biological replicates from 10 different tissues and stages in physiological conditions; a collection from physiological conditions of *Solanum pimpinellifolium*, including 8 libraries [137]; and a collection from physiological conditions of *Solanum lycopersicum* cv. *Ailsa Craig*, including 20 libraries [147], were downloaded from *NCBI SRA* archive [15].

RNAseq Data Processing

Raw *Illumina* reads were cleaned from adaptor sequences and those with a quality lower than Q20 were discarded using *trim galore* [47]. Reads shorter than 20bp were also discarded. Filtered and cleaned reads were indexed by *bowtie2* [148] and mapped onto the Tomato pseudomolecules using *Tophat2* [52], with default parameters (up to 2 mismatches and intron length of 50 to 50,000 nt). Ambiguous matches were filtered out, i.e. reads with multiple matches on the genome were eliminated.

In terms of the differentially expression analyses, *DESeq* [79] a *Bioconductor R* package using negative binomial distribution and a significance threshold of false discovery rate (FDR) <= 0.05 was used.

2.6.2 SPOT-ITN Data Collections for Pollen

Three different NGS data collections of *MACE*, *Small-RNAs* and *MethSeq* from the three developmental stages of Pollen under physiological and heat shock stress conditions that was on the basis of *SPOT-ITN* experiment described in the section 2.2 were provided from the *GenXPro* Company (Frankfurt, Germany)

Table 9: the SPOT-ITN NGS data collections per techniques representing the number of replicates and conditions for each developmental stage

Library Type	Tissue/Stage	Number of Replicates	Conditions
MACE	Tetrad	3	Control & Heat Stress
Small-RNA	Post-meiotic	3	Control & Heat Stress
MethSeq	Mature	3	Control & heat Stress

MACE data analysis

The three stage/tissue/condition samples of *MACE* sequences *(Tetrad, Post-meiotic, and Mature)* were mapped versus the tomato genome (version *SL2.40*) [55, 137] using *Tophat2* (version v2.0.11) [52] considering the default parameters. All the reads mapped multiple times on the genome were discarded from the consequent analyses.

2.6.2.1.1 Annotation Based

Conventionally, using gene annotation (*ITAG 2.3*) [55, 137], the abundance of transcripts in each samples for each gene loci was calculated using *HTSeq-count* package [68]. *DESeq R packages* [79] was then used for the differentially expression analyses of the pair-wise tissue/stage/condition comparisons. To account for multiple testing, an FDR<0.05 was considered.

2.6.2.1.2 Annotation Free

Based on several issues we observed in the tomato gene annotation that will be discussed in details in chapter 3; and also due to the fact that a large number of reads were mapping on the void regions of the genome (regions with no gene annotation), we also decided to analyze the data in a different way independent

from the gene annotation. We developed a pipeline called *Tracker* (see **Error! Reference source not found.**), in which the detection of changing sites (e.g.: gene locus, methylation site, small-RNA cluster etc.) can be performed independent from the annotation.

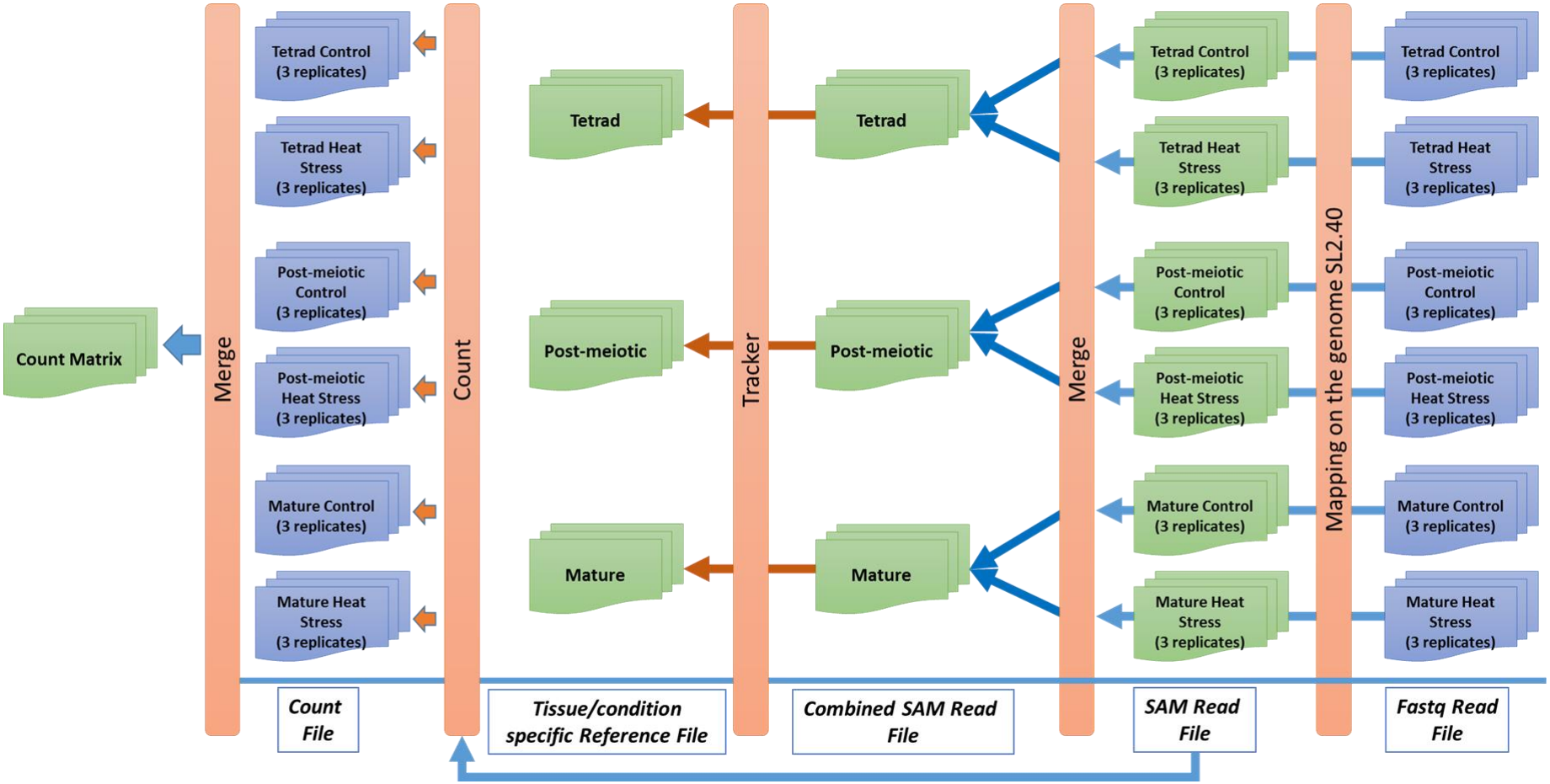


Figure 14: MACE differentially expression analyses using Tracker pipeline

After the mapping procedure, the tissues/stages/conditions that are supposed to be compared will be merged together (e.g.: all tetrad stages from control and heat together) (Figure 14). Using the *Tracker* pipeline, each merged collection is subjected to the cleaning (SAM file re-organization to make it a correct format for the consequent analyses), sorting (sorting the file based on the chromosome, start and end location), indexing (collapsing the identical reads in the sense of location, length, mapping type, quality etc.) and creating track references by collapsing the index files for the selected input. At the end, for each merged collection, a file including all the genomic regions with expression signaling (assembled if overlapping) is created in a tabular format. In this case, a reference annotation specific for the comparison of the merged tissues/stages/conditions will be provided for the quantification purposes. Eventually the counting on the basis of the corresponding reference annotation files for each of the replica for that stage is performed using the *HTSeq-count*. The results are then organized in a count matrix for the consequent analyses.

MethSeq data analyses

The *MethSeq* data was also analyzed on the basis of the CCGG sites and the *Tracker* based analyses. The procedure is defined as follow:

2.6.2.1.3 Annotation Based

The three stage/tissue/condition samples of *MethSeq* sequences were mapped to the tomato genome (*SL2.40*). For each genomic *HpaII* site (CCGG), reads starting at this position were quantified in each library. To account for multiple testing, the sites with FDR<0.05 were considered differentially methylated.

2.6.2.1.4 Annotation Free

As well as the annotation free analyses for the *MACE* data, the same procedure for the *MethSeq* data was considered. In this case, the methylation quantification is not always based on a single CCGG site, but if close enough (less than 100 nt neighboring), a cluster of adjacent CCGG sites will be considered for the methylation quantification (CpG islands detection and quantification).

2.6.3 Integration Process

Aligned with the main objective of the *SPOT-ITN* project to understand the mechanism implied in the heat stress during the pollen developmental stages in tomato, we further integrated the *MACE* and *MethSeq* data on the basis of our annotation free analyses. Using the *Overlapper (section…),* we intersected the detected regions differentially changing their expression and methylation status versus the each other. This supports the understanding of the expression and methylation mechanism when pollen is under heat stress during the developmental stages.

We also demonstrated the chromosomal distribution of the differentially expressed and methylated sites using the Map Chart [149].

The results of our annotation free analyses, and its application for the integration of *MACE* and *MethSeq* data to understand the mechanism of heat stress in tomato pollen will be presented in chapter 3.

Small-RNAs Data Analyses

Due to the availability of the Small-RNA collection in the frame of the *SPOT-ITN*, and the analyses I conducted for the paper "*The Role of TE-Derived Small Interfering RNAs in Tomato Pollen Development*", here we present the small-RNA analyses of the classes 21, 22 nt and 24 nt which can be expanded to other

size classes if necessary. The pipeline used in this analyses (Figure 15) is the adjusted version of the general Small-RNA pipeline I designed and presented in section 5.1.2

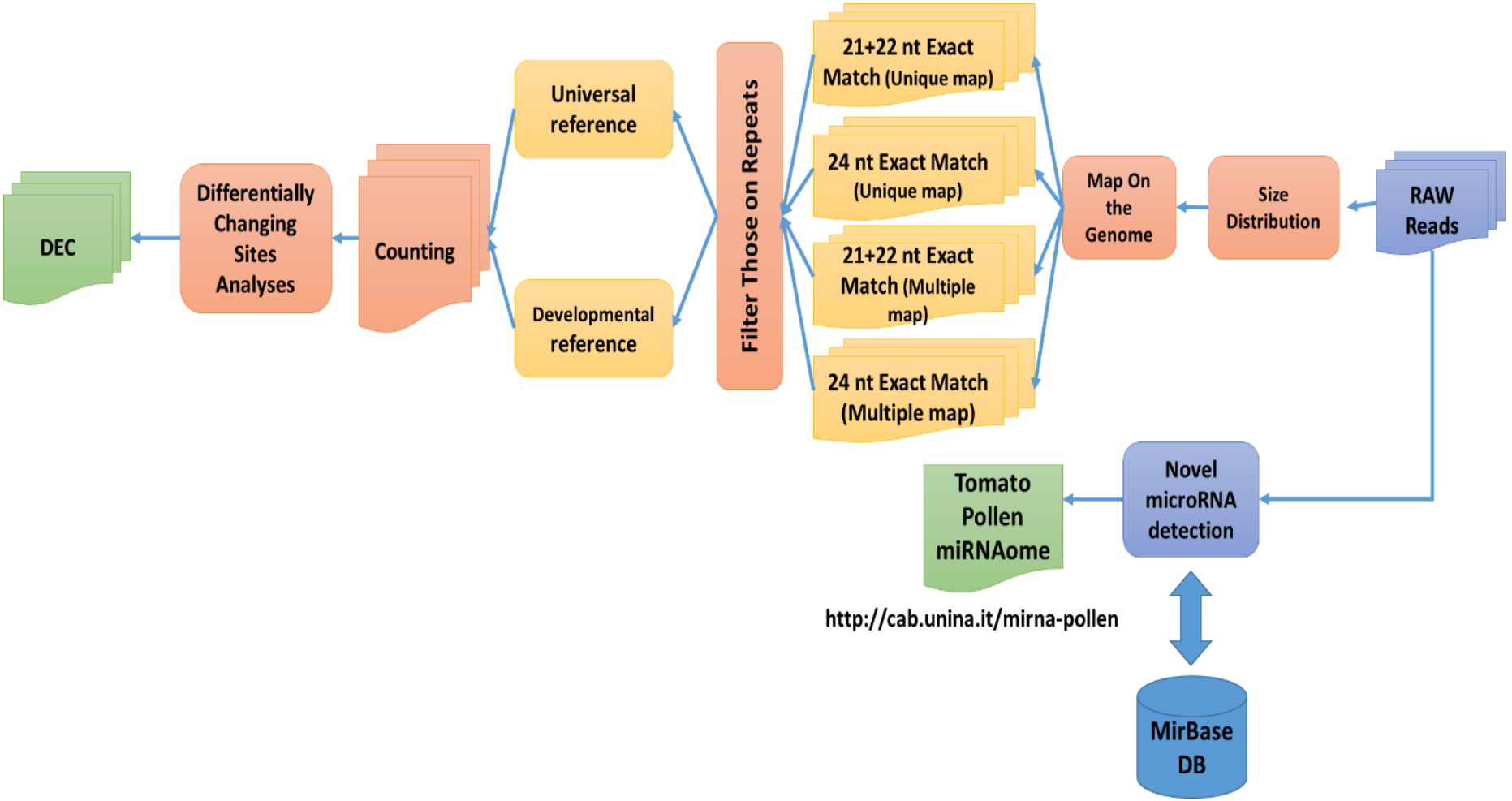


Figure 15: Small-RNA bioinformatics pipeline schema used for this analyses (DEC= differentially expressed clusters)

The Small-RNA sequences were mapped onto the tomato genome (version *SL2.40*) using *Tophat2* (version v2.0.11) considering the default parameters. Sequences with exact match were considered. To extract the *Small-RNA* sequences located in repeat regions, mapped reads were intersected with the *SL2.3 repeat aggressive* annotation. None overlapping (even 1nt overlap is considered) reads were not considered in downstream analyses.

A genomic clustering analysis was used to generate a reference of small-RNA clusters along the chromosomes, which was adapted from [32]. The mapped small-RNAs (21-22nt or 24nt separately) adjacent to each other (less than 100 nts) were clustered together in one group. Only clusters comprising more than one small-RNA read were considered for downstream analysis. The small-RNA abundance in a cluster (unique and multiple separately) was calculated using *featureCounts* [69] package for each of the Small-RNA classes (21-22 nt or 24nt) and samples at respective stages (Tetrad, Post-meiotic, and Mature). Differential expression analyses of small-RNA abundance in each cluster was carried out using *DESeq* [79] using the FDR <= 0.05.

Finally, the *MACE* data and *MethSeq* data were reconciled to investigate "The Role of TE-Derived Small Interfering *RNAs* in Tomato Pollen Development".

Micro-RNA Detection and analyses

The micro-RNA analyses of the sample was carried out by *GeneXPro* Company (Frankfurt, Germany). The results are organized in the *"Tomato Pollen miRNAome"* published at [150] and web accessible via (http://cab.unina.it/mirna-pollen) see 6.1 in ANNEX II.

2.7 Gene Ontology and Enrichment

Gene to GO Terms associations for tomato were defined by combining two major reference collections: i) the GO reference collection of *S. lycopersicum* downloaded from the *BioMart* database [102] (release June 2014), ii) the results of the *S. lycopersicum* mRNA sequences *Blast2Go* [26] versus the *NCBI* non-redundant database (nr). The two datasets were then combined removing duplicated terms. The GO annotations associated to the *iTAG* genes were uploaded into the devoted section of the platform.

2.8 Platforms

Aligned with my objectives in the frame of the *SPOT-ITN* project, and to set-up a genome centric bioinformatics infrastructure to properly organize and offer resources and tools for the genomic analyses, different platforms in a unified and integrated infrastructure was designed and implemented. Here, I present materials, methods and the architecture used to set up the major platforms and partitions of this collection.

All the platforms presented in this work are implemented in a three-tier architecture schema: 1) Data Tier; 2) Logic Tier; 3) Presentation Tier. The platform works as a web based application running on the .Net Framework 4.0, querying embedded databases, designed and organized in a relational model and implemented in *MySQL*, version 5.6.14 *InnoDB* engine [151]. All key fields and query dependent tuples were indexed using the *BTree* indexing algorithm [152].

2.8.1 Tomato Genome Platform

As presented in Figure 16, different resources for tomato were collected, processed and organized into dedicated databases. The platform includes

manifold resources crosslinked to each other providing several services in the form of query pages, web services and visualization interfaces. The tomato genome platform we present is a multi-level genome based infrastructure which is currently available under the *SPOT-ITN* Bioinformatics platform accessible via (http://cab.unina.it/*SPOT-ITN*-bioinfo/tracks/trck-search.aspx).

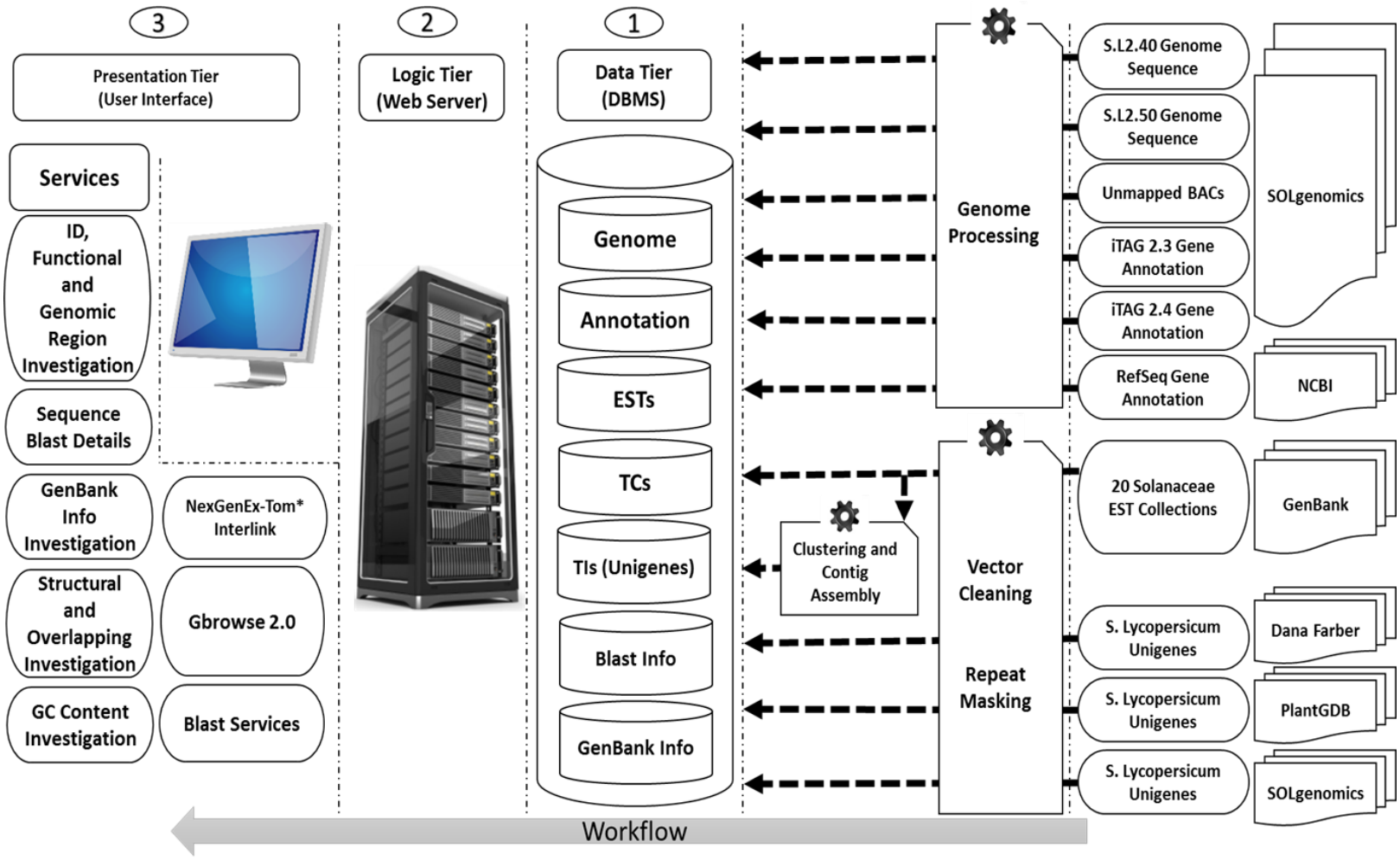


Figure 16: The tomato genome platform architecture, workflow and data processing schema

The platform includes the EST and TC collections from the 20 Solanaceae species (section 2.4), the three unigne collections for the *SGN*, *Dana-Farber* and *PlantGDB* together with their functional annotation (2.4.5), and gene annotations available for tomato *iTAG* and *RefSeq* (2.2.1 and 2.2.2) processed for both genome versions, the *SL2.40* and *SL2.50* of tomato were included in the platform.

The platform also is cross-linked to the expression platform which will be presented in the next section. It also implements a *GBrowse [141]* database and its dedicated interfaces.

2.8.2 Tomato Gene Expression Platform

We implemented *NexGenEx*- as a role based platform which enables the exploration of NGS based transcriptome collections. The platform was designed to provide enhanced tools for straightforward genome-wide gene expression analyses.

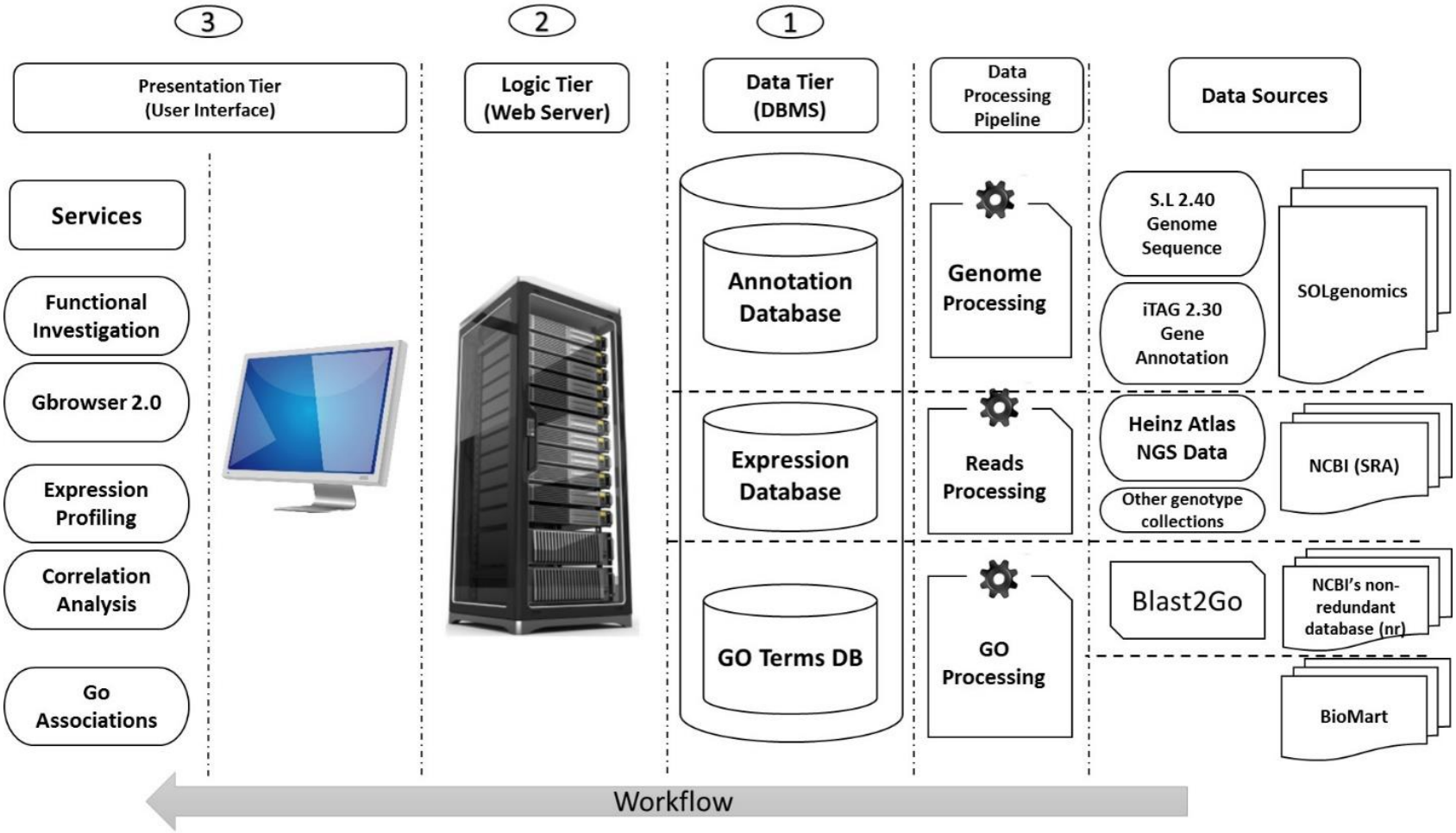


Figure 17: NexGenEx-Tom platform architecture, workflow and data processing schema

Figure 17 describes the main data processing pipelines necessary to define the data to be included in the platform. The three processed NGS collections from *S. lycopersicum* cv. *Heinz*, *S. pimpinellifolium*, and *S. Ailsa Craig* (presented in 2.6.1) are available for the gene expression investigation. The Go Terms collections (combined *Blast2Go* and *BioMart* redundant removed represented in section 2.5.5) for the genome version 2.40 of tomato were included into the devoted database. Several services and tools are provided in the platform.

2.8.3 Orthologs Platform

Here, the orthologs platform schema and its data processing pipeline is presented in details (Figure 18). The platform, at the current setting, includes three different collections (one public and 2 private) which are described as follow:

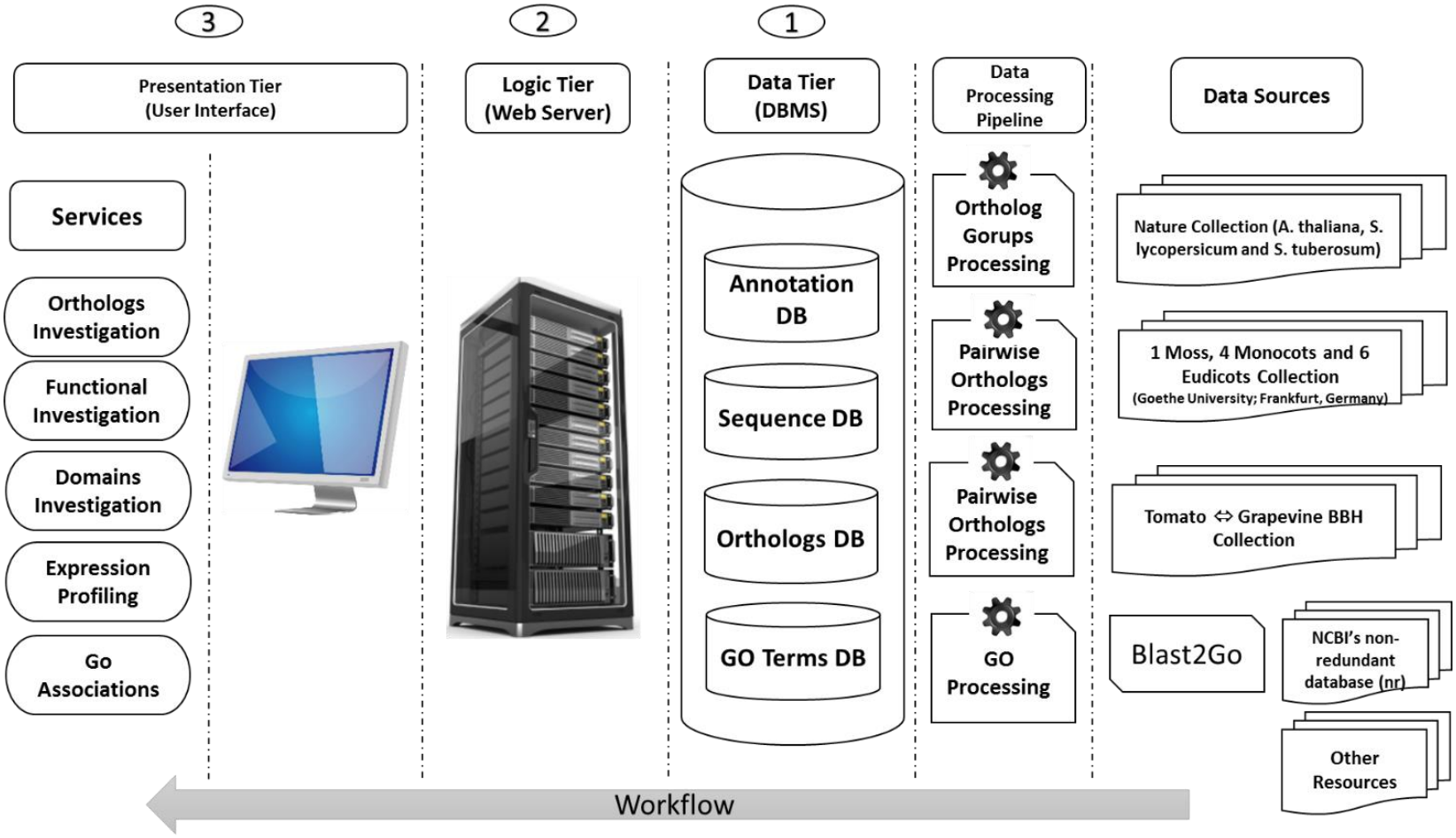


Figure 18 Orthologs platform architecture, workflow and data processing schema

Nature Collection

The orthologs collection released with the tomato genome sequence release [137] including 3 different species of *A. thaliana*, *S. lycopersicum* and S. *tuberosum* was considered.

Private Collection (Frankfurt, Germany)

An orthologs collection including *A. thaliana*, *S. lycopersicum* and another 11 different plant species, one moss, four monocots and six eudicots were received from Geothe University Frankfurt, Germany as a private collection (public upon publication) produced in the frame of *SPOT-ITN* project. The protein sequence collections are based on Phytozome v.9 [153] database.

2.9 Summary

As presented in this chapter, the materials used, and the methods to collect, process, reconcile and analyze them was presented in details. In terms of the bioinformatics platforms developed, the architecture used and the data sources included in each partition was also presented.

3 Results and Discussion: Platforms, data processing pipelines and applications

3.1 Introduction

This section of my thesis is divided into three sub categories in which the main results of my work during the PhD are presented. Initially, the major bioinformatics tools that I developed for the data processing and analyses are presented in brief. The presentation of the tools is not a manual neither an exhaustive discussion. In this section the introduction to the tool, the motivation and the idea behind it, and the advantages it offers are presented. In many of the analytical steps in my work, the tools and pipelines presented here are deployed to achieve the results. In the second section, the major bioinformatics platforms and databases designed and implemented (mainly in the frame of the SPOT-ITN) are presented in details. By providing snapshots and descriptions, here I try to demonstrate the functionality of the platforms and the services they offer. Eventually, the application of the methods and tools designed, the capability of the platforms, and the analyses conducted on different collections are presented for some example case studies.

3.2 Major Bioinformatics Tools Developed

3.2.1 Tracker

Motivation

In *NGS* data analyses, it often happens that the genome sequence is not well annotated or the reads mapped on the reference sequence refer to a location in which no feature is predicted. Moreover, in most cases due to the necessity of pairwise comparisons between read counts among specific tissues/stages/conditions of interest, having a customized reference genomic feature annotation provided by unexpected tags along the genome can be an advantage to trace interesting regions, since many of the reads map there and no annotated feature is described. To this end, the availability of a software applications allowing users to customize the feature description traced by the *NGS* data mapped on the genome can be helpful.

Description

The *Tracker* pipeline is a multi-level software application which allows the organization, cleaning, sorting, indexing and creation of reference tracks (expanding and collapsing the overlapping fragments into one reference regions, called a track, and keeping the trace files) of the mapped reads. The pipeline results into a report of several statistical information regarding each generated track. *Tracker* is written in *C#* under *mono IDE*.net and *Java* languages and runs under both *Windows* and UNIX environments.

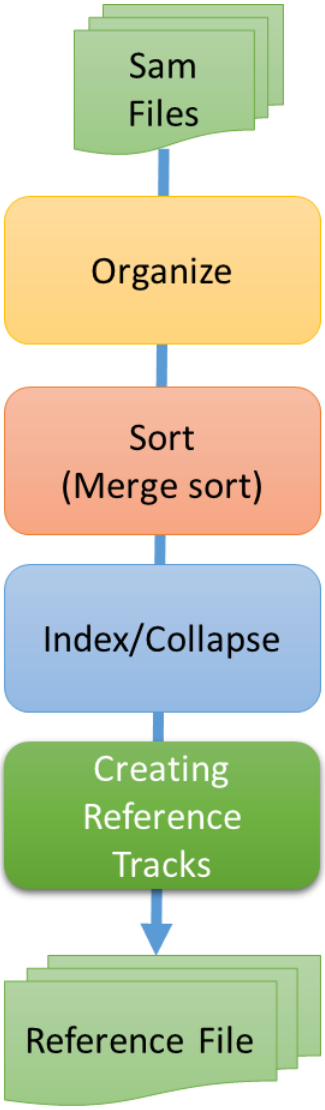
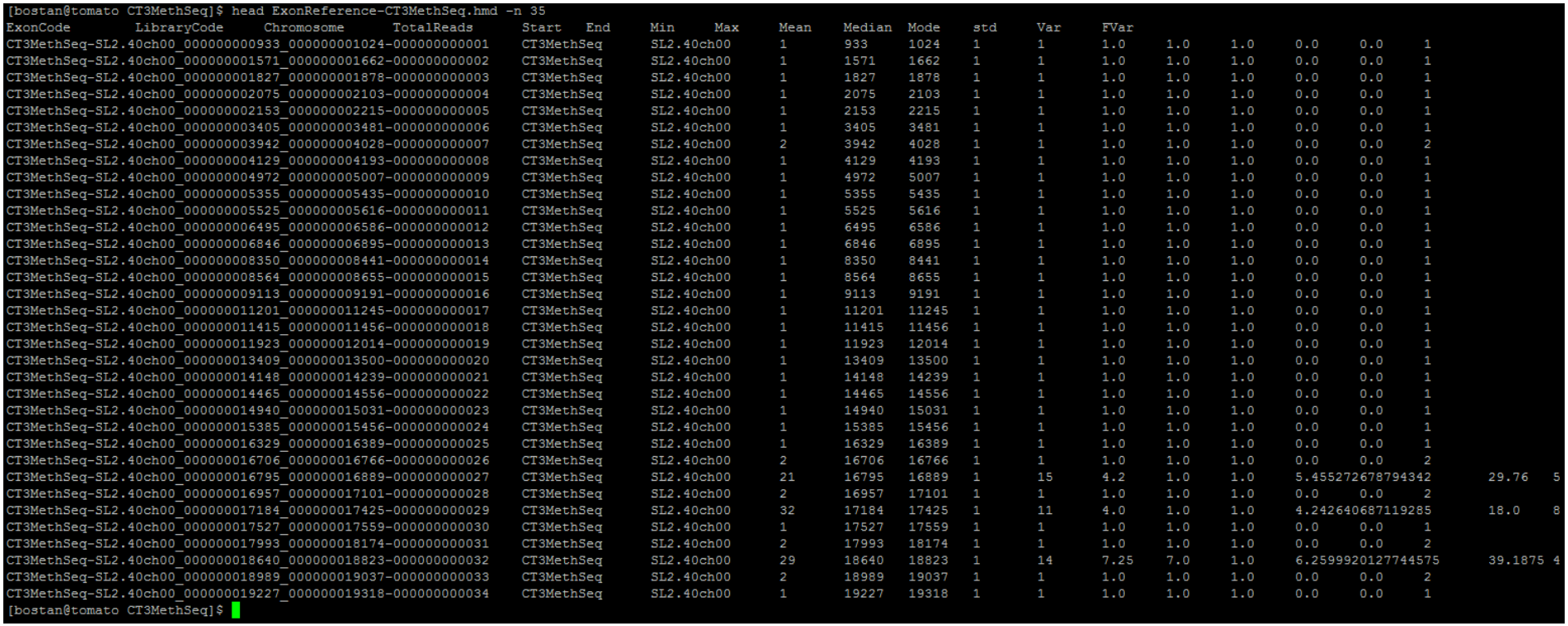


Figure 19: A brief workflow of Tracker pipeline with the possibility of intersecting the data for the modelling purposes

As presented in Figure 19: A brief workflow of *Tracker* pipeline with the possibility of intersecting the data for the modelling purposes, a general workflow of the *Tracker* pipeline is presented. The pipeline accepts as input file the sam (sequence alignment map) files [154]. Depending on the parameters specified as input arguments (-pr y/n: considering the read as one consecutive track or to divide it skipping the intronic regions), the tool aims to detect the splicing event and organize the reads into bins according to the *CIGAR* codes [18], i.e. the processed reads into fragments or the read files are sorted. The

sorted files is indexed (the identical overlapping fragments/reads with similar start, end, quality score, *CIGAR* code etc.) will be collapsed into one representative sequence keeping the trace of the collapsing sequences. Consensus sequences (putative reference tracks) are then created from the indexed sequences resulting from the previous step. A reference file including several statistical information for each track, such as length, total reads, min reads, max reads as the pick, average of frequency, standard deviation of frequency, median, variance and fragment variance of the frequency in each index, is also provided as a final output.



```
[bostan@tomato CT3MethSeq]$ head ExonReference-CT3MethSeq.hmd -n 35
ExonCode        LibraryCode     Chromosome      TotalReads      Start   End     Min     Max     Mean    Median  Mode    std     Var     FVar
CT3MethSeq-SL2.40ch00_000000000933_000000001024-000000000001   CT3MethSeq      SL2.40ch00      1       933     1024    1       1       1.0     1.0     1.0     0.0     0.0     1
CT3MethSeq-SL2.40ch00_000000001571_000000001662-000000000002   CT3MethSeq      SL2.40ch00      1       1571    1662    1       1       1.0     1.0     1.0     0.0     0.0     1
CT3MethSeq-SL2.40ch00_000000001827_000000001878-000000000003   CT3MethSeq      SL2.40ch00      1       1827    1878    1       1       1.0     1.0     1.0     0.0     0.0     1
CT3MethSeq-SL2.40ch00_000000002075_000000002103-000000000004   CT3MethSeq      SL2.40ch00      1       2075    2103    1       1       1.0     1.0     1.0     0.0     0.0     1
CT3MethSeq-SL2.40ch00_000000002153_000000002215-000000000005   CT3MethSeq      SL2.40ch00      1       2153    2215    1       1       1.0     1.0     1.0     0.0     0.0     1
CT3MethSeq-SL2.40ch00_000000003405_000000003481-000000000006   CT3MethSeq      SL2.40ch00      1       3405    3481    1       1       1.0     1.0     1.0     0.0     0.0     1
CT3MethSeq-SL2.40ch00_000000003942_000000004028-000000000007   CT3MethSeq      SL2.40ch00      2       3942    4028    1       1       1.0     1.0     1.0     0.0     0.0     2
CT3MethSeq-SL2.40ch00_000000004129_000000004193-000000000008   CT3MethSeq      SL2.40ch00      1       4129    4193    1       1       1.0     1.0     1.0     0.0     0.0     1
CT3MethSeq-SL2.40ch00_000000004972_000000005007-000000000009   CT3MethSeq      SL2.40ch00      1       4972    5007    1       1       1.0     1.0     1.0     0.0     0.0     1
CT3MethSeq-SL2.40ch00_000000005355_000000005435-000000000010   CT3MethSeq      SL2.40ch00      1       5355    5435    1       1       1.0     1.0     1.0     0.0     0.0     1
CT3MethSeq-SL2.40ch00_000000005525_000000005616-000000000011   CT3MethSeq      SL2.40ch00      1       5525    5616    1       1       1.0     1.0     1.0     0.0     0.0     1
CT3MethSeq-SL2.40ch00_000000006495_000000006586-000000000012   CT3MethSeq      SL2.40ch00      1       6495    6586    1       1       1.0     1.0     1.0     0.0     0.0     1
CT3MethSeq-SL2.40ch00_000000006846_000000006895-000000000013   CT3MethSeq      SL2.40ch00      1       6846    6895    1       1       1.0     1.0     1.0     0.0     0.0     1
CT3MethSeq-SL2.40ch00_000000008350_000000008441-000000000014   CT3MethSeq      SL2.40ch00      1       8350    8441    1       1       1.0     1.0     1.0     0.0     0.0     1
CT3MethSeq-SL2.40ch00_000000008564_000000008655-000000000015   CT3MethSeq      SL2.40ch00      1       8564    8655    1       1       1.0     1.0     1.0     0.0     0.0     1
CT3MethSeq-SL2.40ch00_000000009113_000000009191-000000000016   CT3MethSeq      SL2.40ch00      1       9113    9191    1       1       1.0     1.0     1.0     0.0     0.0     1
CT3MethSeq-SL2.40ch00_000000011201_000000011245-000000000017   CT3MethSeq      SL2.40ch00      1       11201   11245   1       1       1.0     1.0     1.0     0.0     0.0     1
CT3MethSeq-SL2.40ch00_000000011415_000000011456-000000000018   CT3MethSeq      SL2.40ch00      1       11415   11456   1       1       1.0     1.0     1.0     0.0     0.0     1
CT3MethSeq-SL2.40ch00_000000011923_000000012014-000000000019   CT3MethSeq      SL2.40ch00      1       11923   12014   1       1       1.0     1.0     1.0     0.0     0.0     1
CT3MethSeq-SL2.40ch00_000000013409_000000013500-000000000020   CT3MethSeq      SL2.40ch00      1       13409   13500   1       1       1.0     1.0     1.0     0.0     0.0     1
CT3MethSeq-SL2.40ch00_000000014148_000000014239-000000000021   CT3MethSeq      SL2.40ch00      1       14148   14239   1       1       1.0     1.0     1.0     0.0     0.0     1
CT3MethSeq-SL2.40ch00_000000014465_000000014556-000000000022   CT3MethSeq      SL2.40ch00      1       14465   14556   1       1       1.0     1.0     1.0     0.0     0.0     1
CT3MethSeq-SL2.40ch00_000000014940_000000015031-000000000023   CT3MethSeq      SL2.40ch00      1       14940   15031   1       1       1.0     1.0     1.0     0.0     0.0     1
CT3MethSeq-SL2.40ch00_000000015385_000000015456-000000000024   CT3MethSeq      SL2.40ch00      1       15385   15456   1       1       1.0     1.0     1.0     0.0     0.0     1
CT3MethSeq-SL2.40ch00_000000016329_000000016389-000000000025   CT3MethSeq      SL2.40ch00      1       16329   16389   1       1       1.0     1.0     1.0     0.0     0.0     1
CT3MethSeq-SL2.40ch00_000000016706_000000016766-000000000026   CT3MethSeq      SL2.40ch00      2       16706   16766   1       1       1.0     1.0     1.0     0.0     0.0     2
CT3MethSeq-SL2.40ch00_000000016795_000000016889-000000000027   CT3MethSeq      SL2.40ch00      21      16795   16889   1       15      4.2     1.0     1.0     5.455272678794342       29.76   5
CT3MethSeq-SL2.40ch00_000000016957_000000017101-000000000028   CT3MethSeq      SL2.40ch00      2       16957   17101   1       1       1.0     1.0     1.0     0.0     0.0     2
CT3MethSeq-SL2.40ch00_000000017184_000000017425-000000000029   CT3MethSeq      SL2.40ch00      32      17184   17425   1       11      4.0     1.0     1.0     4.242640687119285       18.0    8
CT3MethSeq-SL2.40ch00_000000017527_000000017559-000000000030   CT3MethSeq      SL2.40ch00      1       17527   17559   1       1       1.0     1.0     1.0     0.0     0.0     1
CT3MethSeq-SL2.40ch00_000000017993_000000018174-000000000031   CT3MethSeq      SL2.40ch00      2       17993   18174   1       1       1.0     1.0     1.0     0.0     0.0     2
CT3MethSeq-SL2.40ch00_000000018640_000000018823-000000000032   CT3MethSeq      SL2.40ch00      29      18640   18823   1       14      7.25    7.0     1.0     6.2599920127744575      39.1875 4
CT3MethSeq-SL2.40ch00_000000018989_000000019037-000000000033   CT3MethSeq      SL2.40ch00      2       18989   19037   1       1       1.0     1.0     1.0     0.0     0.0     2
CT3MethSeq-SL2.40ch00_000000019227_000000019318-000000000034   CT3MethSeq      SL2.40ch00      1       19227   19318   1       1       1.0     1.0     1.0     0.0     0.0     1
[bostan@tomato CT3MethSeq]$
```

Figure 20: Tracker sample reference tracks output file reporting the track id, track type, genomic location of each track, total reads, total length, min reads, max reads as the pick, average of frequency, standard deviation of frequency, median, variance and fragment variance of the frequency in each index.

Example usage: Gene Annotation and Revision

The *Tracker* pipeline can be used for several purposes in the NGS data analyses. Here, an example protocol pipeline for validation and revision of gene annotation supported by NGS data is presented.

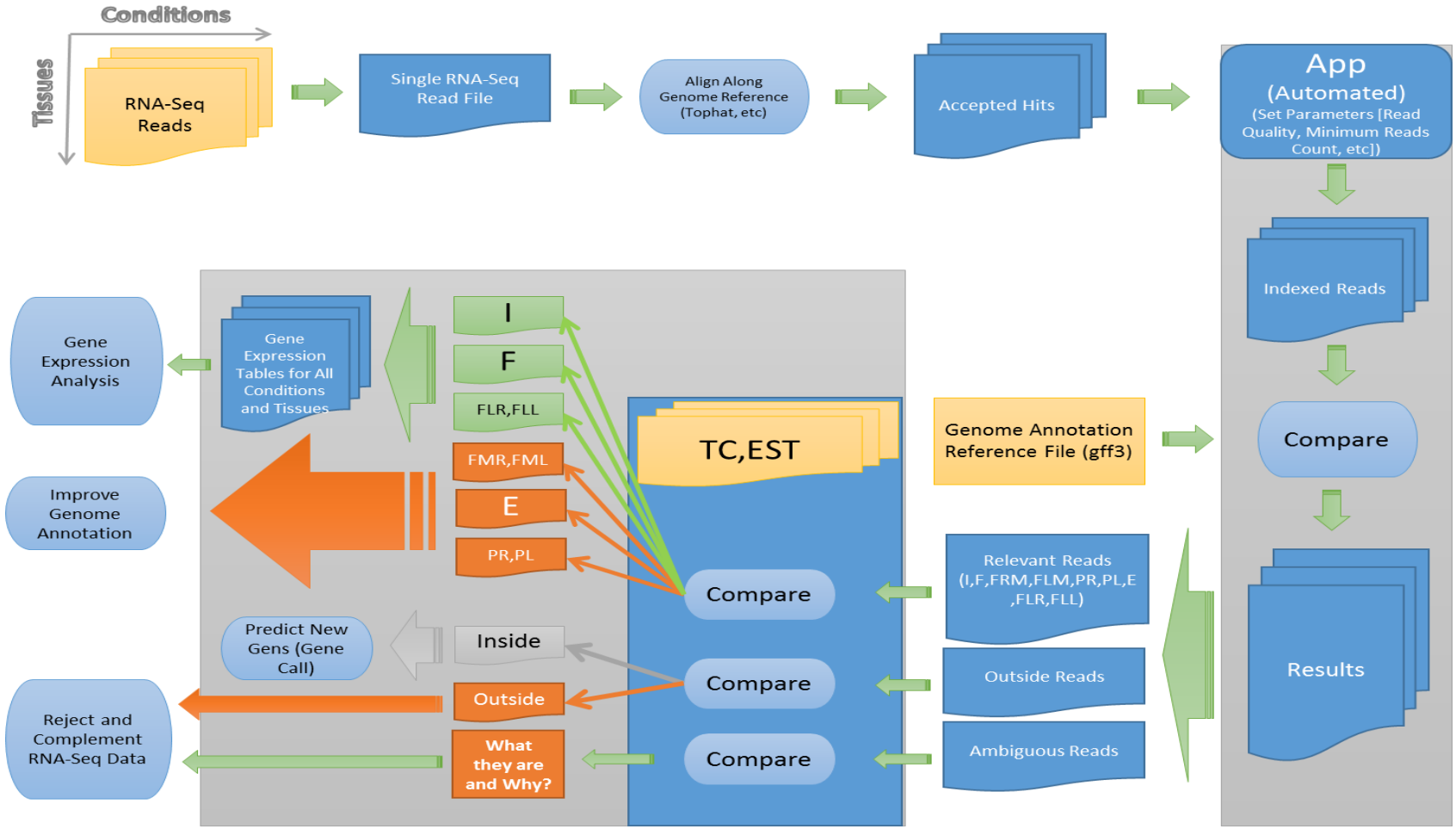


Figure 21: Validation, detection and correction of genes annotation

As presented in Figure 21: Validation, detection and correction of genes annotation, the track references are built from the *RNAseq* data files for the desired tissue/stage/condition. The reference regions (indeed representing possible exons since the reads come from transcripts) are then cross checked with the official gene annotation, available experimental transcript collections (EST, TC, Unigenes) for the specific species and the transcripts/genes lying in the same regions, if available. The majority of the exon references should normally confirm the exons from the gene annotation or transcripts experimental defined by ESTs, TC or Unigenes. Those newly detected tracks with differences from the official gene annotation can help to confirm the current gene annotation. Those not overlapping with any annotated genomic feature but having wide representation from the reads (number of reads contributing to the tracks and length) can be considered as putative novel exons or genes not yet predicted in the current annotation.

The pipeline can be also used for the annotation free analyses. An example application of this approach is presented in sections 2.6.2.1.2 and 2.6.2.1.4.

3.3 Contiger

Motivation

As it was discussed in the "Tentative Consensus and Assembly section…" generation of contigs or consensus sequences from the overlapping genomic features is a necessity in many bioinformatics analyses. There are tools such as *ClusterBed* implemented in *BedTools* package [128] able to define clusters of genomic features mapped on the genome. Though such tools already exists, the need for having more efficient tool with the possibility of keeping the different trace files and also transferring of all the info from the input file to the output file made us to develop *Contiger*.

Description

Contiger is a tool in which the possibility of assembling the overlapping genomic features or the adjacent tracks (specifying the neighboring distance) can be done in a fast way. Moreover *Contiger* is efficient with the memory consumption since the number of records per memory can be specified as an input parameter. It keeps the trace of assembled features into a parent feature including the coverage index, and the information available in the input file are all transferred into the output for easier and more efficient tracing purposes. The tool works with any tab delimited file containing the chromosome name (or reference sequence identifier) and genomic start and end positions.

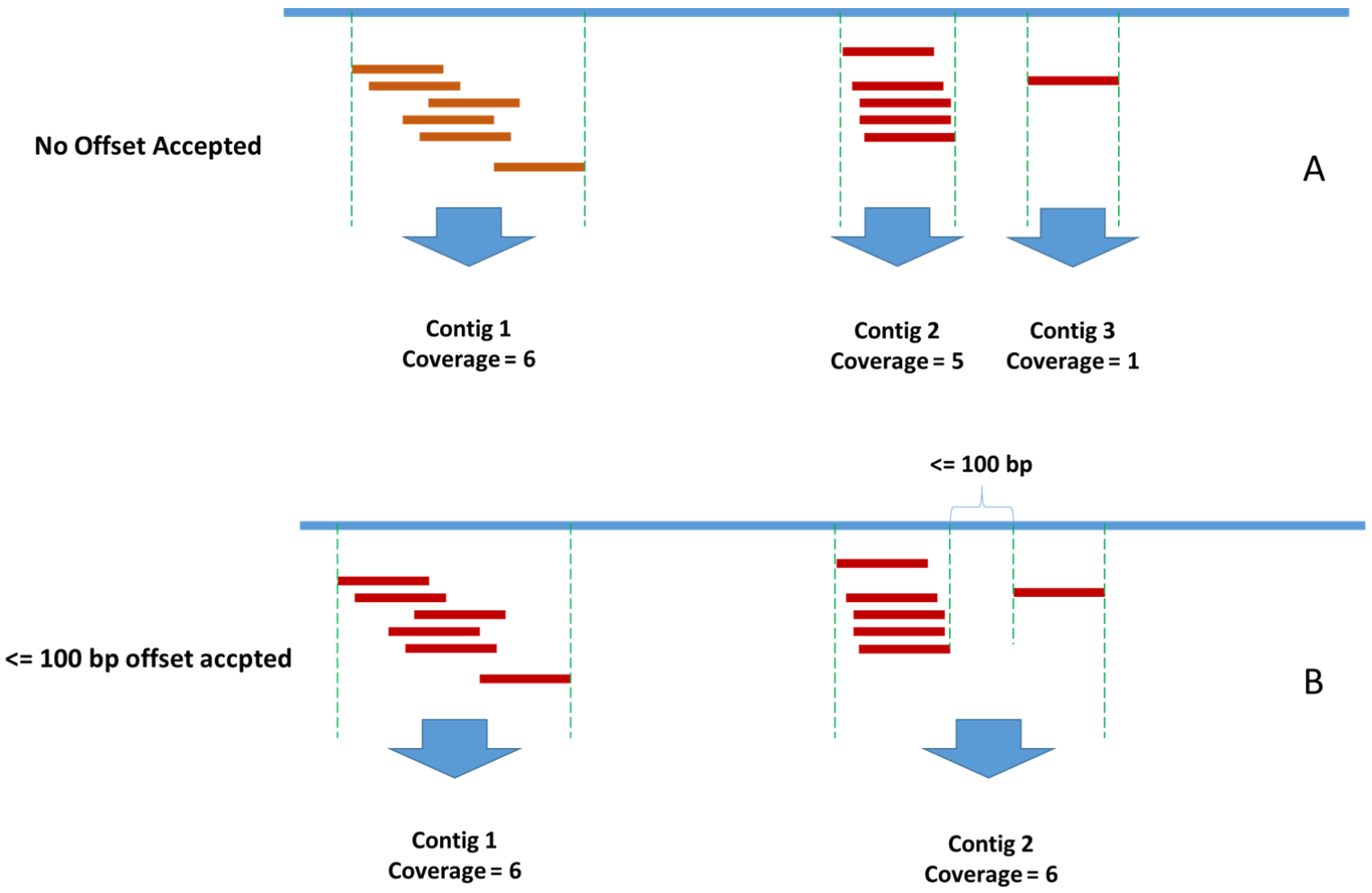


Figure 22:The schema of contig generation by Contiger

Figure 22:The schema of contig generation by Contiger represents the two type of contig generation A) only for the overlapping and B) using the offset distance to create cluster of features. The tool is written in *C#* programming language under *mono IDE* and can be run under both *Windows* and *UNIX* environments.

3.3.1 Overlapper

Motivation

In transcriptome data analyses, the overlapping of different tracks or features can be important to draw useful conclusions. As an example, targeting of different micro-RNA sequences on the coding region, or the overlapping of the methylated sited on transposon elements or tracing a specific distance from the gene transcription start site (TSS) requires the overlapping of thousands or millions (in case of NGS data) of genomic regions versus each other. This requires appropriate tools to carry out the most appropriate intersection.

Description

Overlapper is a tool allowing the intersection of different genomic features (two collections per time). . Due to the sorting algorithm implemented in the tool (merge sort), it can handle bulky files in a memory efficient routine (user can define the load of memory) with the possibility of defining offsets for each start or end position. Possibility of choosing the specific features (e.g. only those records representing mRNA or exon feature etc.) to be compared is another advantage implemented in this software which also reduces the time of processing while discarding the non-relevant records.

The software produces three output files. The log file summarizes the procedure and the overlapping statistics for the whole analyses. The Flags file providing the details information on the overlapping features including all the info from

the source collections together with the type of overlapping and the percentage of their overlap. The Overlaps file provides a summary of each input genomic region with the number of overlaps (if any, otherwise zero). *Overlapper* does not require any specific file format or order in the columns of the input file as far as it is tab delimited and includes the genomic locations (chromosome, start and end) of the track sequences. It transfers all the genomic information from the starting files to the output. It can also summarize each of the overlapping separately together with a summary table indicating how many overlaps a genomic feature intersects. The tool is written in *c#* under *mono* IDE and can run under both *Windows* and *UNIX* environments.

3.3.2 RNAseq Analyses Pipeline

Motivation

In NGS data analyses, adapter removal, trimming of low quality bases, mapping of the short reads on the genome and counting of the mapped reads overlapping a genomic feature are the routine and fundamental approaches to carry out before the differential expressed genes analyses [60]. Normally each experiment contains several stages/tissues/conditions including different number of replicates for each one. In some cases, different adapters and barcoding sequences are used for each replicate. In general, biologists use bash files to automatize the process, but this requires all the parameters should be set one by one in each command line for each step. This can elongate the time required for the analyses drastically and reduce the precision due to human errors. On the other hand, the work should be done in a sequential way unless the commands are distributed manually on the cluster or grid machines.

Description

To facilitate the NGS data analyses procedure and automatize all the steps from raw data to the expression count matrices, we developed a pipeline which implements different external tools (such as trim galore for trimming [47], *cutadapt* for adapter sequence removal [155], *tophat2* [52] and *bowtie2* [148] for the short reads mapping and Htseq-count for the features counting [68]) for each step in parallel to scrounge the time and increase the accuracy. The pipeline also allows the possibility of filtering out ambiguous reads (multiple mapped on the genome or overlapping with more than one genomic feature) during the process. The pipeline can manage both single and paired-end libraries with the possibility of having different adapter sequences for each individual replicate. It also produces different output files such as the complete SAM [154] file (from the *Tophat* bam file), the bam and index bam files to be directly used for the NGS data visualization in a genome browser [141] like platform, and the sam file cleaned from the ambiguous reads.

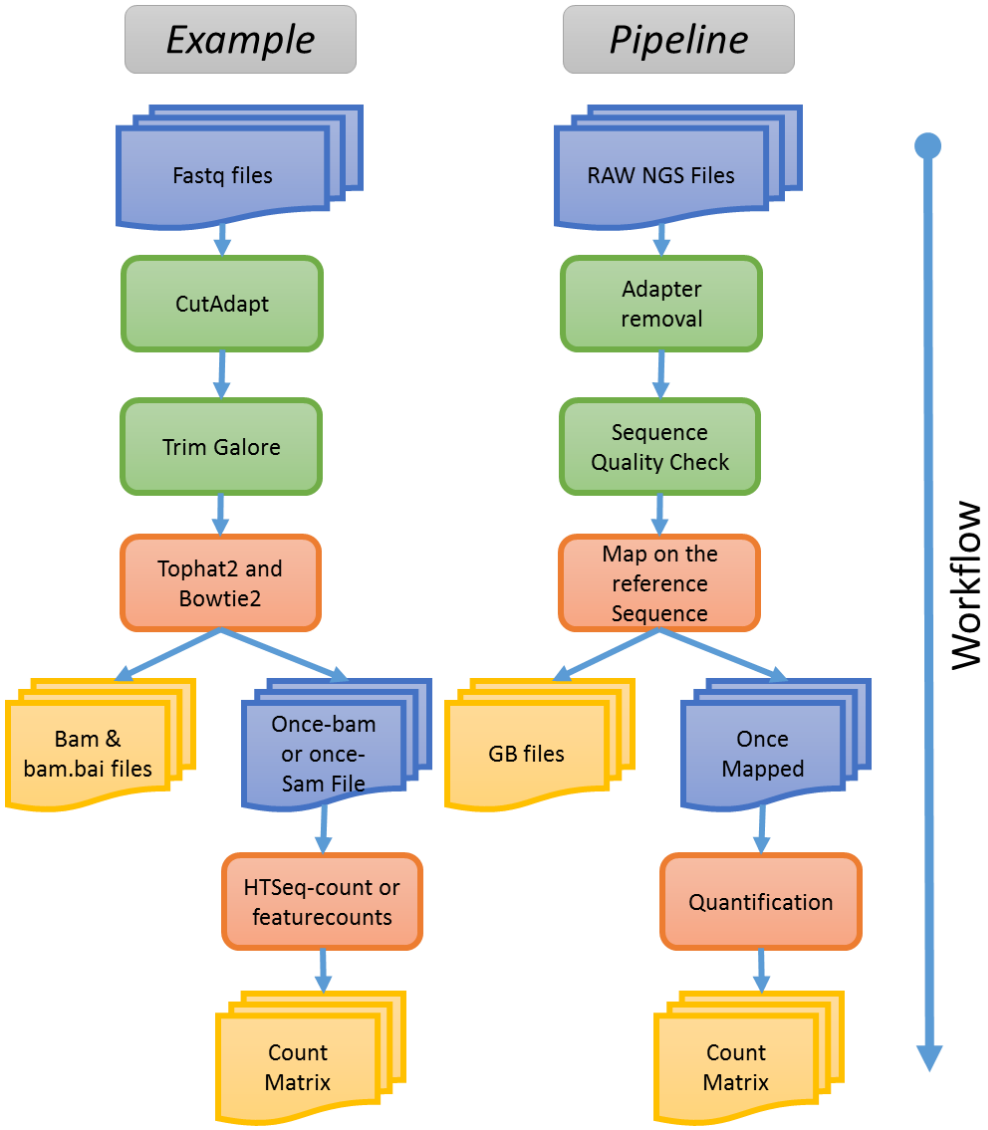


Figure 23: The automated and parallelized RNAseq processing pipeline schema

The tool has been written in *c#* under *mono* IDE and can be run under both *Windows* and *UNIX* environments.

3.3.3 Differentially Expression Analyzer

Motivation

Detection of the genes differentially expressed or suppressed between two conditions is one of the common practice to highlight information useful to understand the mechanisms underlying a biological process [156]. There are several tools and packages (such as *DESeq* [79], *EdgeR* [157] etc.) developed for the DEG analyses which allows the comparison of two different conditions to identify the genes significantly changing the expression level. Most of these packages are implemented in an R environment and require basic knowledge of the R scripting language to conduct the comparisons. Moreover, the need to analyses different combination of pairwise comparisons independently requires time and human work, or in the best case, proper bash files scripting for each set of comparisons.

Description

To facilitate the Differentially Expression Analyses for multiple pairwise comparison, we developed a user-friendly *Windows* based application allowing to analyze multiple pairwise comparisons in an automatized way to enhance the procedure and decrease the human effort. The software application allows the pairwise comparison of a list of conditions (each can have one to several replicates) to be processed in an automatized way with different filtering possibilities implementing both *EdgeR* and *DESeq* packages. Aside from the list of DEG genes, it also creates several plots such as expressed genes Venn diagrams for each comparison, MA plots, Depression plot, box plot of the expressions.

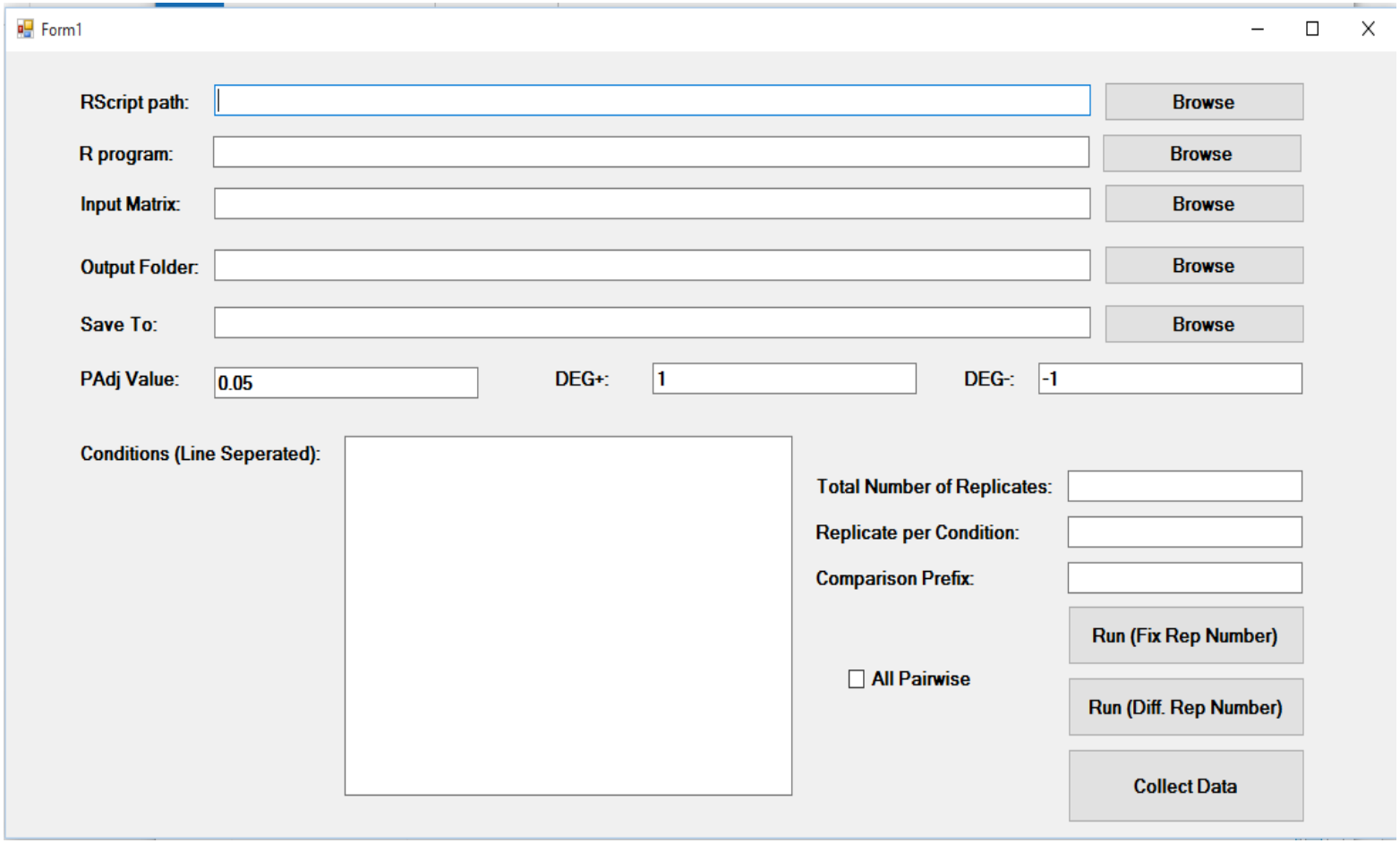


Figure 24: Snapshot of the DEG analyses Windows application interface allowing th epairwise comparison of different stages/tissues/con ditions in an automated and efficient way

Snapshot of the differentially expression analyzer is presented in Figure 24: Snapshot of the DEG analyses *Windows* application interface allowing th epairwise comparison of different stages/tissues/conditions in an automated and efficient way. The software is using *RScript* from R environment and a dedicated R package that their path should be specified in the tool.

It also requires an expression matrix representing the expression levels for the genomic features for different conditions (1.2.5.1.4). User should choose the output directory where the results should be saved, the path to save the running script, the FDR value to be used for the multiple testing correction, and the list of conditions in the expression matrix.

1- If the number of replicates are fixed for each tissue/stage/condition, users can specify the total number of replicates and the number of replicates per each condition to accelerate the process.
2- If the number of replicates per tissue/stage/condition are not fixed, the number of replicate per each tissue/sample/condition should be specified in front of the tissue name in the condition box.

In case the user wants to run all the pairwise comparisons for the available tissues/stages/conditions, the pairwise checkbox should be checked before running the analyses.

The final files (DEG list) can be collected using the "collect data" button to accelerate the process. Results will be reported separated together all in one file flagged by the comparison names.

The tool is written in *c#* under *mono* IDE and can run under the *Windows*.

3.4 Platforms

With the aim of setting up a multi-dimensional (multi-genome including different "omics" data levels) genome reference based computational suit,

different omics data collections and the result of the analyses conducted on them were charged into dedicated databases allowing to access to these processed information via user-interfaces. Different platform sections offering several query pages and online tools to easily explore and exploit the available data were developed.

The working version of this multi-level infrastructure is currently implemented as the *SPOT-ITN Bioinformatics platform* (http://cab.unina.it/*SPOT-ITN-bioinfo*) enriched with several public and private collections to support the objective of the project described at 1.1.

Here, the application and utility of the major platforms implemented in this infrastructure are presented in details.

3.4.1 Genome Platform

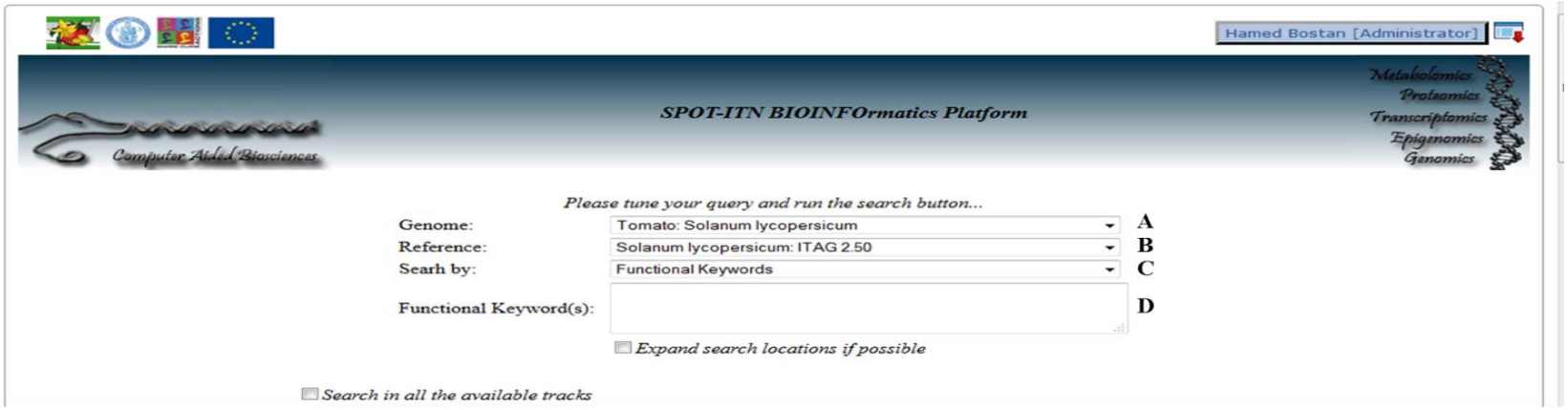
We setup a genome based computational environment to organize multilevel data for the tomato transcriptome. The platform currently includes the genome and transcriptome levels. The platform is set-up on the basis of both versions of the tomato genome reference sequences and their associated *iTAG* and *RefSeq* gene annotations. It also includes the EST and TC from different Solanaceae species, all the available unigene collections for tomato, and their functional annotation. It also entails several other annotation tracks available for tomato (see 2.4). The tomato genome platform offers a cross link to the *NexGeneEx-Tom [54]* database allowing the gene expression investigation and profiling. It allows straightforward and comprehensive genomic center investigations on high quality data resources using several advanced user-interfaces. A *Gbrowse* [158] database and associated interface are also embedded in the platform. The Expression data from the collections available in the *NexGenEx-Tom* are also available in the implemented *Gbrowse* database for further gene expression profiling and visualization purposes.

Tomato Genome Platform

A Genome Platform, for the tomato, was implemented allowing extensive transcriptome data investigation in the tomato genome space. The platform collects different resources and reconcile high quality data in a genomic center infrastructure in which, different transcriptomic data levels can be independently or collectively investigated and visualized. Cross comparison between different transcriptional datasets and levels is easily possible using the platform interfaces and the *Gbrowse* plugin implemented in the platform.

3.4.1.1.1 User Interface and Database access

A Graphical User Interface (GUI) is designed to provide access to the resources available in the genome platform. The query page in the genome platform can be tuned with different options to facilitate the user's investigation (e.g.: choosing different genome versions, choosing different transcript collections to be investigated, searching by "*Gene ID(s)*", "*functional keywords*", "*genomic region*", and "*Protein ID*").

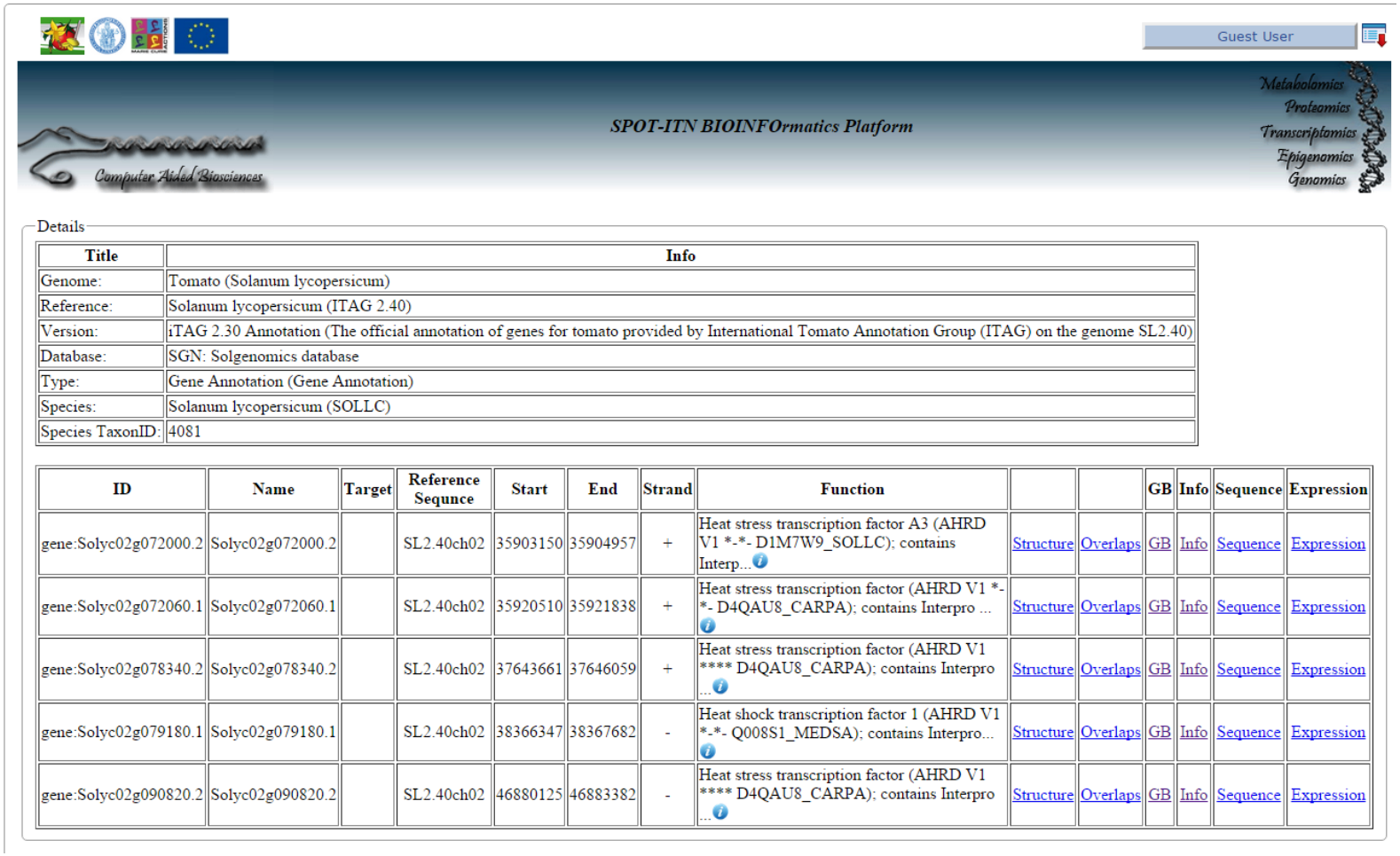


E	Gene Annotation	Tentative Consensus	Tentative Consensus	Tentative Consensus	Express Sequence Tag	Express Sequence Tag	Express Sequence Tag	Express Sequence Tag	Express Sequence Tag	Express Sequence Tag	Express Sequence Tag	Express Sequence Tag	Express Sequence Tag
	Solanum lycopersicum	Solanum lycopersicum	Solanum tuberosum	Solanum melongena	Solanum lycopersicum	Solanum tuberosum	Capsicum annumm	Nicotiana tabacum	Solanum melongena	Solanum torvum	Solanum pennelli	Solanum chacoense	Nicotiana benthamiana
	SGN	SOLEST	SOLEST	SOLEST	SOLEST	SOLEST	SOLEST	SOLEST	SOLEST	SOLEST	SOLEST	SOLEST	SOLEST
Chromosome02	5	10	7	0	107	53	5	5	0	0	0	0	0
Chromosome03	3	22	12	5	97	169	0	0	10	10	0	0	0
Chromosome04	2	5	0	0	15	26	0	0	0	0	0	0	0
Chromosome06	2	25	0	0	111	20	5	0	0	0	0	0	0
Chromosome07	2	10	5	0	61	15	0	5	0	0	0	0	0
Chromosome08	4	25	5	0	99	35	0	0	0	0	5	0	0
Chromosome09	4	10	5	0	132	56	0	0	0	0	5	5	0
Chromosome10	1	0	0	0	0	0	0	0	0	0	0	0	0
Chromosome11	1	0	0	0	0	0	0	0	0	0	0	0	0
Chromosome12	3	10	10	5	235	225	0	0	10	2	0	0	5

Figure 25: The snapshot of the genome platform query page together with the sumamry of results for a query (here HSF keyword in all the available track collections mapped on the tomato genome version 2.50 for all)

As it is shown in Figure 25, A) the genome reference species (e.g.: tomato, potato etc.), B) the version of the genome (e.g.: *SL2.40*, SL2.5 etc.), C) the query type (by ID, functional keyword, genomic region, protein ID) can be specified to run the query (D). Depending on the collections available for each genome reference and version, a list of tracks will be available to be chosen for the investigation. The tomato genome platform is enriched with the annotations for the tomato (section 2.2) and the supportive transcript collections (section 2.4), and the available expression data (2.4.1, 2.4.2, and 2.4.3) in the genome browser.

By tuning the mentioned parameters and running the query, a chromosomal distribution of the hits matching the users query categorized by each collection will be summarized in details (Figure 25). Detailed information for the hits found can be obtained by clicking on each hit number in the summary table (Figure 26).



Details

Title	Info
Genome:	Tomato (Solanum lycopersicum)
Reference:	Solanum lycopersicum (ITAG 2.40)
Version:	iTAG 2.30 Annotation (The official annotation of genes for tomato provided by International Tomato Annotation Group (ITAG) on the genome SL2.40)
Database:	SGN: Solgenomics database
Type:	Gene Annotation (Gene Annotation)
Species:	Solanum lycopersicum (SOLLC)
Species TaxonID:	4081

ID	Name	Target	Reference Sequnce	Start	End	Strand	Function			GB	Info	Sequence	Expression
gene:Solyc02g072000.2	Solyc02g072000.2		SL2.40ch02	35903150	35904957	+	Heat stress transcription factor A3 (AHRD V1 *-*- D1M7W9_SOLLC); contains Interp...	Structure	Overlaps	GB	Info	Sequence	Expression
gene:Solyc02g072060.1	Solyc02g072060.1		SL2.40ch02	35920510	35921838	+	Heat stress transcription factor (AHRD V1 *-*- D4QAU8_CARPA); contains Interpro ...	Structure	Overlaps	GB	Info	Sequence	Expression
gene:Solyc02g078340.2	Solyc02g078340.2		SL2.40ch02	37643661	37646059	+	Heat stress transcription factor (AHRD V1 **** D4QAU8_CARPA); contains Interpro ...	Structure	Overlaps	GB	Info	Sequence	Expression
gene:Solyc02g079180.1	Solyc02g079180.1		SL2.40ch02	38366347	38367682	-	Heat shock transcription factor 1 (AHRD V1 *-*- Q008S1_MEDSA); contains Interpro...	Structure	Overlaps	GB	Info	Sequence	Expression
gene:Solyc02g090820.2	Solyc02g090820.2		SL2.40ch02	46880125	46883382	-	Heat stress transcription factor (AHRD V1 **** D4QAU8_CARPA); contains Interpro ...	Structure	Overlaps	GB	Info	Sequence	Expression

Figure 26: Result representation of the specific collection tracks (here iTAG 2.3 predicted genes) by clicking on the number of hits found from the summary table (figure 23)

As presented in Figure 26, a sample snapshot of hit's details for a specific query are presented. In the results section, the information regarding the transcript collection and its mapping reference, the genomic information for each hit (such as chromosome, start, end, strand and functional annotation) and several hyperlinks to provide further investigational options for that track are provided. ***Structure*** hyperlink provides the sub features details of each transcript (UTRs, exons, CDS, introns) (Figure 27).

Details

ID	Name	Feature Type	Target	Parent	Start	End	strand
gene:Solyc02g072000.2	Solyc02g072000.2	gene			35903150	35904957	+
mRNA:Solyc02g072000.2.1		mRNA		gene:Solyc02g072000.2	35903150	35904957	+
exon:Solyc02g072000.2.1.1		exon		mRNA:Solyc02g072000.2.1	35903150	35903672	+
five_prime_UTR:Solyc02g072000.2.1.0		five_prime_UTR		mRNA:Solyc02g072000.2.1	35903150	35903456	+
CDS:Solyc02g072000.2.1.1		CDS		mRNA:Solyc02g072000.2.1	35903457	35903672	+
intron:Solyc02g072000.2.1.1		intron		mRNA:Solyc02g072000.2.1	35903673	35903768	+
exon:Solyc02g072000.2.1.2		exon		mRNA:Solyc02g072000.2.1	35903769	35904957	+
CDS:Solyc02g072000.2.1.2		CDS		mRNA:Solyc02g072000.2.1	35903769	35904779	+
three_prime_UTR:Solyc02g072000.2.1.0		five_prime_UTR		mRNA:Solyc02g072000.2.1	35904780	35904957	+

Figure 27: Snapshot of the representation of track structure, genomic coordinates and feature's parent-ship for a specific track.

Overlaps hyperlink lists all the tracks having overlap with this transcript on the genomic loci with in all the available collections. The ***Sequence*** link provides the sequence information (sequence, GC-Content, etc.) of the selected track in details (Figure 28).

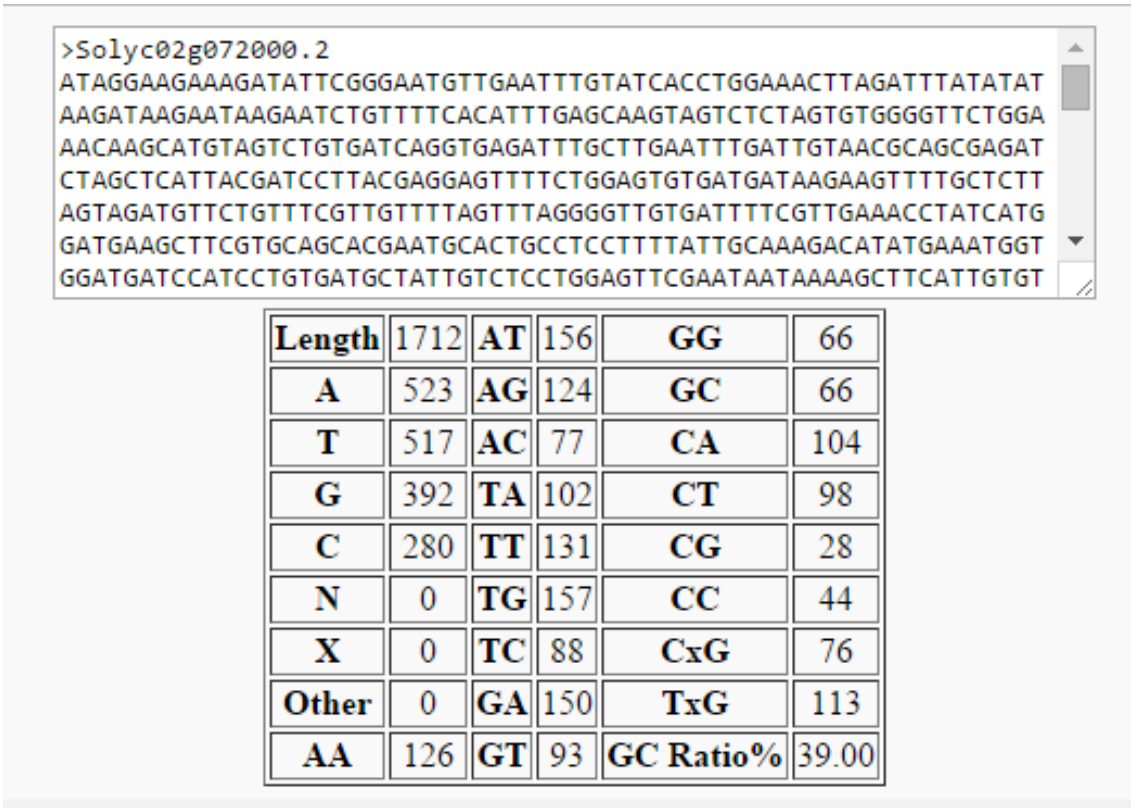


Figure 28: Snapshot of track sequence and GC-content information for a specific track.

GB hyperlink transfer the user to the genomic region of the selected track on the genome browser for the visualization purposes (Figure 29).

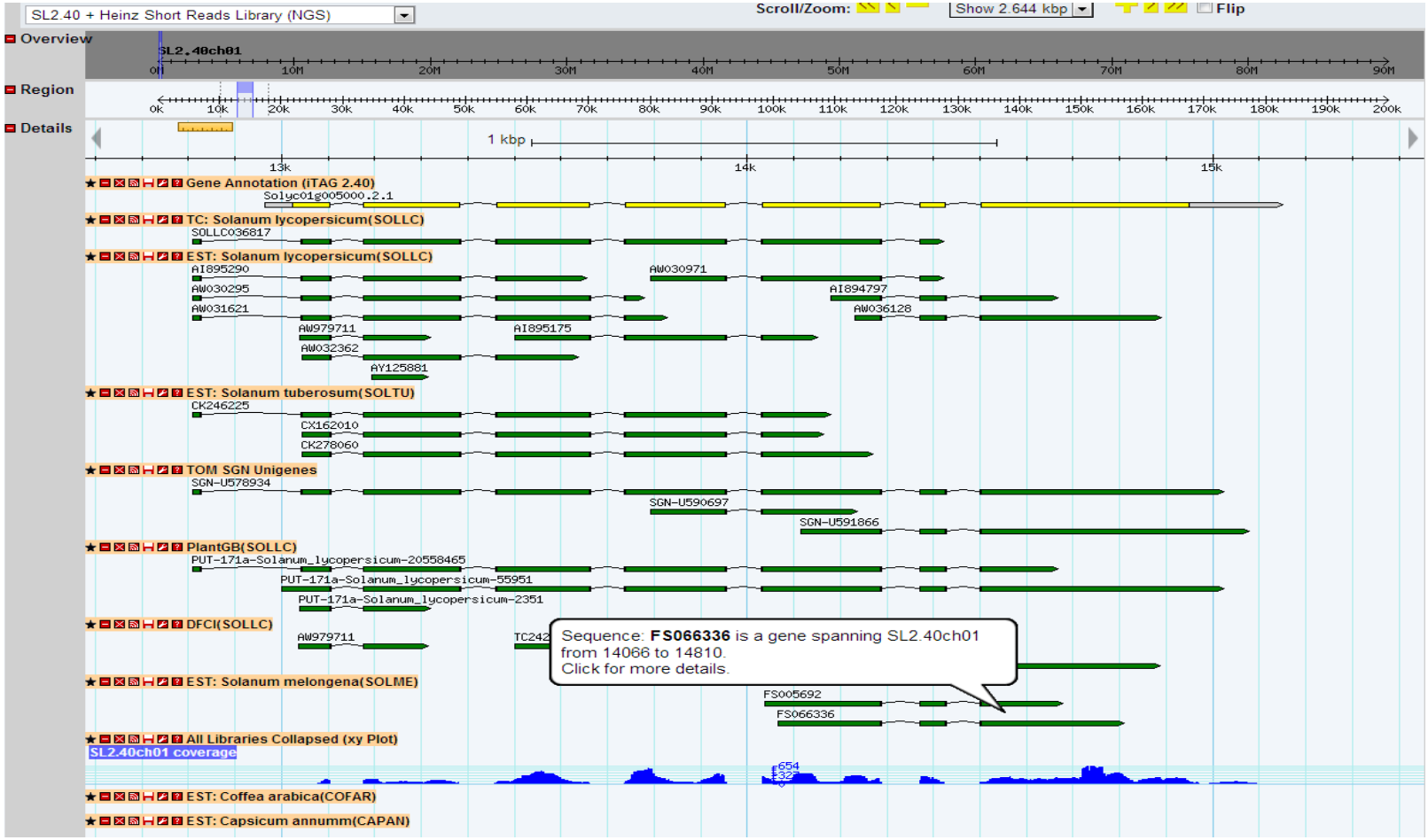


Figure 29: Snapshot of Gbrowse visualization of a specific genomic region with different track types (iTAG gene annotation, EST and TC tracks from tomato and potato, and RNAseq expression xyplot in the Heinz atlas collection) on that region

Info option provides the mapping information and the blast functional report regarding the selected track; and ***Expression*** hyperlink provides the RPKM expression of the selected track in all the available corresponding tissue/stages retrieved from the NexGenEx- database.
The genome platform presented, for the tomato at its current setting, provides flexible tools and facilities to further investigate the genomic space of the selected species.

3.4.2 NexGenEx-

We implemented *NexGenEx-* as a role based platform which enables the exploration of NGS based transcriptome collections. The platform was designed to provide enhanced tools for straightforward genome-wide gene expression analyses. A *Gbrowse* [141] database and associated interface are also embedded in the platform.

NexGenEx-Tom

NexGenEx-Tom is the dedicated partition for the organization of results from tomato NGS based transcriptomes. The platform was published in *BMC Plant Biology journal* on 2014 under the name of "*NexGenEx-Tom: a gene expression platform to investigate the functionalities of the tomato genome*", and is freely accessible via (http://cab.unina.it/nexgenex-tom). An instant implementation of the platform is also available in the *SPOT-ITN* Bioinformatics platform (http://cab.unina.it/spot-itn-bioinfo/expression/exp-search.aspx) to support the objectives of the project.

At the current release, the platform includes the processed gene expression datasets form one atlas and two other main publicly available *RNAseq* collections quantified and normalized with the main normalization approaches

(see Normalization). The platform is also enriched with collective and processed Gene Ontology datasets from the main resources with the possibility of cross link to the associated database for further investigations. Aside from the main motivation and objective of the platform to provide rapid and comprehensive investigational access to the processed NGS data collections with different quantification measures, tools for the clustering and correlation analyses, and the GO term to gene association and Go Enrichment assessment are also implemented in the web interface. Using a dedicated *Gbrowse* [141] interface, the platform has access to all the transcriptomics tracks available in the tomato genome platform mentioned before.

3.4.2.1.1 User Interface and Database Access

In Figure 30, we report the query page of the *NexGenEx*- platform. The figure shows the main sections users are provided with when consulting the platform content. In the *S. lycopersicum* cv. *Heinz* dedicated partition (which is accessible selecting ***genome*** "Tomato", ***reference*** "*S. lycopersicum* Version 2.40") the three available gene expression collections (*Heinz*, *Ailsa Craig*, S. *pimpinellifolium*) can be selected in the ***collection*** field. Crosslinks to reference raw collections and to the papers presenting them are also provided.

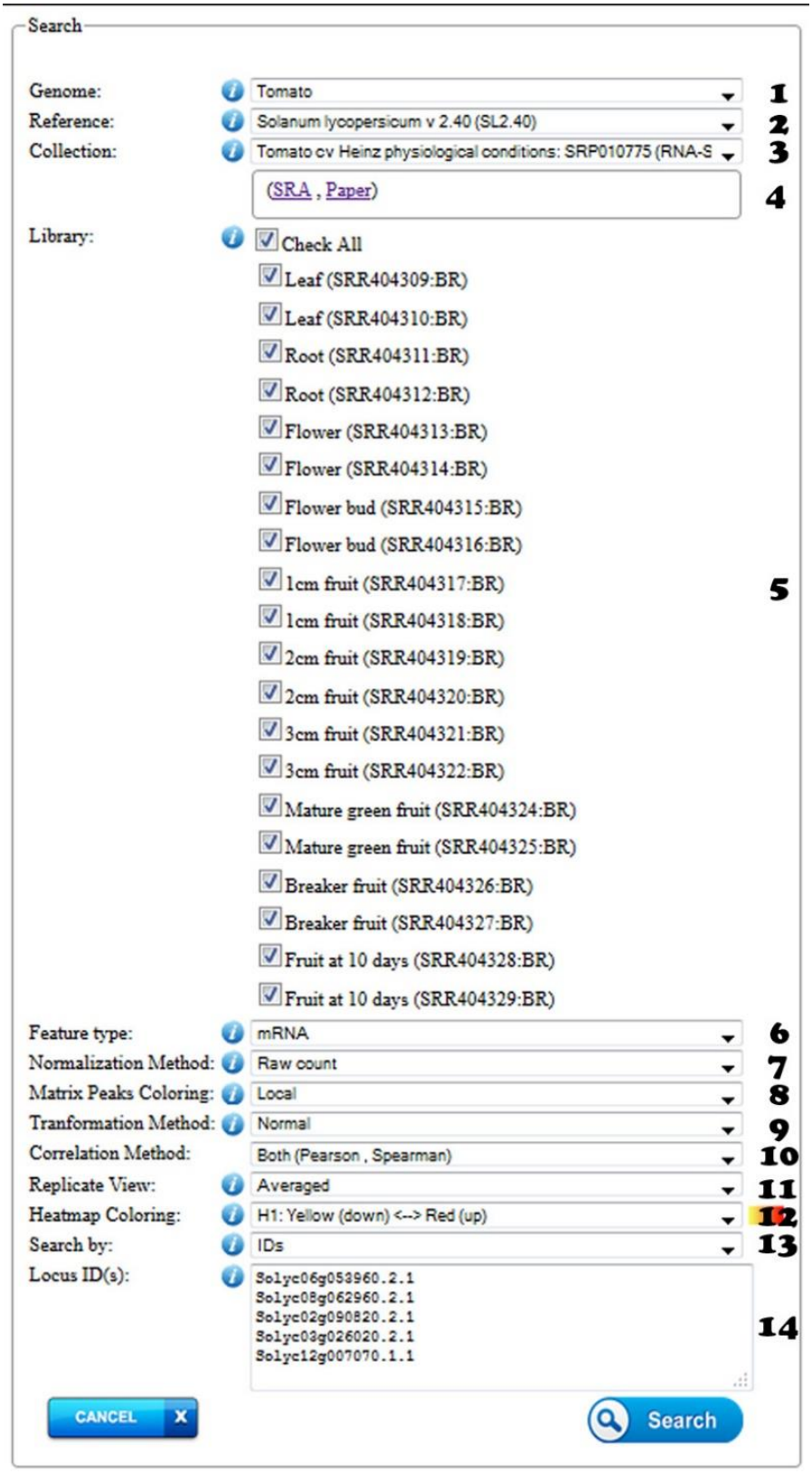


Figure 30: Main sections provided in the NexGenEx- platform query form. 1) Genome: the genome of interest, e.g. Tomato, 2) Reference: it indicates the reference genotype or cultivar

*sequences of interest, e.g. Solanum lycopersicum cv Heinz, version 2.40; 3) NGS collections available in the platform, e.g. the Heinz Illumina based RNA-seq collection in physiological condition; 4) link to the data source and paper for this collection; 5) available libraries (replicates/stages/tissues) included in the collection; 6) Feature type: represents the reference genome feature selected for read counting (e.g. mRNA, represents the exons in the locus); 7) Normalization method; 8) Matrix peak coloring, which defines the approach for color coding of the expression levels. This option assigns color frequencies to the cells of a heatmap view comparatively with the expression levels within the query result set (**local**) or within the whole selected libraries (**global**). 9) Transformation method: expression levels or their log2 transformed results can be accessed; 10) Correlation method: Pearson product-moment correlation coefficient or Spareman's correlation coefficient or Both; 11) Replicate view: defines the expression level by each libraries (**Separate**) or averaged between identical replicates (**Average**); 12) Heatmap coloring: different heatmap coloring combinations are provided for expression level visualization; 13) Search in: searchable fields can be one/multiple locus ids (IDs), or simple/multiple functional keywords with advanced selection options (Keyword), or genome regions (Region), and 14) the search area (Locus IDs/Functional keyword/Region): is the text area in which IDs or functional keywords may be listed, or a specific region of the genome may be specified. Accepted formats are described in the information pop-up from the website interface. "**Info"** buttons are available to support the users.*

The *NexGenEx-Tom* platform enables users to investigate expression of the reference tomato genes in different tissues and developmental stages from different collections in physiological conditions. Users can exploit the platform to investigate on a specific gene, or a set of genes. The query can be based on a list of Gene Identifiers (*IDs*) in the form of *Solyc* identifiers (e.g *Solyc01g00500*), or by indicating one or more functional keywords, or by specifying the boundaries of a chromosome region (indicating the specific directionality of transcription by selecting the *strand* option). Complex queries can be defined as indicated in the "***info***" links.

The web-based list of results is organized in an accordion view in which each result set can be investigated in its corresponding section/tab.

Here, an example query including a list of 27 heat shock factor genes in tomato and the corresponding result views are presented.

3.4.2.1.1.1 Annotation of the structure and functional annotation

By running the query in the system (Figure 30), the list of the resulting loci associated to the query, including their functional annotation and accessory information from the current gene annotation is reported (Figure 31).

Functional Annotation

Functional annotation and locus position

Functional Annotation

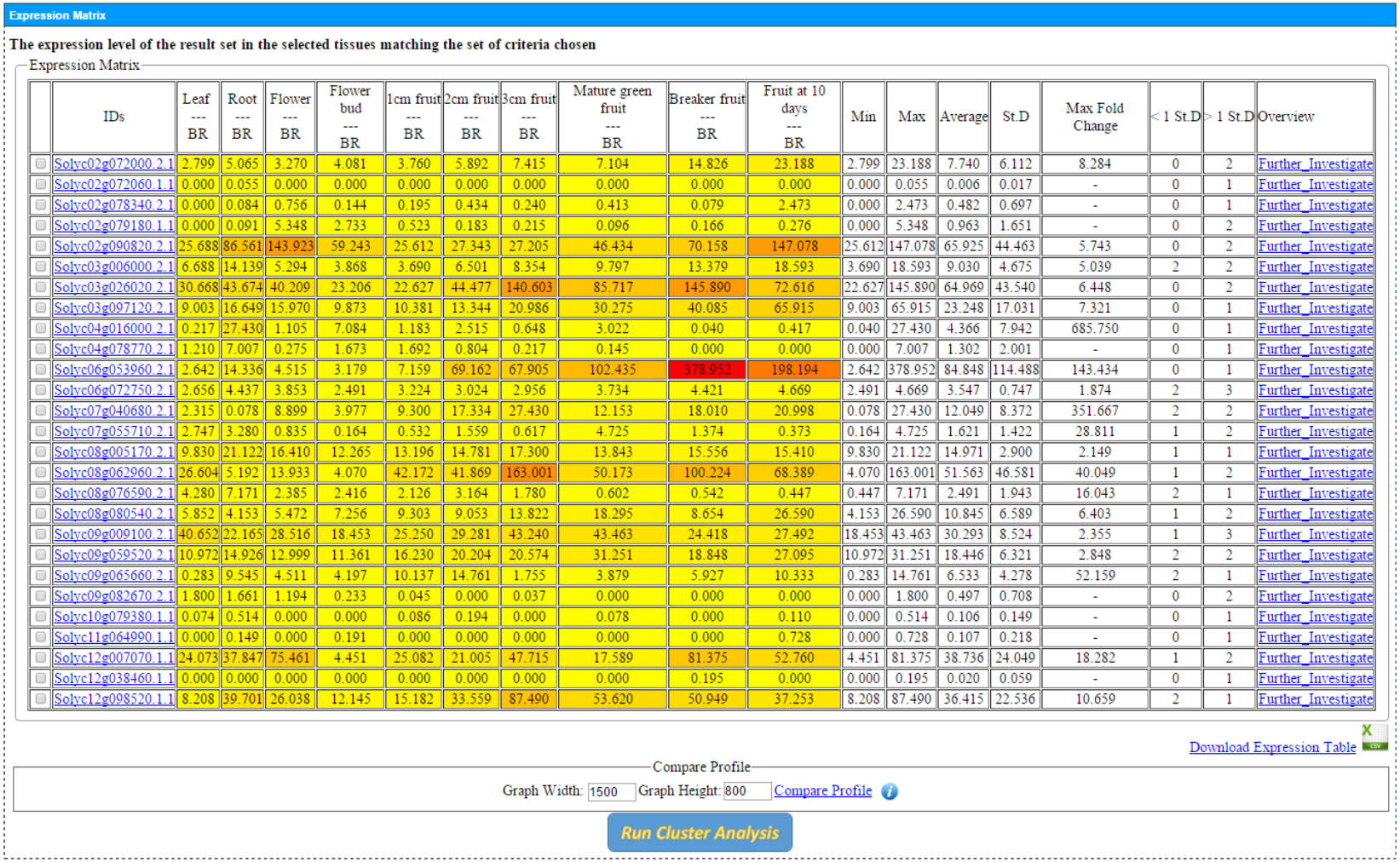
ID	ParrentID	Location	Start	End	Strand	Function
mRNA:Solyc02g072000.2.1	Solyc02g072000.2.1	SL2.30ch02	35903150	35904957	+	Heat stress transcription factor A3 (AHRD V1 *-*- D1M7W9_SOL...
mRNA:Solyc02g072060.1.1	Solyc02g072060.1.1	SL2.30ch02	35920510	35921838	+	Heat stress transcription factor (AHRD V1 *-*- D4QAU8_CARPA)...
mRNA:Solyc02g078340.2.1	Solyc02g078340.2.1	SL2.30ch02	37643661	37646059	+	Heat stress transcription factor (AHRD V1 **** D4QAU8_CARPA)...
mRNA:Solyc02g079180.1.1	Solyc02g079180.1.1	SL2.30ch02	38366347	38367682	-	Heat shock transcription factor 1 (AHRD V1 *-*- Q008S1_MEDSA...
mRNA:Solyc02g090820.2.1	Solyc02g090820.2.1	SL2.30ch02	46880125	46883382	-	Heat stress transcription factor (AHRD V1 **** D4QAU8_CARPA)...
mRNA:Solyc03g006000.2.1	Solyc03g006000.2.1	SL2.30ch03	678142	679948	+	Heat stress transcription factor A3 (AHRD V1 *-*- D1M7W9_SOL...
mRNA:Solyc03g026020.2.1	Solyc03g026020.2.1	SL2.30ch03	7810489	7812280	+	Heat stress transcription factor (AHRD V1 *-*- D4QAU8_CARPA)...
mRNA:Solyc03g097120.2.1	Solyc03g097120.2.1	SL2.30ch03	52901766	52904929	-	Heat stress transcription factor A3 (AHRD V1 *-** D1M7W9_SOL...
mRNA:Solyc04g016000.2.1	Solyc04g016000.2.1	SL2.30ch04	6594909	6598451	-	Heat stress transcription factor (AHRD V1 ***- D4QAU8_CARPA)...
mRNA:Solyc04g078770.2.1	Solyc04g078770.2.1	SL2.30ch04	61036586	61037903	+	Heat stress transcription factor (AHRD V1 *-*- D4QAU8_CARPA)...
mRNA:Solyc06g053960.2.1	Solyc06g053960.2.1	SL2.30ch06	33333411	33336335	-	Heat stress transcription factor A3 (AHRD V1 ***- D1M7W9_SOL...
mRNA:Solyc06g072750.2.1	Solyc06g072750.2.1	SL2.30ch06	41255352	41258348	+	Heat stress transcription factor A3 (AHRD V1 *-** D1M7W9_SOL...
mRNA:Solyc07g040680.2.1	Solyc07g040680.2.1	SL2.30ch07	46702761	46704429	+	Heat stress transcription factor A3 (AHRD V1 **** D1M7W9_SOL...
mRNA:Solyc07g055710.2.1	Solyc07g055710.2.1	SL2.30ch07	60972388	60973952	-	Heat stress transcription factor A3 (AHRD V1 *-*- D1M7W9_SOL...
mRNA:Solyc08g005170.2.1	Solyc08g005170.2.1	SL2.30ch08	111412	116839	-	Heat stress transcription factor A3 (AHRD V1 *-*- D1M7W9_SOL...
mRNA:Solyc08g062960.2.1	Solyc08g062960.2.1	SL2.30ch08	49589145	49591151	-	Heat stress transcription factor A3 (AHRD V1 *-** D1M7W9_SOL...
mRNA:Solyc08g076590.2.1	Solyc08g076590.2.1	SL2.30ch08	57710679	57714096	-	Heat stress transcription factor A3 (AHRD V1 *-** D1M7W9_SOL...
mRNA:Solyc08g080540.2.1	Solyc08g080540.2.1	SL2.30ch08	60985869	60987278	-	Heat stress transcription factor (AHRD V1 *-** D4QAU8_CARPA)...
mRNA:Solyc09g009100.2.1	Solyc09g009100.2.1	SL2.30ch09	2445341	2448016	-	Heat stress transcription factor A3 (AHRD V1 ***- D1M7W9_SOL...
mRNA:Solyc09g059520.2.1	Solyc09g059520.2.1	SL2.30ch09	50372011	50379351	-	Heat stress transcription factor A3 (AHRD V1 *-** D1M7W9_SOL...
mRNA:Solyc09g065660.2.1	Solyc09g065660.2.1	SL2.30ch09	59473864	59475995	+	Heat stress transcription factor A3 (AHRD V1 ***- D1M7W9_SOL...
mRNA:Solyc09g082670.2.1	Solyc09g082670.2.1	SL2.30ch09	63781968	63784228	+	Heat stress transcription factor A3 (AHRD V1 *-** D1M7W9_SOL...
mRNA:Solyc10g079380.1.1	Solyc10g079380.1.1	SL2.30ch10	60254409	60255720	+	Heat stress transcription factor (AHRD V1 ***- D4QAU8_CARPA)...
mRNA:Solyc11g064990.1.1	Solyc11g064990.1.1	SL2.30ch11	47389718	47391840	-	Heat stress transcription factor (AHRD V1 **** D4QAU8_CARPA)...
mRNA:Solyc12g007070.1.1	Solyc12g007070.1.1	SL2.30ch12	1512824	1514090	+	Heat stress transcription factor A3 (AHRD V1 ***- D1M7W9_SOL...
mRNA:Solyc12g038460.1.1	Solyc12g038460.1.1	SL2.30ch12	35761710	35761970	+	Pre-rRNA-processing protein PNO1 (AHRD V1 ***- C6HSF7_AJECH)...
mRNA:Solyc12g098520.1.1	Solyc12g098520.1.1	SL2.30ch12	64340677	64344407	-	Heat stress transcription factor A3 (AHRD V1 *-*- D1M7W9_SOL...

Figure 31: Snapshot of the annotation structure and functional annotation

The result set includes the Gene ID, the genomic location the locus on the genome, the strand of the transcript and the functional annotation associated to the resulted gene. A hyperlink from the gene to its genomic locus, visualized by the *GBrowse*, is available for each gene.

3.4.2.1.1.2 Expression matrix and profiling

As presented in Figure 32, the expression levels of each queried gene, based on the pre-settings of the query option provided by the user, can be investigated by the selected libraries, in the form of read counts per each locus, median normalized counts or RPKM. As an optional parameter, the average expression level of the replicates from each library can be investigated (***Replicate view*** set to ***Average***, in the query options).



Expression Matrix

The expression level of the result set in the selected tissues matching the set of criteria chosen

Expression Matrix

	IDs	Leaf --- BR	Root --- BR	Flower --- BR	Flower bud --- BR	1cm fruit --- BR	2cm fruit --- BR	3cm fruit --- BR	Mature green fruit --- BR	Breaker fruit --- BR	Fruit at 10 days --- BR	Min	Max	Average	St.D	Max Fold Change	< 1 St.D	> 1 St.D	Overview
	Solyc02g072000.2.1	2.799	5.065	3.270	4.081	3.760	5.892	7.415	7.104	14.826	23.188	2.799	23.188	7.740	6.112	8.284	0	2	Further_Investigate
	Solyc02g072060.1.1	0.000	0.055	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.055	0.006	0.017	-	0	1	Further_Investigate
	Solyc02g078340.2.1	0.000	0.084	0.756	0.144	0.195	0.434	0.240	0.413	0.079	2.473	0.000	2.473	0.482	0.697	-	0	1	Further_Investigate
	Solyc02g079180.1.1	0.000	0.091	5.348	2.733	0.523	0.183	0.215	0.096	0.166	0.276	0.000	5.348	0.963	1.651	-	0	2	Further_Investigate
	Solyc02g090820.2.1	25.688	86.561	143.923	59.243	25.612	27.343	27.205	46.434	70.158	147.078	25.612	147.078	65.925	44.463	5.743	0	2	Further_Investigate
	Solyc03g006000.2.1	6.688	14.139	5.294	3.868	3.690	6.501	8.354	9.797	13.379	18.593	3.690	18.593	9.030	4.675	5.039	2	2	Further_Investigate
	Solyc03g026020.2.1	30.668	43.674	40.209	23.206	22.627	44.477	140.603	85.717	145.890	72.616	22.627	145.890	64.969	43.540	6.448	0	2	Further_Investigate
	Solyc03g097120.2.1	9.003	16.649	15.970	9.873	10.381	13.344	20.986	30.275	40.085	65.915	9.003	65.915	23.248	17.031	7.321	0	1	Further_Investigate
	Solyc04g016000.2.1	0.217	27.430	1.105	7.084	1.183	2.515	0.648	3.022	0.040	0.417	0.040	27.430	4.366	7.942	685.750	0	1	Further_Investigate
	Solyc04g078770.2.1	1.210	7.007	0.275	1.673	1.692	0.804	0.217	0.145	0.000	0.000	0.000	7.007	1.302	2.001	-	0	1	Further_Investigate
	Solyc06g053960.2.1	2.642	14.336	4.515	3.179	7.159	69.162	67.905	102.435	378.952	198.194	2.642	378.952	84.848	114.488	143.434	0	1	Further_Investigate
	Solyc06g072750.2.1	2.656	4.437	3.853	2.491	3.224	3.024	2.956	3.734	4.421	4.669	2.491	4.669	3.547	0.747	1.874	2	3	Further_Investigate
	Solyc07g040680.2.1	2.315	0.078	8.899	3.977	9.300	17.334	27.430	12.153	18.010	20.998	0.078	27.430	12.049	8.372	351.667	2	2	Further_Investigate
	Solyc07g055710.2.1	2.747	3.280	0.835	0.164	0.532	1.559	0.617	4.725	1.374	0.373	0.164	4.725	1.621	1.422	28.811	1	2	Further_Investigate
	Solyc08g005170.2.1	9.830	21.122	16.410	12.265	13.196	14.781	17.300	13.843	15.556	15.410	9.830	21.122	14.971	2.900	2.149	1	1	Further_Investigate
	Solyc08g062960.2.1	26.604	5.192	13.933	4.070	42.172	41.869	163.001	50.173	100.224	68.389	4.070	163.001	51.563	46.581	40.049	1	2	Further_Investigate
	Solyc08g076590.2.1	4.280	7.171	2.385	2.416	2.126	3.164	1.780	0.602	0.542	0.447	0.447	7.171	2.491	1.943	16.043	2	1	Further_Investigate
	Solyc08g080540.2.1	5.852	4.153	5.472	7.256	9.303	9.053	13.822	18.295	8.654	26.590	4.153	26.590	10.845	6.589	6.403	1	2	Further_Investigate
	Solyc09g009100.2.1	40.652	22.165	28.516	18.453	25.250	29.281	43.240	43.463	24.418	27.492	18.453	43.463	30.293	8.524	2.355	1	3	Further_Investigate
	Solyc09g059520.2.1	10.972	14.926	12.999	11.361	16.230	20.204	20.574	31.251	18.848	27.095	10.972	31.251	18.446	6.321	2.848	2	2	Further_Investigate
	Solyc09g065660.2.1	0.283	9.545	4.511	4.197	10.137	14.761	1.755	3.879	5.927	10.333	0.283	14.761	6.533	4.278	52.159	2	1	Further_Investigate
	Solyc09g082670.2.1	1.800	1.661	1.194	0.233	0.045	0.000	0.037	0.000	0.000	0.000	0.000	1.800	0.497	0.708	-	0	2	Further_Investigate
	Solyc10g079380.1.1	0.074	0.514	0.000	0.000	0.086	0.194	0.000	0.078	0.000	0.110	0.000	0.514	0.106	0.149	-	0	1	Further_Investigate
	Solyc11g064990.1.1	0.000	0.149	0.000	0.191	0.000	0.000	0.000	0.000	0.000	0.728	0.000	0.728	0.107	0.218	-	0	1	Further_Investigate
	Solyc12g007070.1.1	24.073	37.847	75.461	4.451	25.082	21.005	47.715	17.589	81.375	52.760	4.451	81.375	38.736	24.049	18.282	1	2	Further_Investigate
	Solyc12g038460.1.1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.195	0.000	0.000	0.195	0.020	0.059	-	0	1	Further_Investigate
	Solyc12g098520.1.1	8.208	39.701	26.038	12.145	15.182	33.559	87.490	53.620	50.949	37.253	8.208	87.490	36.415	22.536	10.659	2	1	Further_Investigate

Download Expression Table

Compare Profile

Graph Width: 1500 Graph Height: 800 Compare Profile

Run Cluster Analysis

Figure 32: Snapshot of expression matrix in NexGenEx- for a resulted gene set

In addition, a complementary set of statistics, such as the minimum, maximum, average, standard deviation and maximum fold change of the expression levels of each locus in the selected libraries are provided. Moreover, the number of times each of the expression values exceeds the boundaries of one standard deviation from the average is also shown. This value permits to efficiently investigate locus specificities [159]. This section provides a general abstract of the loci expression level behavior in the selected libraries. Moreover, by selecting only 2 different tissues for specific gene sets, bi-comparison of the gene expression fold change are delivered permitting to identify the differential expression levels.

The expression matrix can also be downloaded in csv format.

Overview

Results Overview

Expression everywhere for : ***Solyc06g053960.2.1***

Collection: Tomato cv Heinz physiological conditions

Accession: SRP010775
Technique: RNA-Seq

Normalization\Sample	Leaf	Leaf	Root	Root	Flower	Flower	Flower bud	Flower bud	1cm fruit	1cm fruit	2cm fruit	2cm fruit	3cm fruit	3cm fruit	Mature green fruit	Mature green fruit	Breaker fruit	Breaker fruit	Fruit at 10 days	Fruit at 10 days
Raw count	12	29	106	121	40	32	25	34	43	55	171	687	591	540	637	765	3424	3110	2142	2044
DESeq Method	15.2685	37.5441	90.4227	105.389	41.3659	34.1805	21.7802	30.4491	45.0244	58.663	204.812	759.898	492.381	454.199	721.779	718.657	3424.04	3227.65	1698	1687.43
RPKM	1.5202	3.76305	13.3493	15.3224	4.93959	4.09078	2.65152	3.70704	6.2056	8.11248	30.1405	108.183	70.5018	65.308	102.652	102.218	390.53	367.374	198.483	197.905
Raw Multiple	10	29	111	123	42	41	30	38	34	60	178	707	600	562	661	816	3261	2984	2073	1982

Collection: GenXpro

Accession: MACE
Technique: MACE

Normalization\Sample	Control Mature	Control Mature	Control Mature	Control Post-meiotic	Control Post-meiotic	Control Post-meiotic	Control Tetrad	Control Tetrad	Control Tetrad	Heat Stress Mature	Heat Stress Mature	Heat Stress Mature	Heat Stress Tetrad	Heat Stress Post-meiotic	Heat Stress Post-meiotic	Heat Stress Post-meiotic	Heat Stress Tetrad	Heat Stress Tetrad
Raw count	3	5	8	4	2	1	1	1	0	11	9	3	1	2	3	5	0	2
DESeq Method	1.61841	3.06334	4.70987	2.80288	2.87773	1.09734	1.34512	2.99683	0	3.68722	3.70783	1.40202	2.27293	1.57301	2.1795	2.88019	0	2.35202
TPM	7.92111	14.0104	23.6985	13.2179	14.6245	6.21276	3.16662	5.40167	0	19.3384	18.2192	7.01202	4.95555	7.26293	11.4361	15.8871	0	6.55966

Collection: Tomato cv Ailsa Craig physiological conditions

Accession: SRP008367
Technique: RNA-Seq

Normalization\Sample	3' end sequencing of tomato transcripts	Immature green fruit	Immature green fruit	Immature green fruit	Immature green fruit	Mature green fruit	Mature green fruit	Mature green fruit	Mature green fruit	Breaker fruit	Breaker fruit	Breaker fruit	Breaker fruit	Fully ripe fruit	Fully ripe fruit	Fully ripe fruit	Fully ripe fruit	5' end sequencing of tomato transcripts	5' end sequencing of tomato transcripts	5' end sequencing of tomato transcripts
Raw count	25	178	107	72	80	884	1411	289	768	831	1711	648	1301	312	1415	509	528	3461	595	233
DESeq Method	114.452	132.509	101.128	118.953	100.5	593.534	1217.83	273.265	803.013	1102.35	2071.82	842.462	1173.76	473.408	689.269	432.642	272.078	998.516	819.807	81.5698
RPKM	10.559	16.5015	12.8755	14.7941	12.2395	71.0599	165.903	31.254	110.833	112.432	262.192	83.8481	138.778	53.0556	76.5934	44.2194	28.522	74.1248	44.6482	6.48991

Collection: Solanum pimpinellifolium physiological conditions

Accession: SRP010775
Technique: RNA-Seq

Normalization\Sample	Leaf	Leaf	fruit at 5 days after the breaker stage	fruit at 5 days after the breaker stage	Breaker fruit	Breaker fruit	Immature green fruit	Immature green fruit
Raw count	10	8	2101	2431	2360	1605	1209	1563
DESeq Method	10.841	8.63324	2262.59	2192.84	1765.15	1866.89	1155.71	1465.42
RPKM	0.373277	0.286562	84.0408	85.9477	81.2595	86.1762	57.287	75.9591

Figure 33: Snapshot of the result page for a gene's expression level in all the available collections.

For each resulting gene, a button (***further investigate***) has been implemented which enables users to further investigate its gene expression in different conditions (Figure 33). Indeed, by clicking on this link, the expression level of the corresponding gene will be reported for all the available NGS collections in the platform, in all the available normalized forms, as calculated for each collection associated to the genome reference. This enables users to focus on the locus of interest with a complete overview of its behavior in any possible and available library per collection.

3.4.2.1.1.3 Heatmap visualization

Heatmaps provide a suitable view on gene expression levels. Customizable heatmaps are offered in the platform to highlight high- and low- expressed genes. This graphical approach is exploited in the platform to show the expression levels of one or multiple genes in different conditions. The data can be reported in the form of a matrix, where the level of expression can be marked by a specific color scaling, which may help to highlight high, medium and low levels of expression. The heatmap provided in the *NexGenEx*- platform (Figure 32 and Figure 35) can be defined by a ***local*** or a ***global*** scaling, according to the preferred selection in the Matrix Peak Coloring option (Figure 30). The "***local***" heatmap option provides the expression level coloring ranging from the lowest to the highest expression levels resulting from the query. This facilitates the comparison of the specific gene expression levels in the selected set. The "***global***" coloring option defines the coloring range on the basis of the lowest and highest expression levels in the whole libraries selected during the query. This enables users to identify the gene expression level when compared with the whole expression levels from all the genes in the selected library/ies.

3.4.2.1.1.4 Expression profiling plot

The Expression Profiling plot (Figure 34) shows gene expression variability in different samples from the collection under investigation.

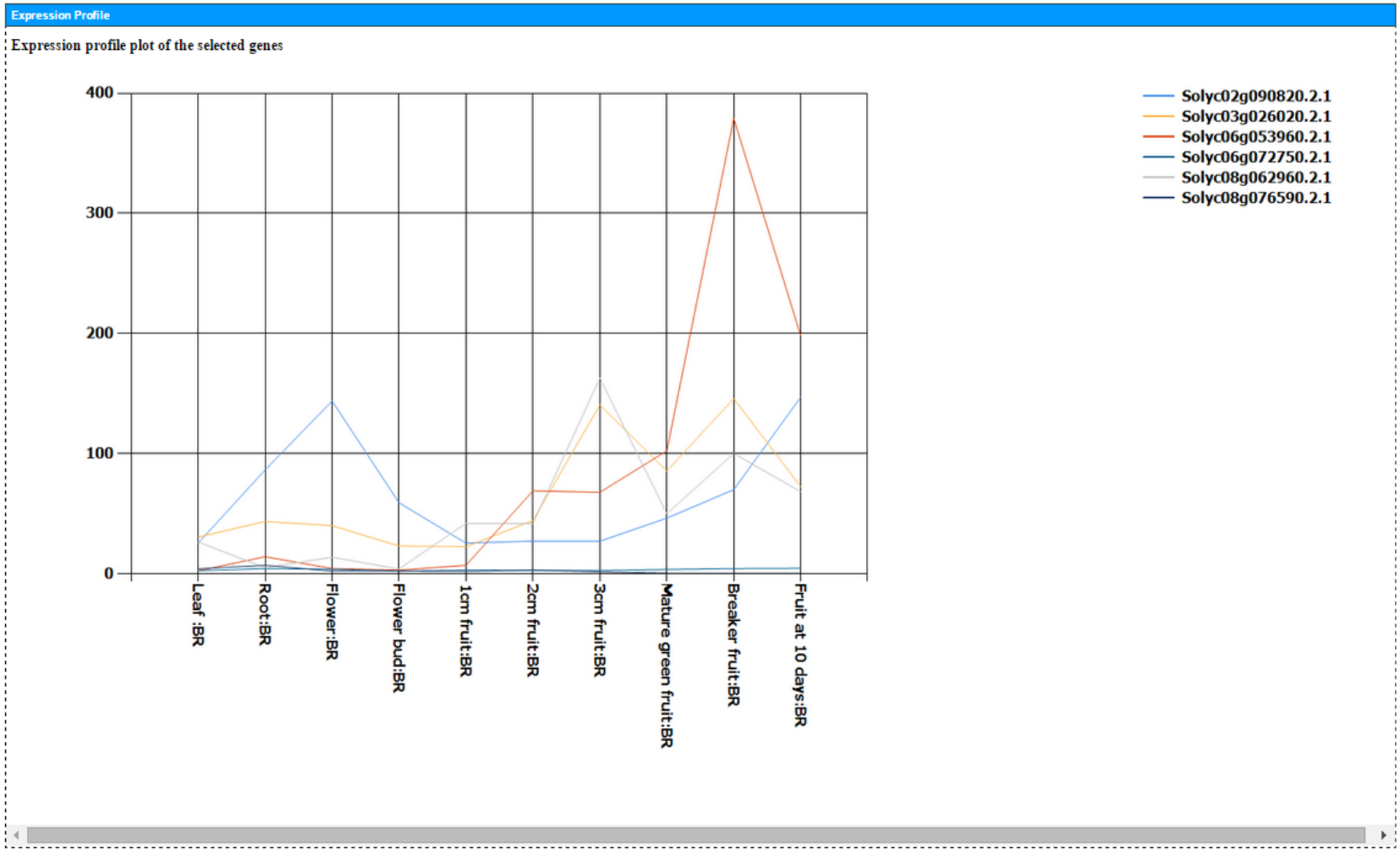


Figure 34: Snapshot of the expression profiling plot for a selected gene set across the selected tissues/stages/conditions

This view depends on the number of libraries selected. The possibility to perform the analysis on specific collections of genes, selected by keyword or ID or by a genome region, allows the comparison of the expression profiles of several genes in a straightforward way.

3.4.2.1.1.5 Correlation Matrix

The Correlation matrix analysis illustrates the correlation between genes on the basis of the selected libraries (Figure 35).

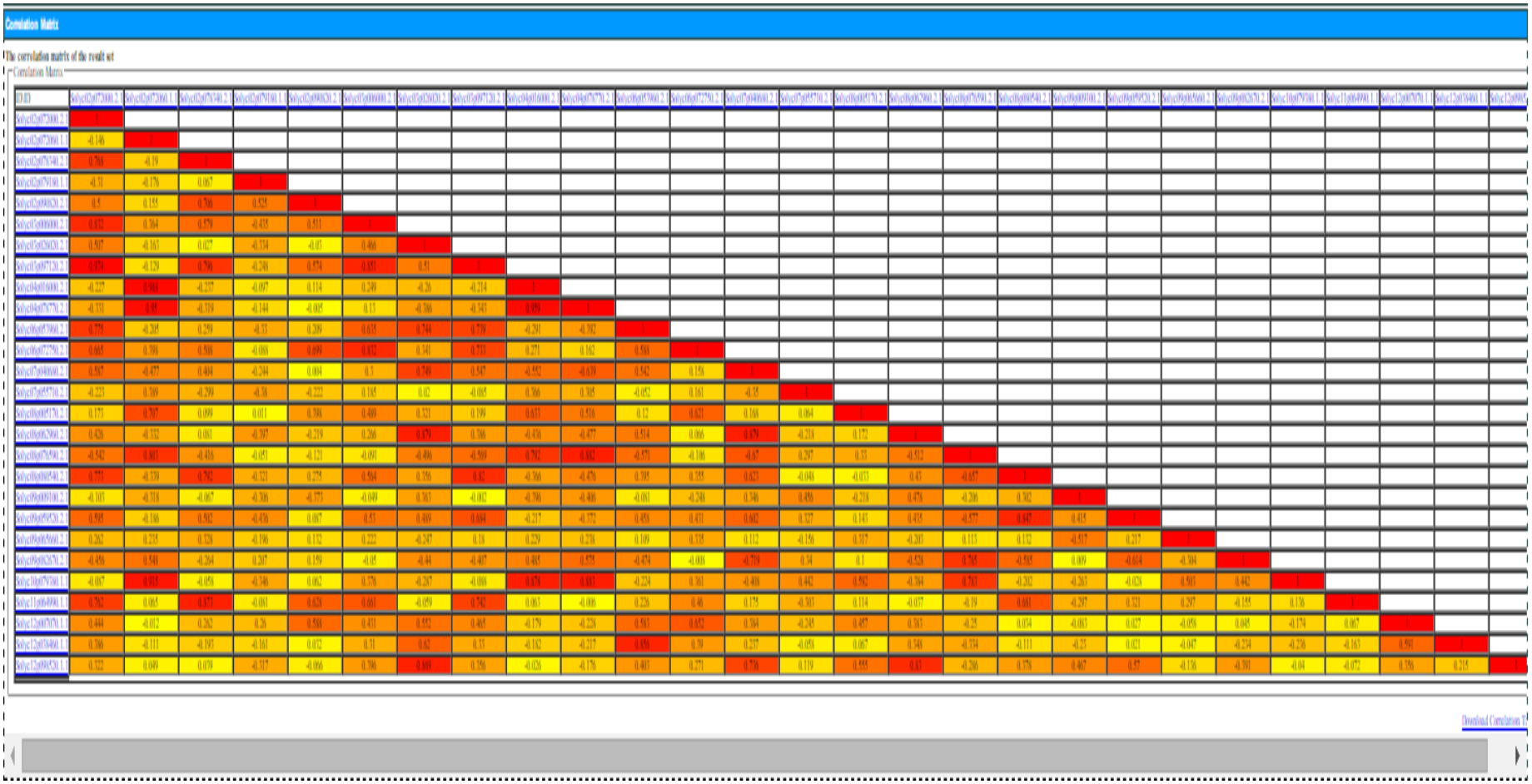


Figure 35: Snapshot of the correlation matrix between the 27 HSF genes in tomato versus each other based on the selected tissues/stages/condition (a matrix of 27 ×27).

The analysis can be based on the *Pearson product-moment correlation coefficient*, or on the *Spearman's rank correlation coefficient*, or both at the same time; and the resulting values fluctuates between -1 to 1, providing the negatively or positively correlated genes (see 1.2.5.1.5).
The correlation matrix can be also downloaded in csv format.

3.4.2.1.1.6 Cluster Analyses

NexGenEx-Tom provides a *k-means* clustering tool in which the clustering of the genes in the result set are easily possible on the bases of their expression profiling across the selected tissue/stages. The *k-means* clustering tool offered in *NexGenEx-Tom* is the online package of *k-means* Cluster Analyzer presented in section 1.2.5.1.6.

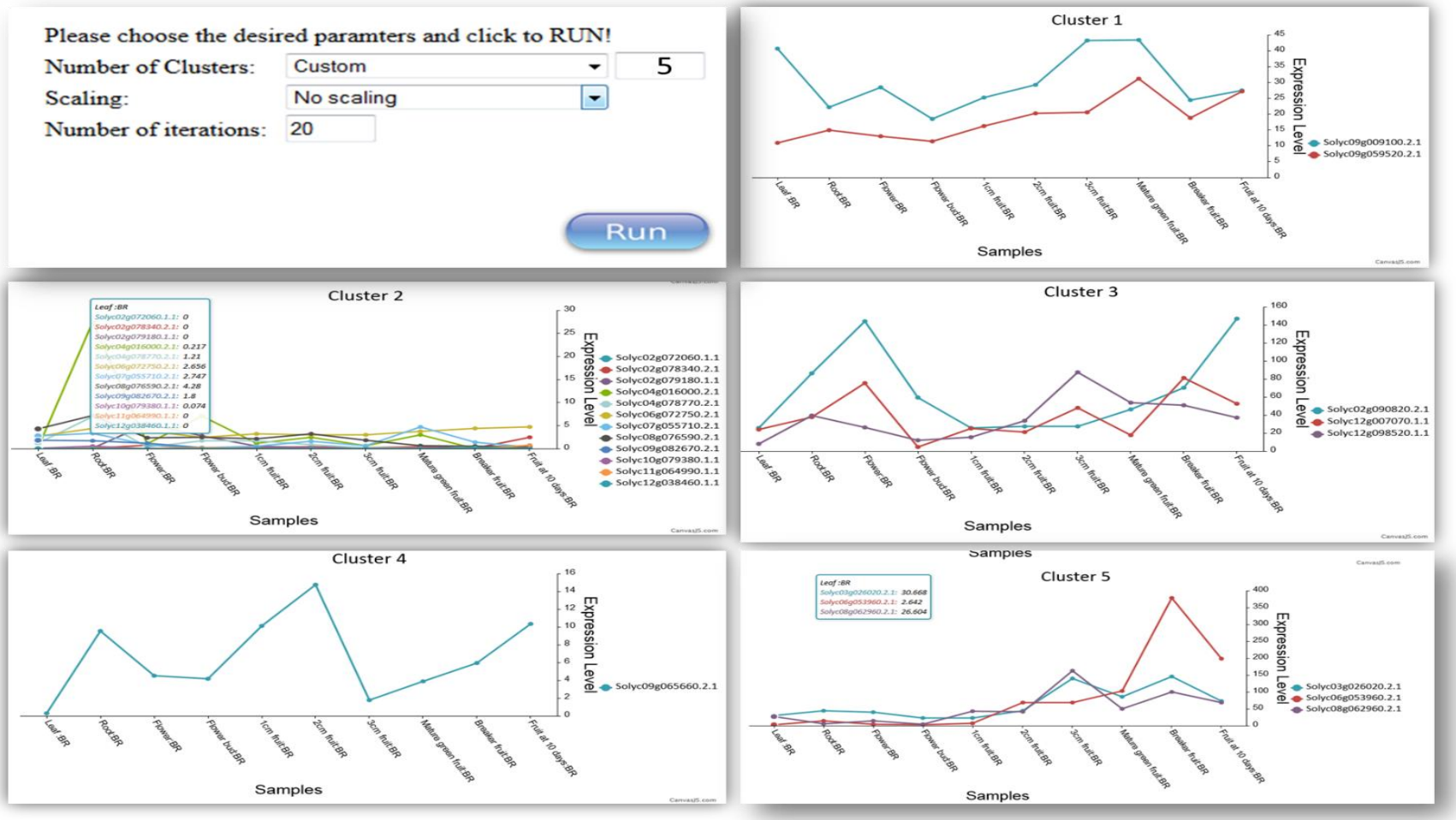


Figure 36: The k-means clustering (k=5) with 20 iterations and no rescaling on the 27 Heat Shock Factor genes in tomato.

As presented in Figure 36: The k-means clustering (k=5) with 20 iterations and no rescaling on the 27 Heat Shock Factor genes in tomato., the cluster analyses on the 27 heat shock factor genes in *iTAG 2.3* annotation, across all the tissues/stages of the *Heinz RNAseq* (see **Error! Reference source not found.**) collection, was carried out organizing the genes with similar expression profile into five (5) distinct clusters. As it can be observed, the genes with similar expression trend are grouped together. The clustering can be performed with different number of clusters on the normal or rescaled gene expression values.

3.4.2.1.1.7 GO Terms Summary Table and their association

As it is shown in Figure 37, a GO Term summary table and the gene to GO Term association to the queried genes is provided to the end users. Figure 37.A shows an example of a resulting GO Term summary table of the list of occurring GO Terms, type of GO (in terms of CC: Cellular component, MF: Molecular Function and BP: Biological Process), GO location and the specific GO descriptions. In addition, the enrichment of the GO in the resulting gene set is sorted by p-value for further investigation purposes.

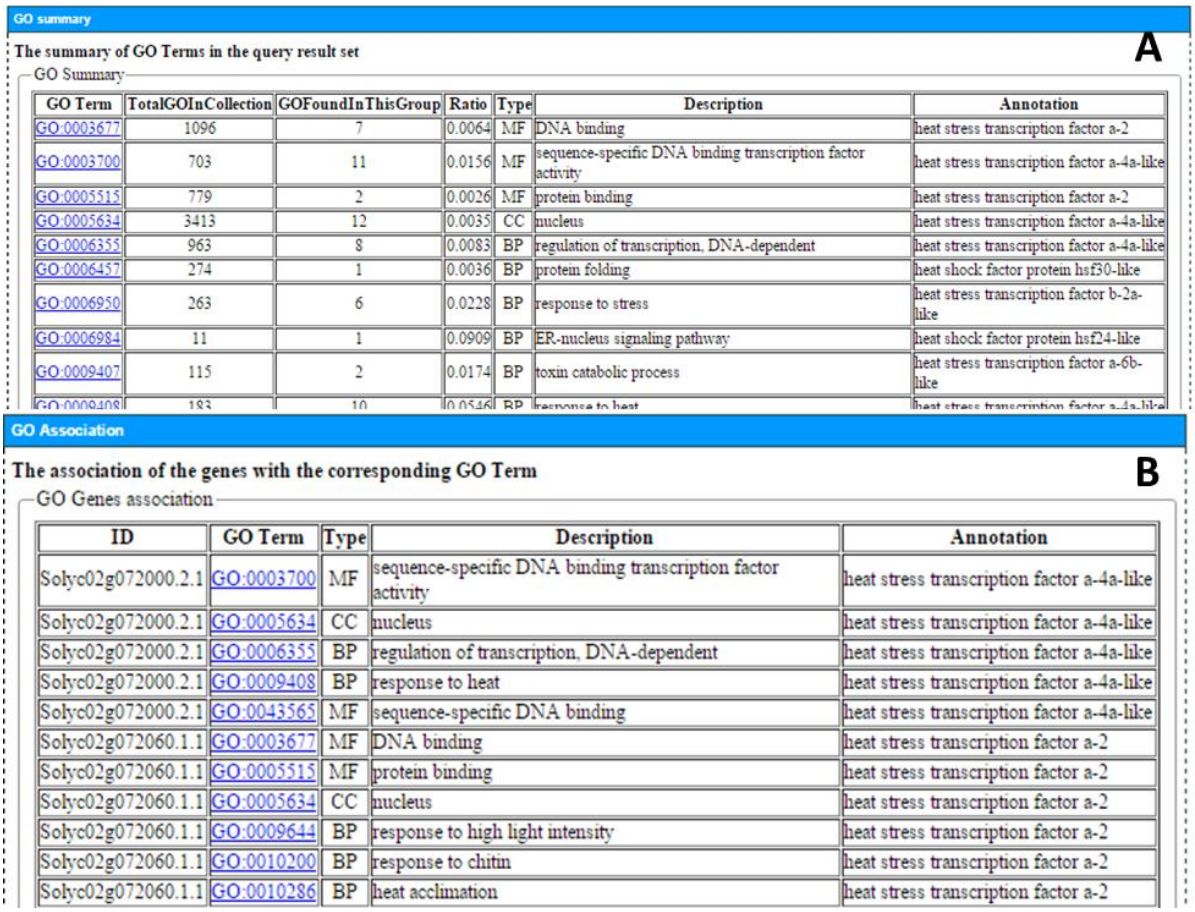


Figure 37: Snapshot of A) the GO enrichment results and B) the GO to gene annotation for a selected gene set.

Moreover, the genes to GO Terms association table (Figure 37.B) also provides the association list of the resulting genes with their GO Terms and their complete description. To further investigate the GO Terms, each GO is also linked to the *AmiGO* ontology and annotation database [160].

3.4.2.1.1.8 Genome Browser Crosslink

NexGenEx-Tom is enriched with an embedded, customized and updated genome browser interface [141]. The genome browser used, *Gbrowse*, permits a genome based investigation of the structure of the gene loci included in the database and can be accessed by the selection of each locus from the query set (Figure 38).

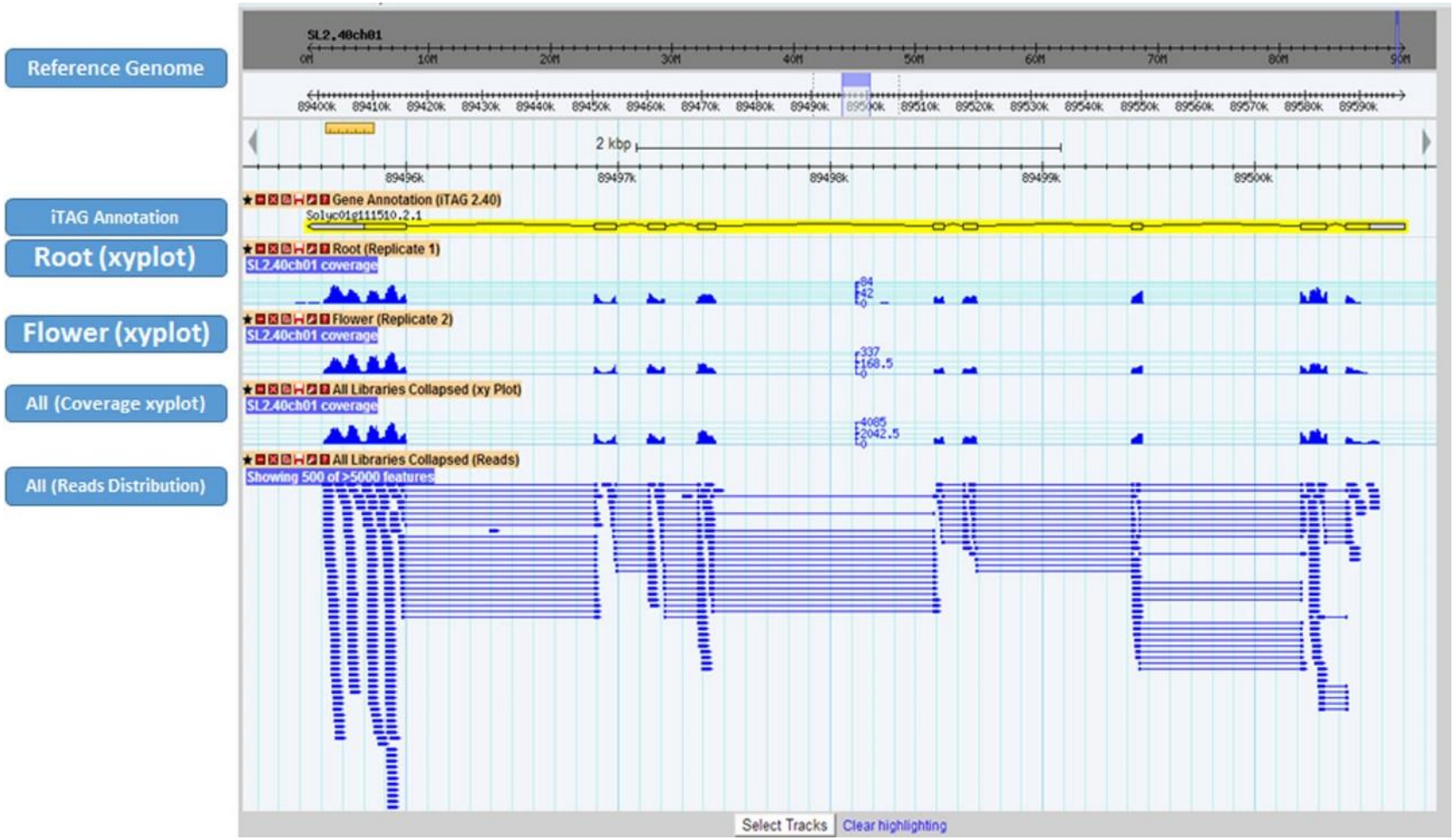


Figure 38: Example of a Gbrowse based view. The gene locus, the xyplot coverage of the NGS reads and their mapping along the selected locus are shown.

Figure 38 shows an example of a *Gbrowse* based view offered by the platform. This view enables users with in-depth investigation of the selected loci and their associated pattern of expression in the form of reads distribution along the genome sequence. Expression profile of the reads mapped on the genome for each tissue/stages is provided in the form of read tracks and coverage plot (*xyplot*). Another track defined on the basis of all the reads from the available libraries (combined) is also added to provide a general overview of the locus expression for each collection. The *iTAG* gene annotation is also accessible through the *Gbrowse* partition. Specifically, the *NexGenEx-Tom Gbrowse* partition is also enriched with all the annotation tracks included in the *Tomato Genome Platform* presented in the previous section.

3.4.3 Orthologs Platform

To allow the comparative analyses between different species, an orthologs platform was designed and implemented. The platform provides different query pages to investigate the ortholog groups and their associated functional and genomic information in details. An instant application of the platform and its interfaces are currently implemented in the *SPOT-ITN* Bioinformatics platform (http://cab.unina.it/*SPOT-ITN*-bioinfo/orthpar/orthpar-search.aspx).

OrthPar-Tom

The *OrthPar-Tom platform*, focused on the tomato orthology with different species. At the current setting, the platform represents the orthology across two collections (see Orthologs Platform), all including the tomato. The platform is also enriched with the protein sequences and their domain information. For tomato genes, the platform is also cross linked to the NexGenEx-Tom for gene expression visualization and investigation.

Here we present its query page, some of its features, and the results representation of this platform for the current collections.

3.4.3.1.1 User Interface and Database Access

The platform provides a simple user interface to conduct the queries.

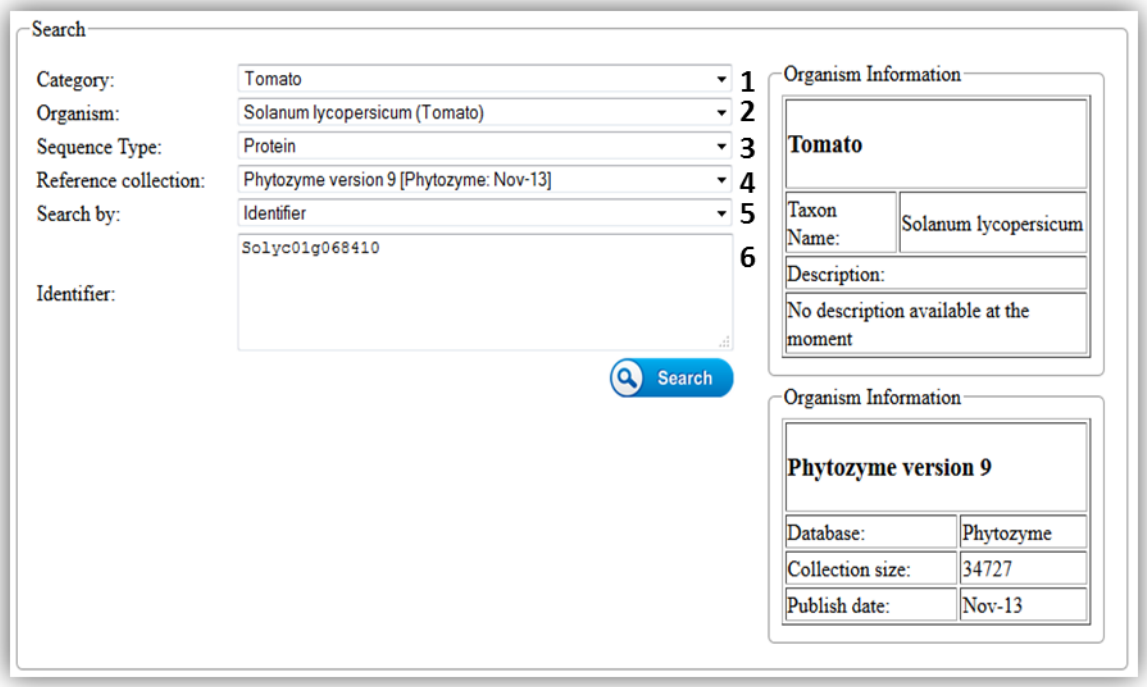


Figure 39: Snapshot of the orthologs platform query page

As presented in Figure 39: Snapshot of the orthologs platform query page, the orthologs platform query page is presented. By choosing the genome of interest to investigate (29), the genome reference version (2), the type of sequence was used for the orthology investigation (protein, transcript or gene) (3), the ortholog collection available in the platform yhe user want to investigate (4), the type of query (ID, functional keyword, or domain keyword) (5) and the keyword text (6); the information regarding the specified ortholog collection will presented

next to the query fields. By running the query in the system, the results matching the query will be provided in details (figure 30).

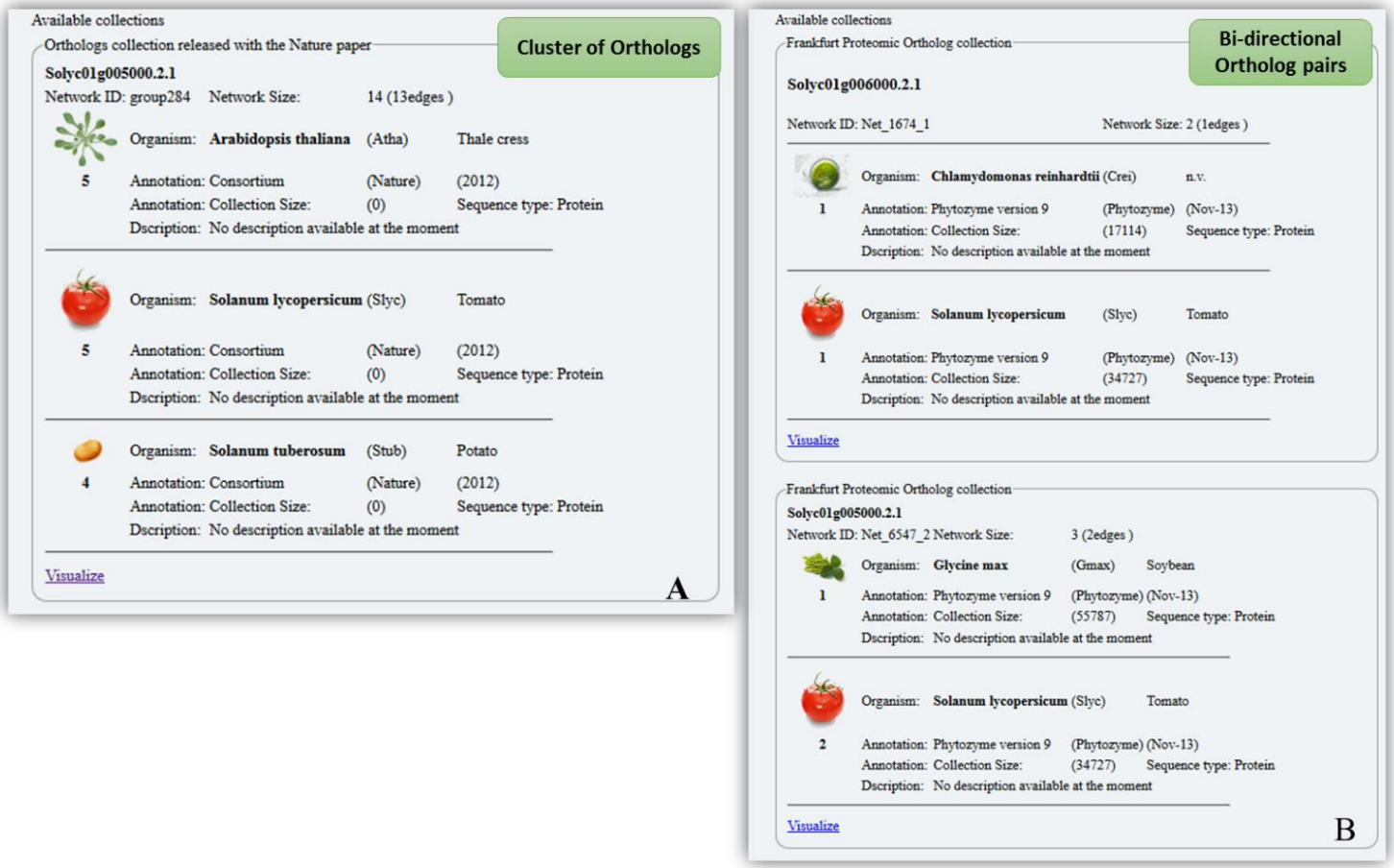


Figure 40: An example output of orthologs platform for A) the orthologs group relationship representation, B) the one-to-one orthologs representation.

Figure 40: An example output of orthologs platform for A) the orthologs group relationship representation, B) the one-to-one orthologs representation., illustrates the two possible ways of ortholog collections representation for a specific query. Snapshot A shows the ortholog group including the results matching the query keyword. As an example, if a gene was matching the query, all the orthologs in the same ortholog group with that id are listed categorized by the species name. The detailed information and the number of members in each species are also listed in details. Snapshot B presents the pairwise orthologs that one of them was matching the query criteria. In this case, the bidirectional ortholog pairs showing the relationship of each pair is presented in details. Visualization of the ortholog graphs are also provided in the platform.

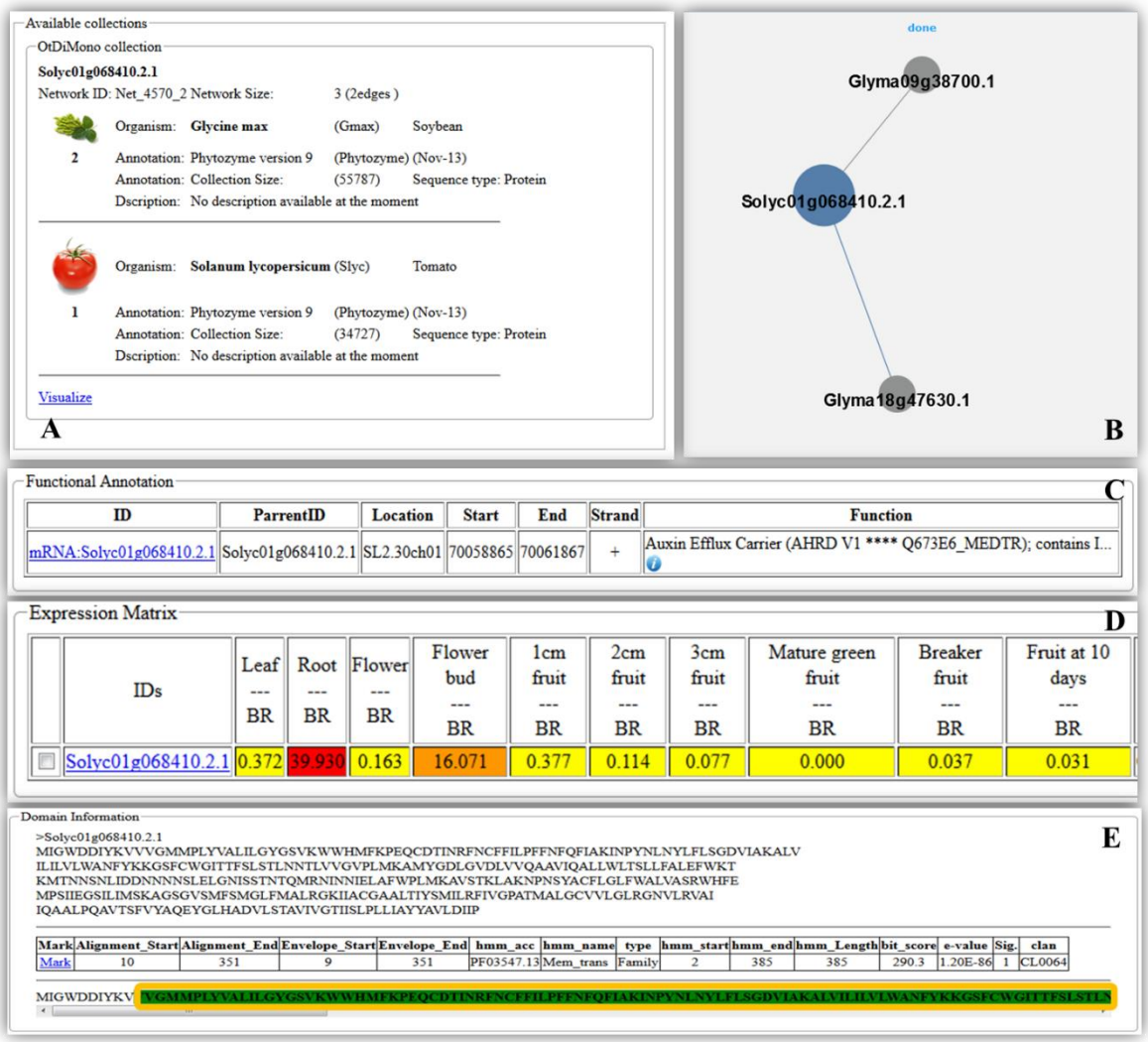


Figure 41: A sample orthologs platform results presentation in orthologs platform. A) orthologs collection, species name, description etc., B) the network visualization of the resulted orthologs, C) functional annotation of the selected ortholog pair, D) the expression profile of the selected ortholog pair in RPKM normalized value, and E) the sequence, and the domain information regarding the ortholog pair sequence.

Figure 41: A sample orthologs platform results presentation in orthologs platform. A) orthologs collection, species name, description etc., B) the network visualization of the resulted orthologs, C) functional annotation of the selected ortholog pair, D) the expression profile of the selected ortholog pair in RPKM normalized value, and E) the sequence, and the domain information

regarding the ortholog pair sequence. shows the possible results representation for an ortholog group. The collection information (A), the network visualization (B), the functional annotation of each element (C), the RPKM expression of the element (retrieved from the NexGenEx- Database) (D), and the sequence and domain information for that element is also presented in details.

3.4.4 Enrichment Tool

To support the functional investigation and GO Enrichment analyses, an online tool (implemented in the bioinformatics infrastructure presented) was implemented with a query interface. The tool includes a user friendly query interface with the implementation of the enrichment analyses on the basis of Fisher Exact Test (see 1.2.5.1.7.1) to provide the GO Enrichment of the selected set. It also provides the Go to Terms association or each gene with the cross link to the *AmiGo* [160] database as for as the *NexGenEx*- platform. Here we present the user interface and the result view for an example application (40 genes highly expressed in *Ascorbic Acid* pathways).

User interface and Database Access

As presented in Figure 42, a graphical user interface to allow the GO functional investigation for a selected set of gene is provided. By selecting the genome and the genome reference of interest, the available gene annotation and the GO dataset associated to that annotation can be selected for the following analyses. The Significance threshold (FDR) for the Fisher Exact Test can be also specified.

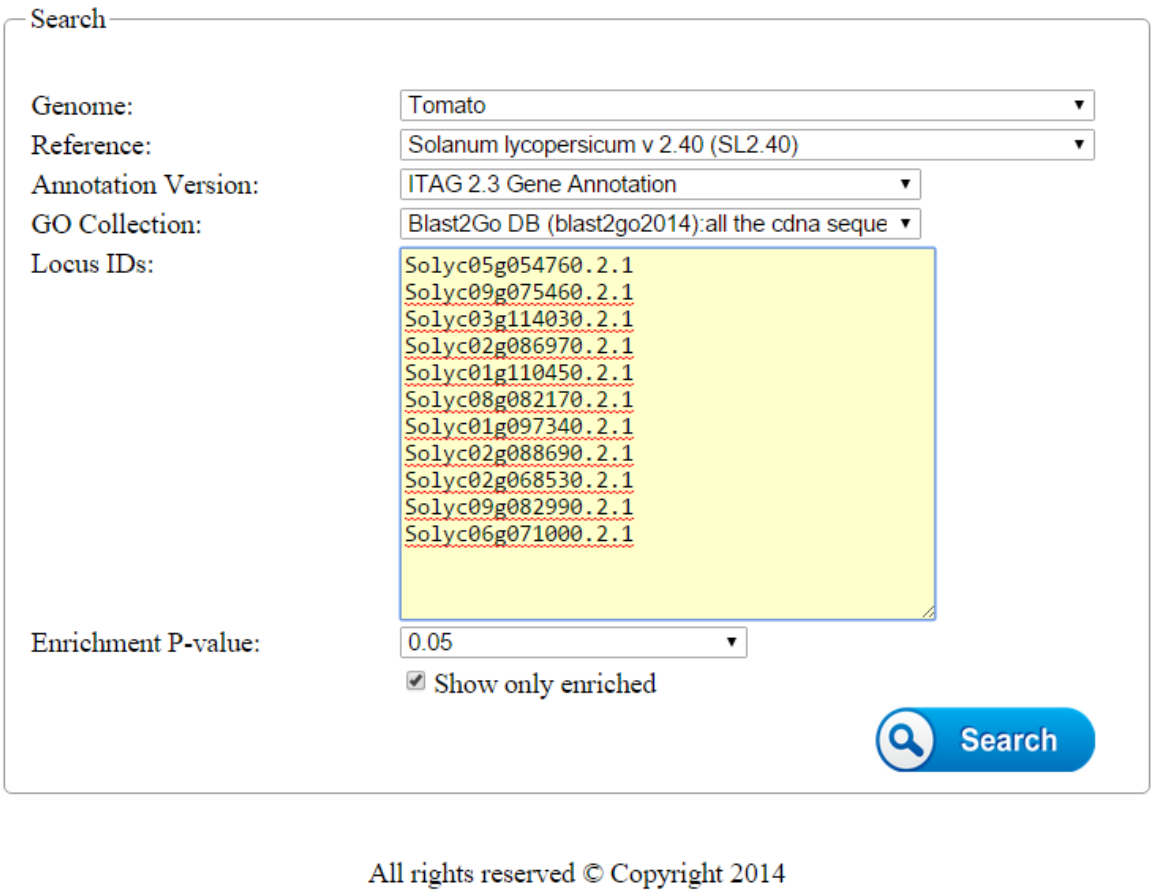


Figure 42: Snapshot of GO Enrichment Analyses tool implemented in the Bioinformatics infrastructure developed

Querying the set of genes (e.g.: 40 genes highly expressed in *Ascorbic Acid* pathway), the list of GOs enriched in this gene set will be provided (Figure 43). the result section also provides the GO ID, adjusted p-value associated to the enriched GO, the number of GOs in the set, the total number of GOs in the collection excluding this set, and the type of the GO (MF= Molecular function, BP= Biological Process, and CC= Chemical Compound). Getting advantage of the information available in the data set, it also provides the description and annotation of the GO term, if available.

GO Summary

Enriched?	GO Term	p-value	nr. InSet	nr. Total	Type	Description	Annotation
Yes	GO:0004028	0.000000	4	9	MF	3-chloroallyl aldehyde dehydrogenase activity	aldehyde dehydrogenase family 2 member mitochondrial-like
Yes	GO:0004029	0.000000	5	6	MF	aldehyde dehydrogenase (NAD) activity	aldehyde dehydrogenase family 2 member mitochondrial-like
Yes	GO:0008928	0.000000	3	1	MF	mannose-1-phosphate guanylyltransferase (GDP) activity	gdp-d-mannose pyrophosphorylase
Yes	GO:0009225	0.000000	4	12	BP	nucleotide-sugar metabolic process	gdp-mannose 3 -epimerase
Yes	GO:0016688	0.000000	4	6	MF	L-ascorbate peroxidase activity	l-ascorbate peroxidase
Yes	GO:0019853	0.000000	6	17	BP	L-ascorbic acid biosynthetic process	gdp-mannose 3 -epimerase
Yes	GO:0055114	0.000000	17	1572	BP	oxidation-reduction process	l-galactose dehydrogenase
Yes	GO:0005829	0.000001	13	1481	CC	cytosol	gdp-mannose 3 -epimerase
Yes	GO:0001758	0.000011	2	1	MF	retinal dehydrogenase activity	aldehyde dehydrogenase family 2 member mitochondrial-like
Yes	GO:0051287	0.000018	4	81	MF	NAD binding	gdp-mannose 3 -epimerase
Yes	GO:0003983	0.000021	2	2	MF	UTP:glucose-1-phosphate uridylyltransferase activity	probable udp-n-acetylglucosamine pyrophosphorylase-like
Yes	GO:0004475	0.000021	2	2	MF	mannose-1-phosphate guanylyltransferase activity	gdp-d-mannose pyrophosphorylase
Yes	GO:0016656	0.000052	2	4	MF	monodehydroascorbate reductase (NADH) activity	probable monodehydroascorbate cytoplasmic isoform 2-like
Yes	GO:0080046	0.000052	2	4	MF	quercetin 4'-O-glucosyltransferase activity	gdp-l-galactose phosphorylase
Yes	GO:0003979	0.000096	2	6	MF	UDP-glucose 6-dehydrogenase activity	uridine diphosphate glucose dehydrogenase
Yes	GO:0006021	0.000096	2	6	BP	inositol biosynthetic process	myo-inositol-1-phosphate synthase
Yes	GO:0050378	0.000096	2	6	MF	UDP-glucuronate 4-epimerase activity	udp-glucuronate 4-epimerase 1-like
Yes	GO:0046686	0.000236	6	485	BP	response to cadmium ion	nadph dependent mannose 6-phosphate reductase

Download GO Enrichemnt

Figure 43: Snapshot of GOs Enriched in the selected gene set including the 1) Enrichment flag (yes/No), GO Term, adjusted p-value of the enrichment test for the corresponding GO, number of GO found in the set (InSet), total number of GOs in the whole set excluding the InSet number, yupe of GO (MF= Molecular function, BP= Biological Process, and CC= Chemical Compound), GO description and annotation.

Each GO is also linked to the *AmiGo* [160] database for the further investigation on the GOs Enriched in the set. The genes to GO association of the selected gene set also is provided including the GO description and annotation (Figure 44).

GO Genes association

ID	GO Term	Type	Description	Annotation
Solyc01g097340.2.1	GO:0005829	CC	cytosol	gdp-mannose 3 -epimerase
Solyc01g097340.2.1	GO:0009225	BP	nucleotide-sugar metabolic process	gdp-mannose 3 -epimerase
Solyc01g097340.2.1	GO:0019853	BP	L-ascorbic acid biosynthetic process	gdp-mannose 3 -epimerase
Solyc01g097340.2.1	GO:0047918	MF	GDP-mannose 3,5-epimerase activity	gdp-mannose 3 -epimerase
Solyc01g097340.2.1	GO:0051287	MF	NAD binding	gdp-mannose 3 -epimerase
Solyc01g106450.2.1	GO:0004033	MF	aldo-keto reductase (NADP) activity	l-galactose dehydrogenase
Solyc01g106450.2.1	GO:0004353	MF	glutamate dehydrogenase [NAD(P)+] activity	l-galactose dehydrogenase
Solyc01g106450.2.1	GO:0005829	CC	cytosol	l-galactose dehydrogenase
Solyc01g106450.2.1	GO:0006520	BP	cellular amino acid metabolic process	l-galactose dehydrogenase
Solyc01g106450.2.1	GO:0010349	MF	L-galactose dehydrogenase activity	l-galactose dehydrogenase
Solyc01g106450.2.1	GO:0016633	MF	galactonolactone dehydrogenase activity	l-galactose dehydrogenase
Solyc01g106450.2.1	GO:0019853	BP	L-ascorbic acid biosynthetic process	l-galactose dehydrogenase
Solyc01g106450.2.1	GO:0050235	MF	pyridoxal 4-dehydrogenase activity	l-galactose dehydrogenase
Solyc01g106450.2.1	GO:0055114	BP	oxidation-reduction process	l-galactose dehydrogenase
Solyc01g110450.2.1	GO:0005829	CC	cytosol	nadph dependent mannose 6-phosphate reductase
Solyc01g110450.2.1	GO:0046686	BP	response to cadmium ion	nadph dependent mannose 6-phosphate reductase
Solyc01g110450.2.1	GO:0047641	MF	aldose-6-phosphate reductase (NADPH) activity	nadph dependent mannose 6-phosphate reductase
Solyc01g110450.2.1	GO:0055114	BP	oxidation-reduction process	nadph dependent mannose 6-phosphate reductase
Solyc01g111510.2.1	GO:0005739	CC	mitochondrion	l-ascorbate peroxidase
Solyc01g111510.2.1	GO:0005774	CC	vacuolar membrane	l-ascorbate peroxidase
Solyc01g111510.2.1	GO:0005778	CC	peroxisomal membrane	l-ascorbate peroxidase

Figure 44: Snapshot of GO to Gene association for the selected gene set.

All the result sets (GO enrichment and Go to Gene association) can be downloaded independently in the excel format for the further references.

3.5 Applications

Setting up integrative resources to allow exploration and exploitation of the data, improvement of quality and enhancement of data access, and eventually converting the data into meaningful information are the main key aspects that made us to pursue these efforts. As presented up to know, various data analysis pipelines, databases and web platforms were designed and implemented. Each tool or resource, or their combination, support specific objectives in which a specific biological question is addressed.

At the light of the *SPOT-ITN* objectives, some example applications are here presented to highlight results achieved and the relevance of such tools and integrated resources.

We focused on understanding the data we were dealing with.

Efforts on the mining of the EST and TC collections we included in the tomato genome platform, on the quality of the available versions of the tomato genome and its annotations were reported. Since we identified several issues in the tomato gene annotation, we also present our findings and results for the gene annotation improvement and revision.

These preliminary investigations were also useful for our understanding of data from the heat stress response in the tomato pollen developmental stages based on transcriptome analyses. Then, an extensive analyses on the role of TE-derived Small-RNAs interfering RNAs in pollen developmental stages is presented.

3.5.1 Experimental Transcript Collections

ESTs vector and repeats cleaning

As it was discussed in section "*ESTs, TCs and TIs data processing*", each EST collection was subjected to vector removal and repeat masking process. This procedure supports the production of high quality and clean datasets representing the transcriptomics data for each species. Table 10 presents the proportion of remaining EST sequences after this quality check and filtering in our database, which were then used for the further analysis.

Table 10: The number of sequences in the starting datasets. A) the starting number of EST sequences in each species collection downloaded from the reference database without any processing, B) the starting number of EST sequences for each collection species after vector

removal and repeat masking together with the percentage of remaining sequences with respect to the initial dataset.

Species Code	Starting Seq. No.	Vector&Repeat Removed Seq. No (%)
TOBAC	334809	330311 (98.66%)
SOLLC	298370	297451 (99.69%)
SOLTU	250127	249974 (99.94%)
COFAR	174275	172921 (99.22%)
CAPAN	118651	118597 (99.95%)
SOLME	98089	98086 (100%)
COFCA	69066	68806 (99.62%)
NICBE	56180	56019 (99.71%)
PETHY	50705	50605 (99.8%)
SOLTO	28743	28731 (99.96%)
SOLHA	26019	25916 (99.6%)
NICLS	12537	12533 (99.97%)
SOLPN	10946	10935 (99.9%)
NICSY	8583	8574 (99.9%)
SOLCH	7752	7731 (99.73%)
SOLPH	2099	2099 (100%)
SOLLP	1008	1007 (99.9%)
CAPCH	442	437 (98.87%)
NICAT	355	352 (99.15%)
SOLPE	69	69 (100%)

As presented in Table 10, at worst less than 0.2 % of each collection was discarded due to the quality check and sequence cleansing process. The remaining proportion of each collection as then used for the downstream analysis. In addition to highlight the impact and contribution of each collection species, we can observe that among all the 20 Solanaceae EST collections available in the platform, *Nicotiana tabacum, Solanum lycopersicum, Solanum tuberosum, Coffea arabica* and *Capsicum annumm* represent a large collection of EST sequences (more than 75% of the total ESTs in all the 20 collections)

while the other species contribute the other 15% of the EST coverage in our datasets.

EST to TC Assembly

With respect to the aim of having more reliable transcriptome datasets confirmed with multiple sequences for each consensus (section 1.2.5.1.3), here a complete overview of the EST to TC assembly and the remaining singletons, with the number of protein matches for each specific set per collection species, is provided in details.

Table 11: The proportion of EST to TC assembly for each collection species and their protein matches

Species Code	ESTs Assembled	Singletons	No. TC Seq.	EST.Dist. Protein Match	TC.Dist. Protein Match
TOBAC	215889 (65.36%)	118920 (36%)	31199	193737 (58.65%)	18871 (60.49%)
SOLLC	251810 (84.66%)	46560 (15.65%)	21920	214658 (72.17%)	15622 (71.27%)
SOLTU	186296 (74.53%)	63831 (25.54%)	23120	170000 (68.01%)	16115 (69.7%)
COFAR	140157 (81.05%)	34118 (19.73%)	16417	117201 (67.78%)	11487 (69.97%)
CAPAN	94778 (79.92%)	23873 (20.13%)	12247	81090 (68.37%)	8689 (70.95%)
SOLME	80354 (81.92%)	17735 (18.08%)	13500	63585 (64.83%)	8993 (66.61%)
COFCA	52800 (76.74%)	16266 (23.64%)	8287	45561 (66.22%)	5723 (69.06%)
NICBE	38745 (69.16%)	17435 (31.12%)	6817	35317 (63.04%)	4691 (68.81%)
PETHY	38920 (76.91%)	11785 (23.29%)	9569	31736 (62.71%)	6500 (67.93%)
SOLTO	21783 (75.82%)	6960 (24.22%)	4029	20569 (71.59%)	2881 (71.51%)
SOLHA	16276 (62.8%)	9743 (37.59%)	2779	16732 (64.56%)	2034 (73.19%)
NICLS	7348 (58.63%)	5189 (41.4%)	1401	7261 (57.94%)	983 (70.16%)
SOLPN	6673 (61.02%)	4273 (39.08%)	1136	6685 (61.13%)	802 (70.6%)
NICSY	1949 (22.73%)	6634 (77.37%)	701	5134 (59.88%)	527 (75.18%)
SOLCH	1410 (18.24%)	6342 (82.03%)	591	4717 (61.01%)	457 (77.33%)
SOLPH	402 (19.15%)	1697 (80.85%)	178	1609 (76.66%)	159 (89.33%)
SOLLP	430 (42.7%)	578 (57.4%)	112	755 (74.98%)	102 (91.07%)
CAPCH	108 (24.71%)	334 (76.43%)	26	273 (62.47%)	19 (73.08%)
NICAT	44 (12.5%)	311 (88.35%)	19	129 (36.65%)	10 (52.63%)
SOLPE	28 (40.58%)	41 (59.42%)	6	56 (81.16%)	5 (83.33%)

As it is presented in Table 11, each EST collection species was resulted into a collection of assembled EST and Singleton (those did not assemble) sequences. The total number of EST sequences for each dataset contributing in the

assembly and the resulted number of TC sequences is also reported in details. Obviously, those datasets with higher number of EST sequences (e.g. *TOBAC*, *SOLLC* and *SOLTU*) provide higher number of TC assemblies in comparison to those with lower coverage (e.g. *SOLPE*, *NICAT* and *CAPCH*). In addition, the number of EST and TC sequences from each collection specie having at least one match with a protein (blast procedure described in *"BLAST"* section) is presented accordingly. Stats shows that except few data collections, more than about 70% of the TC sequences found a protein match while this statement is not true for the EST sequences. This shows the reliability of the TC sequences in comparison with the ESTs. In addition, the large contribution of the EST sequences in the assembly for each TC collection is another indication of the reliability for these assembled datasets.

EST Mapping

on the bases of the mapping procedure presented in "EST, TC and unignees processing" section, here we present the mapping statistics gained form the mapping of each EST collection species on the both versions of Tomato Genome *SL2.40* and *SL2.50*, and the BAC sequences un-mapped on the genome.

BAC sequences are considered to provide to complement the genome sequences to allow more exhaustive investigations.

Table 12: Overview of the EST collections mapping on the both Genome sequences of Tomato versions ITAG 2.40 and ITAG 2.50, and the unmapped BAC sequences

Species Code	Vector&Repeat Removed EST No.	Mapped On 2.40	Mapped On 2.50	Dist.Mapped On 240	Dist. Mapped On 2.50	Mapped On the Unmapped BACs	Dist. Mapped on the Un-Mapped BACs
SOLLC	297451	294474	294512	273894	273897	14465	11594
SOLTU	249974	150439	150418	142966	142956	6585	5546
TOBAC	330311	38577	38601	34792	34796	1497	1274
CAPAN	118597	31119	31115	28979	28980	1179	1009
SOLME	98086	30996	31048	27718	27733	1394	1182
SOLHA	25916	20396	20375	16889	16891	1265	820
SOLTO	28731	10582	10609	8892	8895	0	0
SOLPN	10935	9384	9403	8492	8492	448	379
PETHY	50605	6077	6084	5580	5583	275	221
NICBE	56019	5749	5750	4917	4913	345	265
SOLCH	7731	3128	3128	3073	3073	135	125
NICSY	8574	1651	1651	1553	1553	64	56
SOLPH	2099	1421	1420	1385	1385	52	48
NICLS	12533	841	843	692	692	57	45
COFAR	172921	881	888	438	438	20	20
COFCA	68806	285	296	160	160	22	16
SOLLP	1007	160	160	149	149	0	0
CAPCH	437	87	87	82	82	2	2
SOLPE	69	29	29	29	29	0	0
NICAT	352	27	27	22	22	0	0

As presented in Table 12, the complete overview of the EST collections mapping on the two versions of *SL2.40* and *SL2.50* tomato genome, and the SL2.3 unmapped BACs (see 2.3.1) is presented. As it can be observed, some collections such as *SOLLC, SOLTU, TOBAC, CAPAN and SOLME* have high coverage while some such as *SOLLP*, *CAPCH*, *SOLPE* and *NICAT* have very low coverage of ESTs mapping on the reference sequences. Aside from the closeness or distance between the two species, this can also be due to the staring number of sequences available in each dataset.

We also intended to report the number of total mapping and the distinct number of transcripts mapping on the genome to provide a brief indication of the redundancy on the reference sequences with respect to the transcript collections. In other words, the ratio between the number of mapping transcripts versus the number of distinctly mapped transcripts can be a parameter to detect the remapping and redundancy of mapping for the transcripts on the reference genome (as much as the ratio higher, the redundancy of mapping higher).

TC Mapping

Here we present a general overview of the TC collections mapping on the both versions of *Tomato Genome SL2.40 and SL2.50, and the BAC sequences unmapped on the genome.*

Table 13: Summary of TC collection species mapping on the both Tomato genome versions ITAG 2.40 and ITAG 2.50, and the unmapped BAC sequences

Species Code	TC Starting Sequence	Mapped On 2.40	Mapped On 2.50	Dist. Mapped On 2.40	Dist. Mapped On 2.50	Mapped On the Unmapped BACs	Dist. Mapped on the Un-Mapped BACs
SOLLC	21920	21068	21074	19754	19753	901	698
SOLTU	23120	12759	12751	12507	12505	496	439
TOBAC	31199	1602	1604	1525	1526	66	64
CAPAN	12247	2239	2239	2187	2187	85	75
SOLME	13500	3212	3212	3104	3105	128	112
SOLHA	2779	2475	2461	2058	2058	118	82
SOLTO	4029	1166	1169	1077	1078	0	0
SOLPN	1136	1036	1034	963	963	46	40
PETHY	9569	1131	1131	1071	1070	59	50
NICBE	6817	444	445	415	415	22	17
SOLCH	591	328	328	321	321	10	10
NICSY	701	172	172	160	160	4	3
SOLPH	178	127	127	126	126	4	4
NICLS	1401	96	96	83	83	3	3
COFAR	16417	26	27	15	15	0	0
COFCA	8287	21	21	12	12	2	2
SOLLP	112	2	2	2	2	0	0
CAPCH	26	3	3	2	2	0	0
SOLPE	6	0	0	0	0	0	0
NICAT	19	0	0	0	0	0	0

The mapping summary of TC collections per each species (Table 13) provides the coverage and proportion of each collection for the tomato genome and the BAC sequences unmapped on the genome reference. Similar to the ESTs, the tomato species of *SOLLC, SOLME, SOLHA*, and *SOLPH* have the highest relative number of transcripts with respect to the starting collection size covering the reference sequences. Moreover, SOLTU also as the closest species to SOLLC also shows the relevant high coverage of mapping on the tomato

genome and the BAC sequences. Respectively, other Solanaceae species also are reported on the sense of mapping with respect to the initial collection size. It is also nice to mention that some TCs are also mapped on the unmapped BAC sequences not considered in the official genome references. This can be another source of information in which the sequence regions not present in the official genome reference can be also investigated.

Unigenes Mapping

A detailed mapping overview of the three unigene collections of *SGN*, *DFCI* and *PlantGDB* are provided here.

Table 14: Overview of Transcript Indices (unigenes) collections from 3 reference websites of SGN DFCI and PlantGDB mapped on both tomato genome versions of ITAG 2.40 and 2.50, and the unmapped BAC sequences

Species Name	*Database*	*Starting Sequence Number*	*Mapped On 2.40*	*Mapped On 2.50*	*Dist. Mapped On 2.40*	*Dist. Mapped On 2.50*	*Mapped On the Unmapped BACs*	*Dist. Mapped on the Un-Mapped BACs*
SOLLC	***DFCI***	52502	50790	52823	45266	46692	2575	1796
SOLLC	***PlantGDB***	56845	55955	57075	49993	51032	2855	1940
SOLLC	***SGN***	42257	41789	42625	36393	37203	2062	1418

As for the TC collections, the overview of the unigene collections mapped on each version of the Tomato Genome and the unmapped BACs are presented in the redundant and distinct manner (Table 14). Interestingly, around 4% of each collection is also mapped on the unmapped BACs which provides the information regarding the transcripts annotated on the regions not anchored in the chromosomes. In the following section, the number of transcripts from each collection mapped uniquely on these BAC sequences will be discussed in details.

TCs and unigenes mapping uniquely on the reference sequences

Table 15: Overview of TC collections uniquely Mapped on the both genome sequences and UnMapped BACs

Species	TC Unique-Map_Chr 2.40	TC Unique-Map_Chr 2.50	TC Unique-Map-BAC 2.40	TC SUM(Unique-Map-Chr&BAC) 2.40	TC Unique-Map-On-BACs-NotOn-Chr 2.40
SOLLC	19272	19270	582	19854	8
SOLTU	12305	12306	386	12691	4
TOBAC	1481	1482	62	1543	3
CAPAN	2142	2142	65	2207	2
SOLME	3026	3027	97	3123	6
SOLHA	1990	1990	61	2051	8
SOLPN	919	919	34	953	
SOLTO	1018	1019	37	1055	1
PETHY	1018	1017	44	1062	2
NICBE	395	395	15	410	1
SOLCH	314	314	10	324	
NICSY	149	149	2	151	
SOLPH	125	125	4	129	
NICLS	76	76	3	79	
COFAR	9	9		9	
SOLLP	2	2		2	
COFCA	8	8	2	10	
CAPCH	1	1		1	
SOLPE				0	
NICAT				0	

Having a transcript mapped uniquely on a genomic region can confirm the

origin of the transcript. Table 15 provides the number of transcripts in each of the TC collections uniquely mapped on a genomic reference. Interestingly, we can observe some transcripts mapped on the unmapped BAC sequences (not included in the official reference) with no other copy or map on any of the Tomato chromosomes.

The same information also is presented in the sense of unigene collections.

Table 16: Unigenes uniqueness mapping overview on the genomes and BACs

Species	Unique-Map_Chr 2.40	Unique-Map_Chr 2.50	Unique-Map-BAC 2.40	SUM(Unique-Map-Chr&BAC) 2.40	Unique-Map-On-BACs-NotOn-Chr 2.40
SOLLC *DFCI*	43852	45189	1440	45292	54
SOLLC *PlantGDB*	48368	49383	1551	49919	32
SOLLC *SGN*	35172	35969	1130	36302	30

Table 16 shows the number of transcript form each unigene collection mapped uniquely on the tomato genome and the unmapped BAC sequences. As well as the TCs, we can see that some transcripts are uniquely mapped on the BAC sequences in which the information is not considered in the official genome reference.

Considering the total number of transcripts from all the TC and unigene collections (Table 7) and those uniquely mapped on the genome (tables of unique **Error! Reference source not found.**), we can conclude that the majority of the transcripts are mapped uniquely on the genome. This is an indication of the transcripts quality, reliability and specificity in each dataset.

The processing of the EST collections to produce cleaned datasets, assembly of

the TC sequences to have more reliable transcripts confirmed by multiple sequences, and having the unigene collections from the major reference databases can provide useful transcript resources for the genomic analyses and investigations. Moreover, the availability of different species collections, all mapped on a unique reference genome (here *S. lycopersicum*), can allow the cross species analyses and functional investigations. These collections are great supports to the assessment, assignment and characterization of the genomic features (e.g. assessing the miss-annotated genes such as split, very long etc.).

3.5.2 Genome Reference and Gene annotations

A reliable genome reference is the basis for genome centered approaches and "omics" analyses. Hence, a good understanding of its quality and content tracing its improvements is fundamental for appropriate investigations. In the light of the *SPOT-ITN* project we set up the tomato genome platform to support the data analyses.

Moreover, we deeply investigated the two different genome versions of tomato. Furthermore we tested the quality of the annotations available (*iTAG* and *RefSeq*) to define a reference annotation too.

ITAG 2.40 vs ITAG 2.50

For tomato, two different versions of the *S. lycopersicum* genome sequences have been currently released (*SL2.40* [137], *SL2.50* [143]). The genome version *SL2.50* was made available on the *SGN* website on 2014 (announced in [143]). It represents the updated version of the first version *SL2.40* [137], release in 2012 by the Consortium.

An overview reveals that the chromosomal lengths in *SL2.50* is increased comparing to the *SL2.40* (Table 17, column length). In other words, long pieces were inserted/added to the new version of the tomato genome (*SL2.50*).

Interestingly, looking at the number of "N" added to each chromosome, which is exactly similar to the corresponding added lengths, it is clear that the added regions are only filled by "*Ns*" to improve the genomic distances in the genome.

Table 17: An overview of SL2.50 versus SL2.40 genome version

SeqID	Length	#A	#T	#C	#G	#N
ch00	0	0	0	0	0	0
ch01	8239200	-1858	1858	-10148	10148	8239200
ch02	5422150	-9705	9705	-13141	13141	5422150
ch03	5946950	-2289	2289	15381	-15381	5946950
ch04	2406630	-16881	16881	3450	-3450	2406630
ch05	853650	-18636	18636	-1162	1162	853650
ch06	3710000	-13028	13028	-10145	10145	3710000
ch07	2776400	0	0	0	0	2776400
ch08	2834000	1299	-1299	-31808	31808	2834000
ch09	4820000	-6147	6147	-3522	3522	4820000
ch10	693200	0	0	0	0	693200
ch11	2916500	21144	-21144	38971	-38971	2916500
ch12	1658950	9569	-9569	24075	-24075	1658950

Looking at the exact proportion of the positive and negative values in each A and T (i.e.; A= -1885 and T= -1885 in ch01), and G and C bases content (C= -10148 and G= 10148 for ch01), a total of the inverted regions per each chromosome can be observed.

We also observed that the changes in the GC, CG and CpG was zero comparing the two versions of the genome. So we concluded that no new genome sequence representing ATGC bases was added in the new version of the tomato genome *SL2.50* comparing the *SL2.40*.

These analyses raised the question if the reorganization of the genome had affected the gene content.

iTAG 2.3 and iTAG 2.4

We also check the differences between the *iTAG 2.4* and 2.3 gene annotations. Comparing the *iTAG 2.3* and *iTAG 2.4* gene annotations to understand the differences and peculiarities of the two annotations, we observed that two genes of *Solyc03g053140.1.1* and *Solyc12g032910.1.1* were discarded from the newer version.

It is important to highlight that due to the addition of "*N*" insertions to the genome, the genes in *iTAG 2.4* were shifted in their genomic position. This means that the position indicated in the GFF3 annotation file is changed on the basis of the insertion lengths occurring before the specific gene on the genome. However that the genomic locus is identical in the sense of the genomic sequence.

RefSeq

The RefSeq GFF3 file was tested for the standard format and compatibility for the visualization in the *Gbrowser*. We also checked the number of the genomic features available in the annotation. The stats on the produced file are as follow:

Table 18: Stats on the RefSeq2.3 genomic features and their coverage

Feature Type	Count	Length
Gene	24528	122350718
mRNA	25946	128880502
Mature Transcrips	646	2399912
Exon	150609	43240563
CDS	142115	33604313

The total coverage of the gene features for the 12 tomato chromosomes in the genome are presented in Table 18. *RefSeq* Annotation does not include the gene prediction for the *S. lycopersicum* unassigned chromosome (chr00).

We also aimed to compare the two available gene predictions for the tomato genome (*iTAG* and RefSeq) to understand their similarities, differences and whether the two annotations for the same genome confirm each other. The result of the analyses are presented below:

iTAG 2.3 vs RefSeq 2.3

Table 19: Stats on the iTAG2.3 genomic features and their coverage

Feature Type	Count	Length
gene	34727	109860926
mRNA	34727	109860926
CDS	157239	35972459
exon	160007	41982942
intron	125280	67877984
five_prime_UTR	13567	1922626
three_prime_UTR	15378	3576943

As presented above the stats on the number of genomic features and their coverage on the genome for the *iTAG 2.3* is presented in Table 19.

In Table 19 *we report the statistics for iTAG 2.3. iTAG* includes 34727 gene loci representing 34727 mRNAs while *RefSeq* 2.3 includes 24528 representing 26592 mRNA loci. As declared by *ITAG*, the official gene annotation for the *tomato* genome (*iTAG 2.3*) does not include the alternative splicing prediction whilst *RefSeq* 2.3 annotation includes the alternative splicing prediction for the genes. Due to this fact (the availability of alternative transcript in *RefSeq* annotation), the number of genes are 10119 but for mRNA 8135 less than *iTAG 2.3* official annotation.

Deeper investigations on both annotations revealed that in total 1062 mRNAs confirm locus and structure (start and end of all the features such as locus, exons, cds etc.) while only 207 mRNAs confirm only the locus (start and end of the transcript) but not the internal structure between the *RefSeq* and *iTAG* gene annotations. Moreover, overlapping the *iTAG 2.3* and *RefSeq* 2.3 gene annotations 1624 mRNAs from *RefSeq* did not overlapping any *iTAG* predicted loci, and 10931 *iTAG 2.3* mRNAs did not overlapping any *RefSeq* 2.3 predicted mRNA. The significant difference between the *iTAG* and *RefSeq* can be also proportional to the total number of predicted mRNAs in each annotation (see Table 18 andTable 19).

iTAG Annotation Issues

To understand the quality of the official tomato gene annotation, we also made some exhaustive assessment on the *iTAG 2.3* in which several issues raised. Here we present the ambiguities, miss-annotations and the issues detected in the *iTAG 2.3* gene annotation for the *S. lycopersicum* genome.

Very Long Genes

Our analyses resulted to the detection of two very long genes in the annotation in which, the genes span for more than 200,000 pb on the genome.

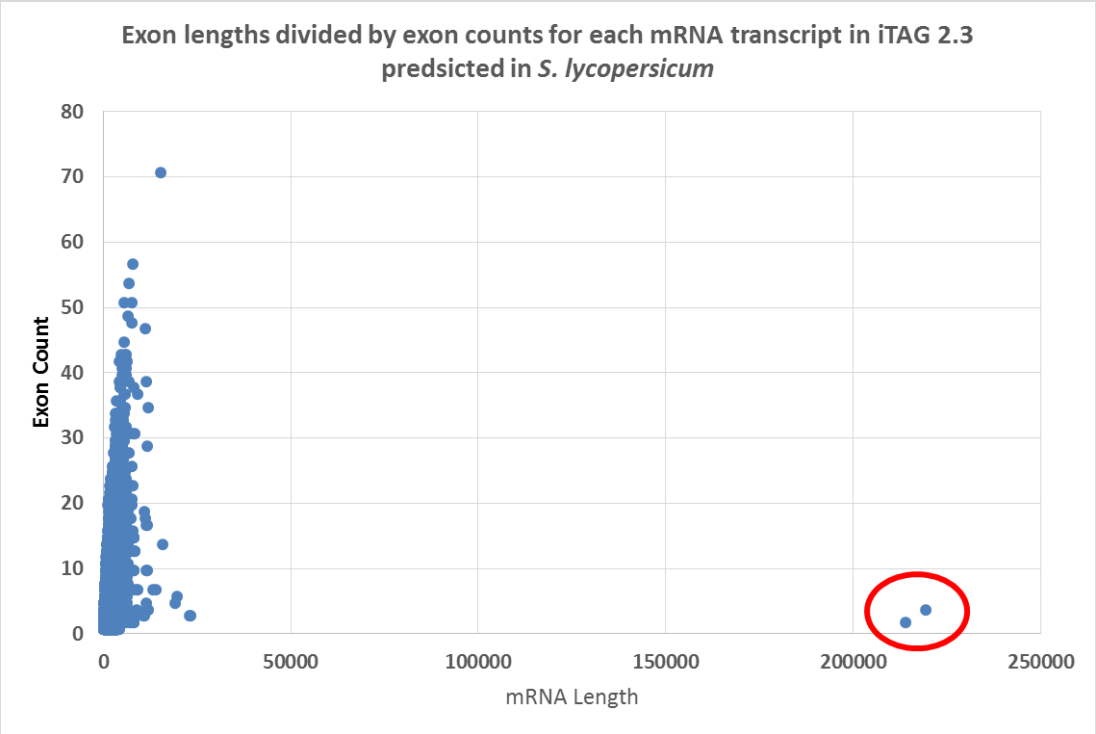


Figure 45: The exon lengths versus exon counts in iTAG 2.3 predicted genes

As it is shown in Figure 45, the number of exons versus the sum of the exons length for each transcript is presented. The plot shows the distribution of the *iTAG 2.3* predicted genes in the tomato genome where the majority of them are less than 25000 bp long. The 2 genes of *Solyc01g110700.2* and *Solyc01g111180.2* possess the length of 244,093 bp and 214,621 bp respectively. In addition, the gene *Solyc01g111080.2*, though much shorter than the 2 mentioned before, also has is a long gene spanning 23681 bp on the genome. A *Gbrowse* snapshot of the three very long genes mentioned are presented in (Figure 46)

Figure 46: A genome browser snapshot of the 3 very long genes and the TC tracks overlapping the locus.

Interestingly, all the three genes are overlapping on their genomic locus also covering multiple *iTAG 3.4* predicted genes. By cross checking these three genes with the *S. lycopersicum* TC collection, the unigene collections form the *SGN*, *PlantGDB* and *DanaFarber*, all available in our *Tomato Genome Platform*; we could not confirm any experimental transcript confirming the predicted structures.

Table 20: Statistics on the UTR's length and overlapping of the 3 very long genes in iTAG predicted genes in S. lycopersicum

Gene ID	Length	3' UTRs Length	5' UTRs Length	Overlaps
Solyc01g110700.2	244,093 bp	64 bp	219069 bp	54
Solyc01g111180.2	214,621 bp	212781 bp	18 bp	49
Solyc01g111080.2	23681 bp	22853 bp	52 bp	5

We also observed that the reason these genes are so long is due to the long UTR regions predicted (Figure 47 and Table 20).

Concerning the overlapping genes, we observed that the gene *Solyc01g091150.2* (118,735 bp) also overlaps 14 other genes, in its intron region, in the annotation (Figure 47).

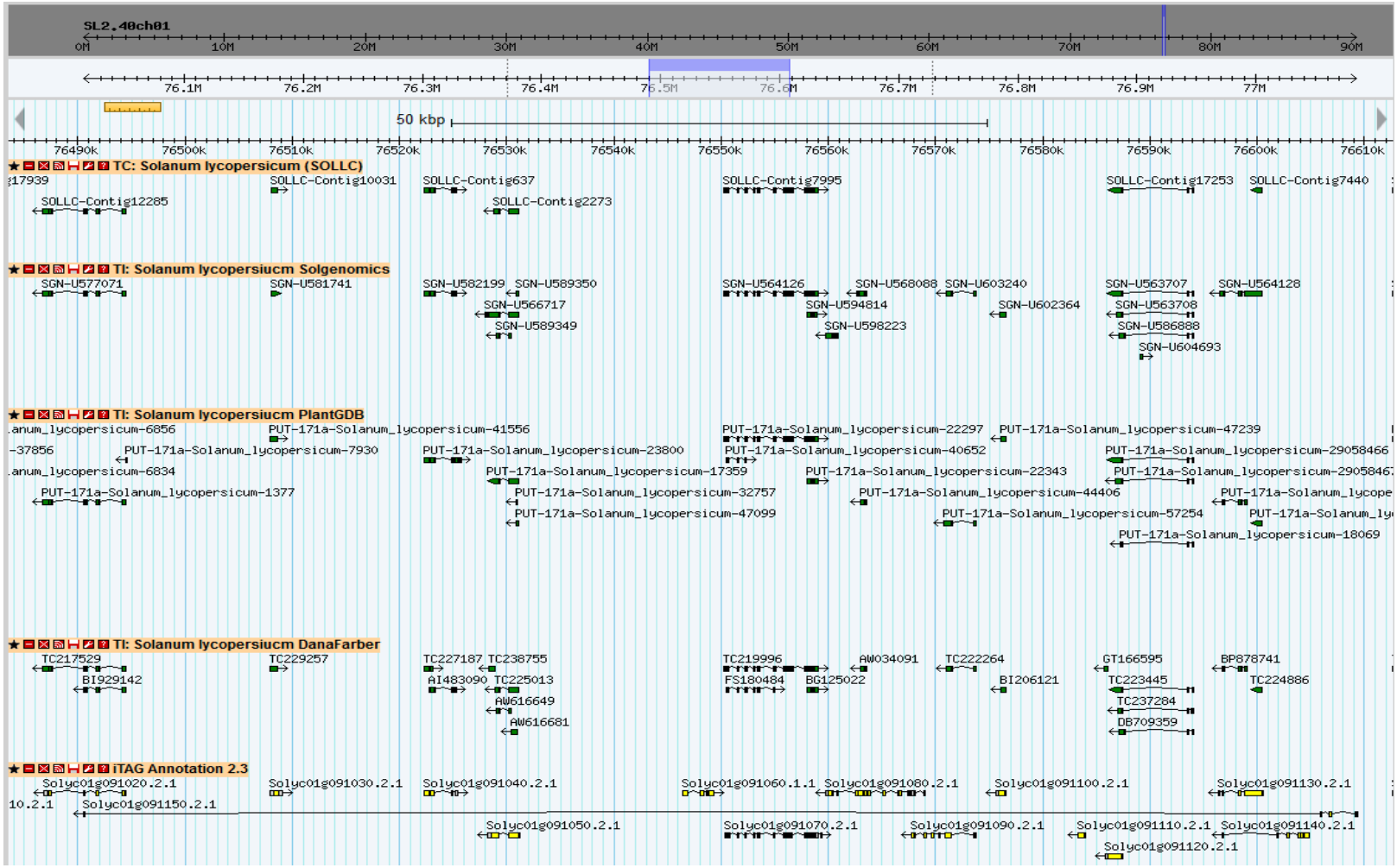


Figure 47: A genome browser snapshot of a long gene covering several iTAG 2.3 annotated loci in its intron

Figure 69 shows the genome browser snapshot for the gene Solyc01g091150.2 overlapping multiple *iTAG 2.3* predicted locus in its intron region. By cross comparing this gene with the available TC and unigene transcript collections in our platform, we could not find any experimental transcript confirming its structure. Hence, probably this gene also is miss-annotated in the *iTAG* annotation.

Three exact Overlapping ***Genes with different CDS regions matching the same Protein***

The further analyze overlapping genes predicted in the *iTAG 2.3* we noticed that three genes (*Solyc01g088200.2.1*, *Solyc01g088210.2.1* and *Solyc01g088230.2.1*), exactly overlapping each other in their exons start and ends but differ completely in their coding regions. An illustration of these three genes overlapping each other on the genome is presented in Figure 48

Figure 48: Demonstration of the 3 iTAG 2.3 predicted genes with exact exonic and different CDS overlapping matching 1 protein consecutively.

Interestingly all the three genes also match the same protein in a consecutive manner (Figure 48). This can be also due to the miss-annotation of one single gene as three different genes on the genome.

Blast of all the iTAG 2.3 mRNAs versus the proteins databank

As the result of our blast analyses for the *iTAG 2.3* mRNA sequences versus the protein databank (downloaded on February 2014), 758 mRNAs with unknown function in the annotation found at least one protein match in the database (Table 21).

Table 21: The blast results of iTAG 2.3 mRNA versus the protein databank

34727 ITAG genes	26059 at least one protein match	758 Unknown genes in ITAG (Can be annotated)
	8671 No protein match	1754 Annotated in ITAG
		6917 Unknown in ITAG

The proteins with match with these 758 unknown annotated mRNAs can provide a putative functional annotation for these transcripts.

Split Genes

The blast analyses of the *iTAG 2.3* mRNA transcripts also revealed split genes in the annotation. We identified 1873 genes that match the same protein in consecutive portions. An example of the split genes is presented in Figure 49

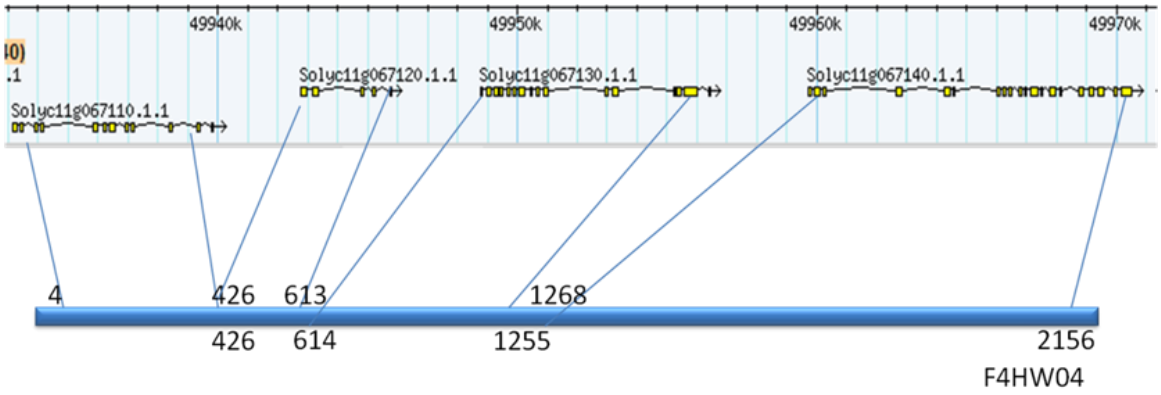


Figure 49: Example of 4 iTAG 2.3 predicted genes matching consecutive regions of a protein.

Figure 49 shows the four *iTAG 2.3* predicted genes, *Solyc11g067110.1.1*, *Solyc11g067120.1.1*, *Solyc11g067130.1.1*, and *Solyc11g067140.1.1*, matching the protein F4HW04 in consecutive regions.

On Repeated Regions (iTAG 2.3 Repeat Aggressive)

The intersection of the *iTAG 2.3* genes and the repeat aggressive annotation resulted to the identification of several genes located in overlapping repeated regions. In Table 22, the number of genes overlapping a specific repeat aggressive class more than 50, 80 and 100% are presented.

Table 22: The summary of iTAG 2.3 predicted genes with the iTAG 2.3 repeat aggressive classes based on the overlapping thresholds of 50, 80 and 100 %of coverage

Repeat Aggr. Class	50%	80%	100%
DNA/En-Spm	2		
DNA/Harbinger	2	1	
DNA/hAT	7	1	1
DNA/MuDR	4	3	2
LINE	8	4	3
Low_complexity	14	2	1
LTR	295	191	143
LTR/Copia	110	85	75
LTR/Gypsy	65	52	47
RC/Helitron	32	12	5
rRNA	46	36	26
Simple_repeat	19	4	
Grand Total	604	391	303

As presented in Table 22, it is interesting to observe that 303 times a gene falls completely inside a repeated region (100% of coverage).

iTAG Remapping onto the tomato genome

By mapping the *iTAG 2.3* mRNA sequences versus the *S. lycopersicum* genome (2.3), we categorized the transcripts into three major groups (once map, multiple map and not mapped on the genome). We also further divided the once and multiple mapped transcripts into two classes of "Confirming their *iTAG* gene structure prediction" and "not confirming their *iTAG* gene structure prediction" (Table 23).

Table 23: The summary of remapping the 34727 iTAG 2.3 mRNA transcripts versus the iTAG 2.40 genome

Total Number of Transcripts (34727)			
Once Mapped	30046	**Confirming Gene Structure**	27968
		Not Confirming Gene Structure	2078
Multiple Mapped	4593	**Confirming Gene Structure**	4165
		Not Confirming Gene Structure	428
No Match	88	**Found by Blast**	62
		Partially Found by Blast	24
		Very Long and Discarded Genes	2

As it can be observed in the Table 23, the majority of *iTAG 2.3* predicted genes (27968) could map uniquely on the genome confirming their own predicted structure. Still 6759 genes either do not confirm their structure or have multiple mappings on the reference genome. To further characterize the remapping status of the annotated gene on each of the chromosomes, we further summarized the results into a bigger table, in which Snapshot from this table is provided in Figure 50

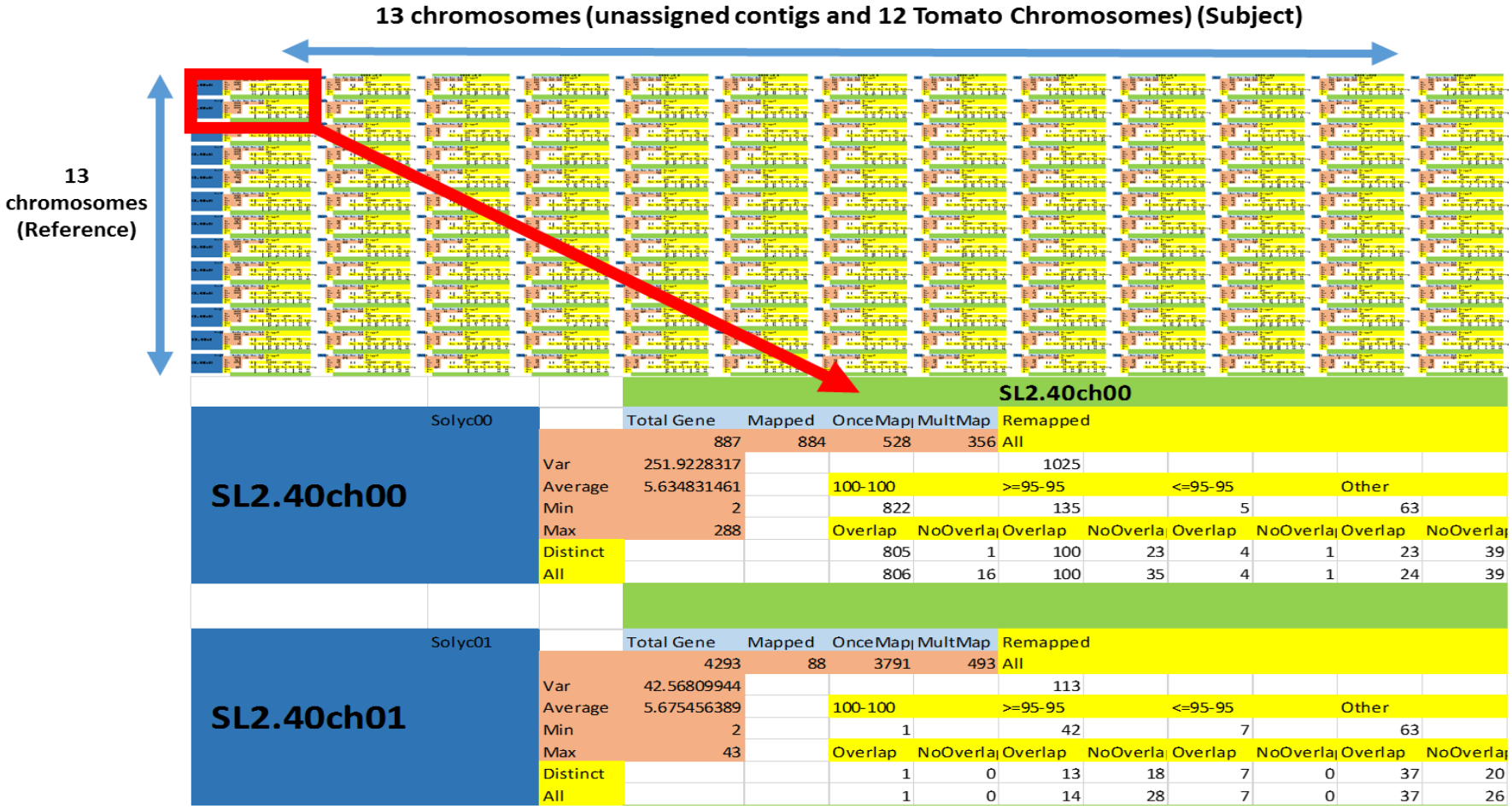


*Figure 50: Snapshot of the 13*13 cross table characterizing the genes remapping from each of the 13 S. lycopersicum chromosomes on oth er chromosomes*

Figure 74 is snapshot of a bigger table which summarizes the behavior of the annotated genes if mapped on all the other chromosomes (See ***ANNEX IV)***. The total number of genes from the starting chromosome, the total, once and multiple number of times they remapped on the target chromosome, the detailed categorization of the remapping statistics on the bases of their identity and coverage when mapping, and whether they overlap an *iTAG* predicted loci or not are listed in a redundant and distinct way (redundant = if the genes mapped four times, four is considered; distinct= if the gene is mapped four times, one is considered). This 13 * 13 dimensional table, summarizing all the *iTAG* genes predicted on the 13 tomato chromosomes versus each other, provides a comprehensive overview of the annotation issues like missing annotation, genome duplication by sequencing miss-assembly, similarity of the genomic regions, and potential new unpredicted gene loci on each chromosome.

To provide a broader overview of the *iTAG 2.3* predicted genes in the sense of their remapping status, Figure 51 demonstrates the frequency of the number of times each gene matched a chromosomal region of the *S. lycopersicum* genome.

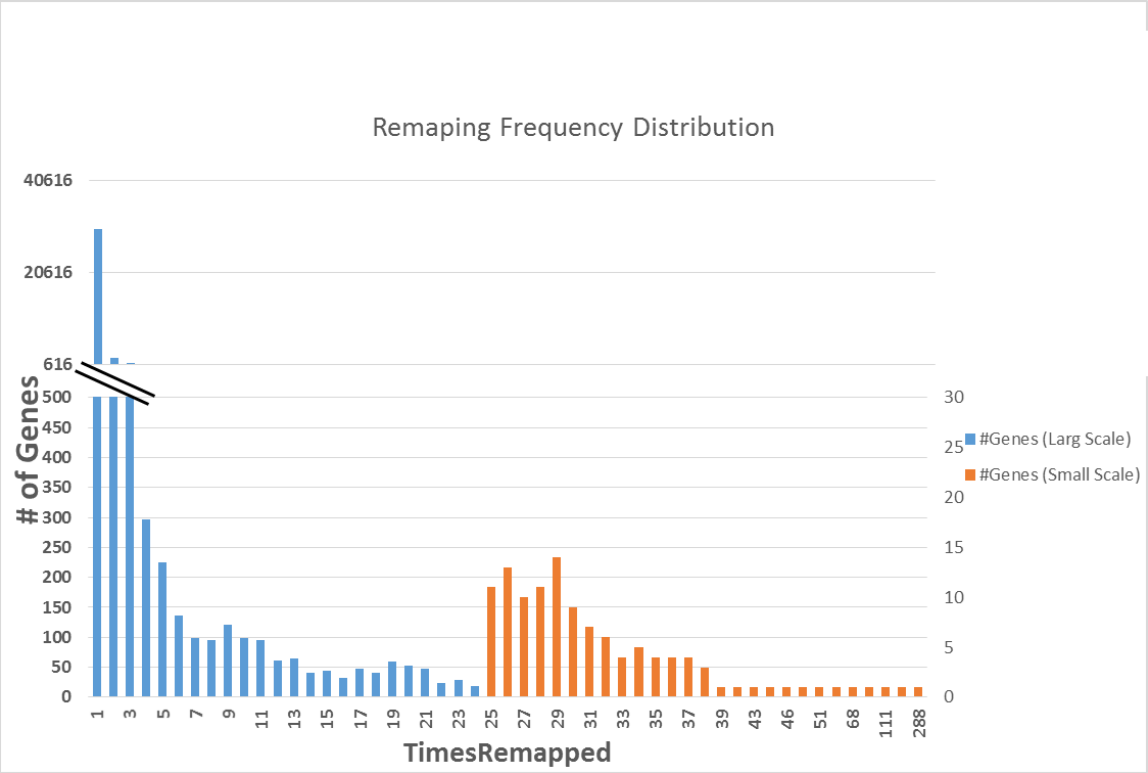


Figure 51: Remapping time distribution of ITAG 2.3 genes divided into two groups of large and small scales

As it is shown in (Figure 51 and also Table 23), 30046 genes have one positioning when mapped on the genome whereas 2088 and 661 genes have two and three times of remapping on the genome, respectively. At the extreme, the 3 genes of Solyc00g005070.1.1, Solyc12g019160.1.1 and Solyc04g047730.1.1 locate on 288 regions, 112 and 111 times, respectively. This highlights that several *iTAG 2.3* genes either have multiple copies or homologous genes not predicted on the genome, or several pieces of DNA sequence were repeated.

In principle, each gene should have only one gene mapped in the locus, overlapping to the loci it is associated to. This is a clear indication of some repeated or highly similar regions on the genome which share a big similarity proportion with several transcript sequences.

After cross comparing the remapped transcripts on their genomic coordinates considering their strands, a list of genes exactly identical in the sense of their

genomic structure (with some difference on their strands) were identified (Table 24).

Table 24: List of identical genes. Per each genes is specified: length, exons number, identical region 50 nt after and before, strand and alignment coverage (Replace with my own data for the overlapping of identical genes considering the strands)

	Sequences length (nt)	# exons	identical (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%
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%
Solyc00g188250.1.1\ Solyc06g053270.1.1	405	1	YES	plus/minus	100%
Solyc03g005530.1.1\ Solyc03g005560.1.1	1587	1	YES	plus/minus	96% (1622/1687)
Solyc07g055360.1.1\ Solyc07g055590.1.1	228	1	YES	plus/minus	98% (320/328)

We identified several independently (differently) predicted genes in *iTAG 2.3* annotation which are identical in the sequence and structure on the genome (also in some cases with differences in the strand).

Unassigned chromosome (Chromosome Zero)

The chromosome zero and its gene content are the unassigned chromosomes and genes in the tomato genome. Figure 52 illustrates the behavior of the transcripts of the genes predicted on the unassigned chromosome of the tomato which were mapped on the other chromosomes with different identities and coverages, categorized into the overlapping and not overlapping with respect to other *iTAG* predicted loci.

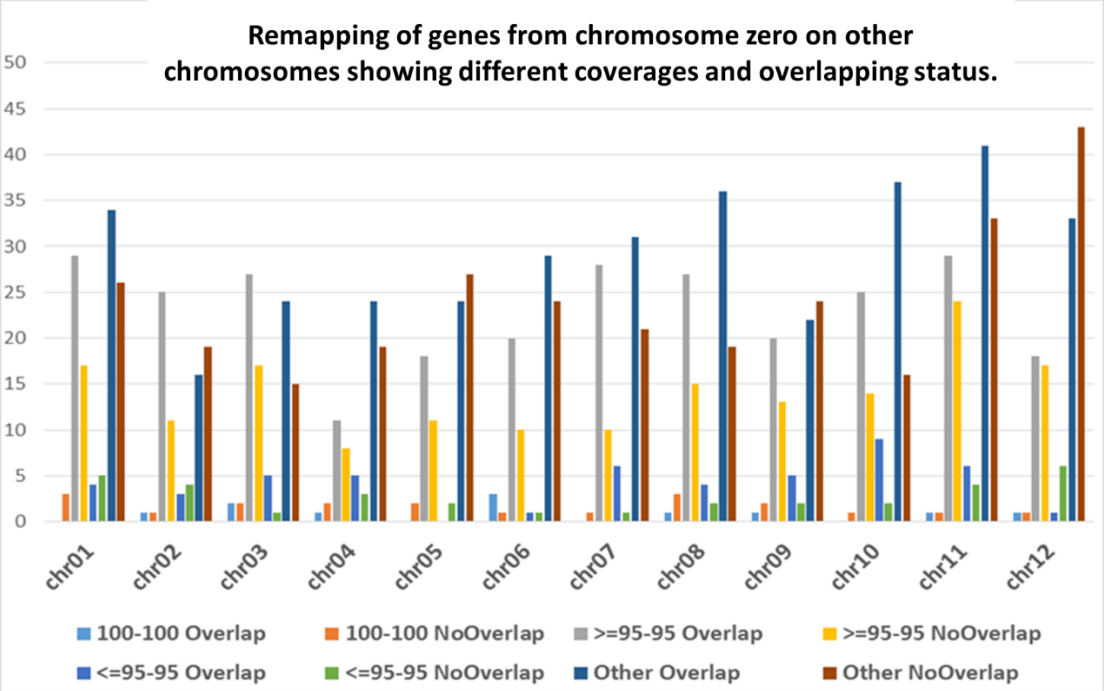


Figure 52: Representation of the remapped genes with different thresholds of identity and coverage remapped on the other chromosomes, categorized into the overlapping and not overlapping with respect to the other iTAG predicted loci

The distribution of the *iTAG 2.3* predicted genes from chromosome zero mapped in the void (un-annotated) regions of the other chromosomes with high level of identity and coverage suggests the possibility of putatively not predicted genes on those regions which were, in contrast, annotated on the unassigned chromosome. In other words, there are in total 79 genes of chromosome zero remapped on the void regions of other chromosomes with the identity and coverage >= 95 that can be putatively predicted on the genome.

Our analyses revealed a comprehensive overview of the *iTAG 2.3* genes in which, several miss-annotated genes (very long, overlapping with multiple genes, split genes, putative new genes, those predicted on the repeat region, and identical genes and genomic regions) were identified. These information are fundamental since they can introduce several biases and miss-leading issues when exploiting the genome information. As an example in the NGS data analyses, the genes overlapping multiple other genes can lead to the ambiguity in the gene expression quantification. In most methods such as *HTSeq-count* (see Quantification), the count for all these gens will be considered as zero. This is also valid for those genes with multiple mapping on the genome since the reads matching multiple locations on the genome, in most cases, will be automatically discarded from analysis. In terms of the split genes, the expression quantification will be highly affected since the complete transcript is not considered. Obviously the quantification of expression for the genes not predicted on the genome (the putative new genes we found) is not also possible unless they are put into consideration. However, in this specific case they would result to be repeated.

3.5.2.1.1 Revised annotations

Therefore we go to the conclusion that the three genes are probably miss-annotated in the annotation.

3.5.2.1.2 ITAG 2.3 Revised annotation

These version of the annotation is on the basis of the official *iTAG 2.3* annotation 2.3 in which.

The added information provided in the revised annotation is in Table 25.

Table 25: Major information segments available in the revised annotation

Label	Description	e.g.
Type of overlapping	It presents the type of overlapping from one predicted transcript versus the other one.	No overlap, partical overlap, inside, over, locus match, locus and structure match.
iTAG 2.3 overlapping loci	In case this transcript is overlapping another *iTAG* loci, the *iTAG* ID of the overlapping loci is listed.	
Code	Regarding the *iTAG 2.3* overlapping Flag, the query length, subject length and the percentage of ovrerlaping is listed.	
RefSeq overlapping loci	Incase this loci is overlapping a *RefSeq* 2.3 loci, the *RefSeq* ID of the overlapping loci is listed	
Remapped Time/code in the id	Number of times transcript maps on the *S. lycopersicum* genome with identity coverage higher than 90 and 80 respectively	
Repeats Overlap	Represents the percentage of overlap for this transcript with each of the *iTAG 2.3* repeats aggressive classes in details.	69% *LTR-RE*,12% *rRNA*

Confirmed by EST/TC/Unigenes	Represents the transcript confirmation by any of the 20 Solanaceae EST, TC collections together with the three universal unigenes collections for the tomato. The number of overlap for each collection and species, the percentage of overlap are listed in details.	
Locus Expression	The maximum level of *RNAseq* expression from the *Heinz* Atlas collection for this trascrip calculated on the basis of its gene locus is presented.	
Exon Expression	The maximum level of *RNAseq* expression from the *Heinz* Atlas collection for each gene locus is presented.	
Exons *GenBank* format	The *GenBank* format of all the exons for the transcript are listed.	(start1,end1,start2,end2,....,startn,end)

The information together with the genomic information available in the original annotation file have been made available both in the Excel and *GFF3* file formats

3.5.2.1.3 ITAG 2.3 Preferred Annotation

Considering the *iTAG 2.3* gene annotation as the reference annotation for the *S. lycopersicum* genome, a GFF3 file including the *iTAG 2.3* predicted genes together with those of *RefSeq*2.3 genes not overlapping any *iTAG 2.3* predicted

loci (1624 mRNAs) are presented to provide a more comprehensive and exhaustive gene annotation for this species. The annotation file is in the standard GFF3 format including all the functional and genomic details provided in the original files.

3.5.2.1.4 RefSeq Preferred Annotations

In contrast with the *iTAG 2.3* preferred annotation, a GFF3 file representing the *RefSeq* 2.3 predicted genes including those of *iTAG 2.3* not overlapping any *RefSeq* 2.3 gene (10931 mRNAs) is provided including all the functional and genomic information available in the source files.

3.5.2.1.5 Impact of different annotations in the NGS data analyses

Here, the expression quantification for each gene from all the 10 tissue/stages of *Heinz* atlas collection for the 4 annotations *iTAG 2.3*, *RefSeq* 2.3, *iTAG 2.3* Preferred and *RefSeq* 2.3 Preferred annotations separating once- and all-mapped reads on the genome are presented in details.

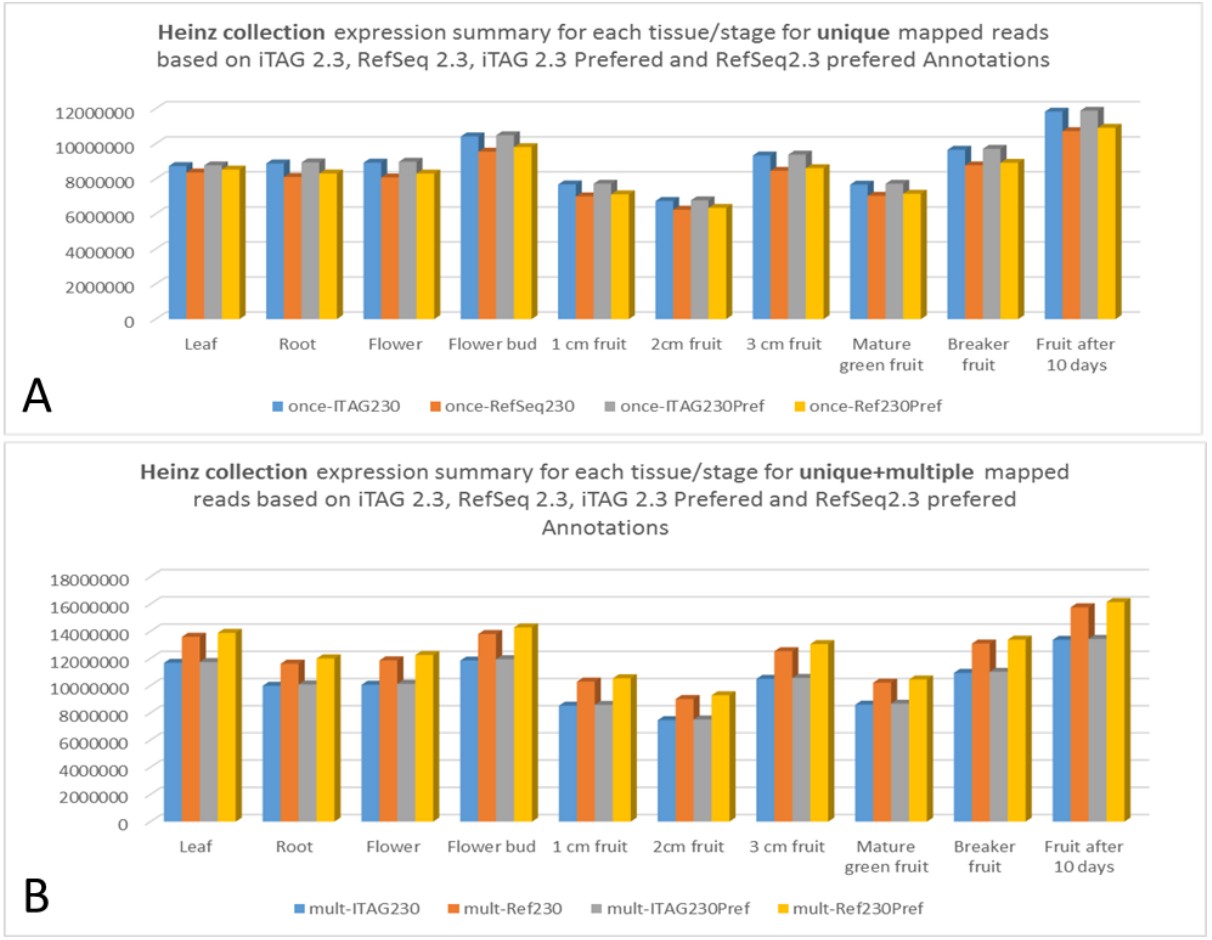


Figure 53: Representation of different read countings (A= once map and B= unique+multiple map) using iTAG 2.3, iTAG 2.3 Preferred, RefSeq 2.3, and RefSeq 2.3 preferred annotations

In Figure 53.A, the reads that uniquely mapped on the genome (once-) are checked using the four annotations of *iTAG*, RefSeq, *iTAG* Preferred and RefSeq preferred. Indeed we report the number of reads mapped in the gene loci. In Figure 53.B, multiple read mapped on the genome and once mapped were considered for the quantification.

3.5.3 NGS Data Analyses

During my PhD, several NGS data analyses were conducted, in collaboration with other research groups or independently for some of my target based analyses, on various collections. Here, some example of findings on major collections useful to highlight features of the *iTAG* 2.3 annotations are reported. The *Heinz* collection, as a representative collection of *RNAseq* from several tomato tissues from the sequenced genome, was used for several investigations and analyses to expand our knowledge on the reference genome defined for this crop species. Here we present some of the results and findings acquired.

Zero level Genes

The data analysis highlighted that among all 34727 *iTAG* annotated genes, 6412 genes showed zero read mapped on the gene loci when considering any of the libraries from each of the replicates from the *Heinz* collection. Interesting to observe that the number of genes that are zero are also 5700 in the paper of tomato genome release [137]. This is probably due to analytical approaches and highlights the importance of clear description of the methods used to reproduce the data. Indeed, overlapping genes are counted as zero from many of the methodologies [68, 69]. Moreover, besides the 0 counting genes, 10025 genes showed expression levels lower than 1 RPKM in all the tissues/stages, falling in the criteria to be defined as not expressed genes. Hence, in total, 24702 genes show expression level higher than 1 RPKM in at least one of the investigated physiological conditions.

Genes specifically expressed in a tissue

We also analyzed the number of genes that showed the expression level higher than specified thresholds (0.3 and 1 RPKM, respectively) only in one condition (defined as specifically expressed). On the other hand, we also reported the list of genes with expression level lower than the given thresholds only in a specific condition (defined as specifically not expressed). In Figure 54, we report the number of specifically expressed and specifically not expressed genes per conditions according to the different thresholds. The statistics shows that comparatively a large number of genes (1106) are specifically expressed in root while a significant number of genes (695) are tissue specifically not expressed in "fruit after 10 days".

	Specifically expressed		Specifically not expressed	
Tissue\Threshold	**0.3**	**1**	**0.3**	**1**
Leaf	99	138	158	158
Root	836	1106	123	160
Flower	120	105	44	46
Flower bud	624	630	45	33
1cm fruit	57	65	48	62
2cm fruit	72	53	18	13
3cm fruit	55	48	38	37
Mature green fruit	51	51	28	8
Breaker fruit	26	19	179	144
Fruit at 10 days	71	64	756	695
All Fruit	9	21	190	235

Figure 54: Tissue specifically expressed and not expressed status for the Heinz NGS collection

We also observed that 21 genes are specifically expressed in the fruit stages while 235 genes are specifically not expressed in the same conditions. The results suggest that probably these genes, not expressed in other tissues but only

fruit) are the fruit specific expressed genes, which can have important roles in the fruit maturation and the processes involved.

3.5.4 Transcriptome analyses for the Heat Stress Response in Tomato Pollen

Gene based MACE data analysis

MACE NGS data provided by GeneXpro were analyzed by a classical gene reference based approach.

The collection was also a precious resource that was made available in the *SPOT-ITN* project since *RNAseq* from the developmental stages of pollen were not available for tomato in the SRA archive (ref). As an example, the *Heinz* collection does not include stages from pollen. To this end, we tried the exploit the data to better understand the peculiarities of this collection in terms of expression profiling.

3.5.4.1.1 Putative pollen specific genes detection in tomato

We crosschecked the information from *Heinz* tissue specific genes with those from *MACE* data analyses.

Table 26: Representation of expression signaling between the Heinz and MACE NGS collections

Total Gene	Heinz	MACE (ALL)
34727	6412	1051
		5361
	26561	662
		25899

No Expression Signal

With Expression Signal

The analyses revealed that among the 6412 genes not expressed in any of the *Heinz* tissues/ stages, 5361 genes showed expression signaling at least in one of the tissues of pollen from our *MACE* collection (control and heat stress conditions). The results suggest the putative pollen specific genes in *S. lycopersicum* which are expressed in pollen stages only and not in any other tissue.

Table 27: Representation of specifically expression of MACE NGS collection between Control and Heat Shock Stress stages

Heinz	MACE(CT)	Different	Common	Different	MACE(HS)
6412	1599	548	1051	290	1341
	4813	**292**	4521	**550**	5071
26561	1465	803	662	478	1140
	25096	478	24618	803	25421

No Expression Signal

With Expression Signal

Comparison of the expressed and not expressed genes between the putative pollen specific genes detected at the previous step also revealed that, 292 of these putative pollen specific genes are specific in the physiological condition while 550 of these putative pollen specific genes are expressed only when the pollen undergo the heat shock.

Due to these evidence, an integrative approach was undertaken to decipher the transcriptome changes during tomato pollen developmental stages and under stress response. As described in the materials and method (see 2.4) and due to the issues observed in the tomato gene annotation, we exploited an "annotation free" approach using *Tracker* to investigate the transcription changes

DEG Analyses

We localized Hot Regions as those identified by the software *Tracker* and resulting as differentially expressed regions when comparing the heat shocked and the physiological conditions of each pollen developmental stage in the *MACE* collection in terms of number of regions showing up- or down-regulation signals. These data were also intersected with the *iTAG 2.3* annotated loci.

Table 28: MACE Annotation free analyses DEG loci detected with their overlapping status with the iTAG annotated loci

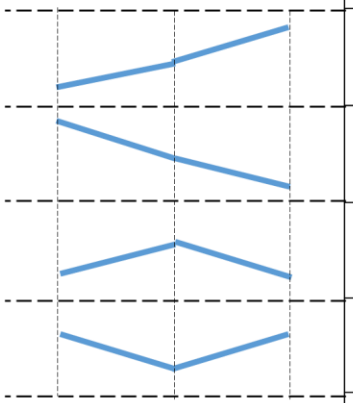
Comparisons (Heat Shock vs. Control)	DEG Loci	Status	Associated to a gene loci	HS Related
Tetrad	15	**UP (10) DW (5)**	6	2
Post-meiotic	71	**UP (57) DW (14)**	32	4
Mature	159	**UP (112) DW (47)**	88	7

Characterize be … change into Associated to gene loci …

As presented (Table 28), 15 (10 up and 5 down), 71 (57 up and 14 down) and 159 (112 up and 47 down) Hot Regions were detected in the pairwise comparison of Tetrad, Post-meiotic and Mature stages in the tomato pollen, comparing the heat shock versus physiological condition. Among all only 126 (6+32+88) regions were overlapping an annotated locus in which 13 (2+4+7) of them are heat shock related genes.

Interestingly, 119 regions (genes or isoforms, see MACE Data) were fallen out of the gene regions in which no functional assignment could be assigned to them from the *iTAG 2.3* official annotation.

We also categorized the differentially expressed/suppressed regions into 4 major trends of (up-up, down-down, up-down, and down up) across the developmental stages (Figure 55).



Trend (TotalDEG)	CT => CPM => CM (5898)		HST => HSPM => HSM (5337)
	All-iTAG-Revised	Common	All-iTAG-Revised
UP-UP	67	9	61
DW-DW	141	2	179
Up-DW	20	0	15
DW-UP	88	15	44
Other	5582		5038

Figure 55: Statistics regarding the MACE detected DEGs with respect to their up- or down-regulation trend

Figure 55 lists the differentially expressed/suppressed regions annotated with the *iTAG 2.3* revised annotation (see 2.4.1) in which, different number of genes showing common and different trends in the sense of expression across the physiological and heat shock conditions are demonstrated.

We also considered the same approach to detect hot methylated or under methylated regions from *MethSeq* data provided by GenGPro Company (Frankfurt, Germany).

The analysis resulted in a strong de-methylated regions when comparing heat stages versus physiological ones. Interestingly, the largest number of regions affected by the phenomena are at the first stage of pollen development indicating that the stress caused a drastic change of methylation status (usually repressing expression) to de-methylated ones (Figure 56.B).

Comparing the expressed regions with those de-methylated (Figure 56.A and B)

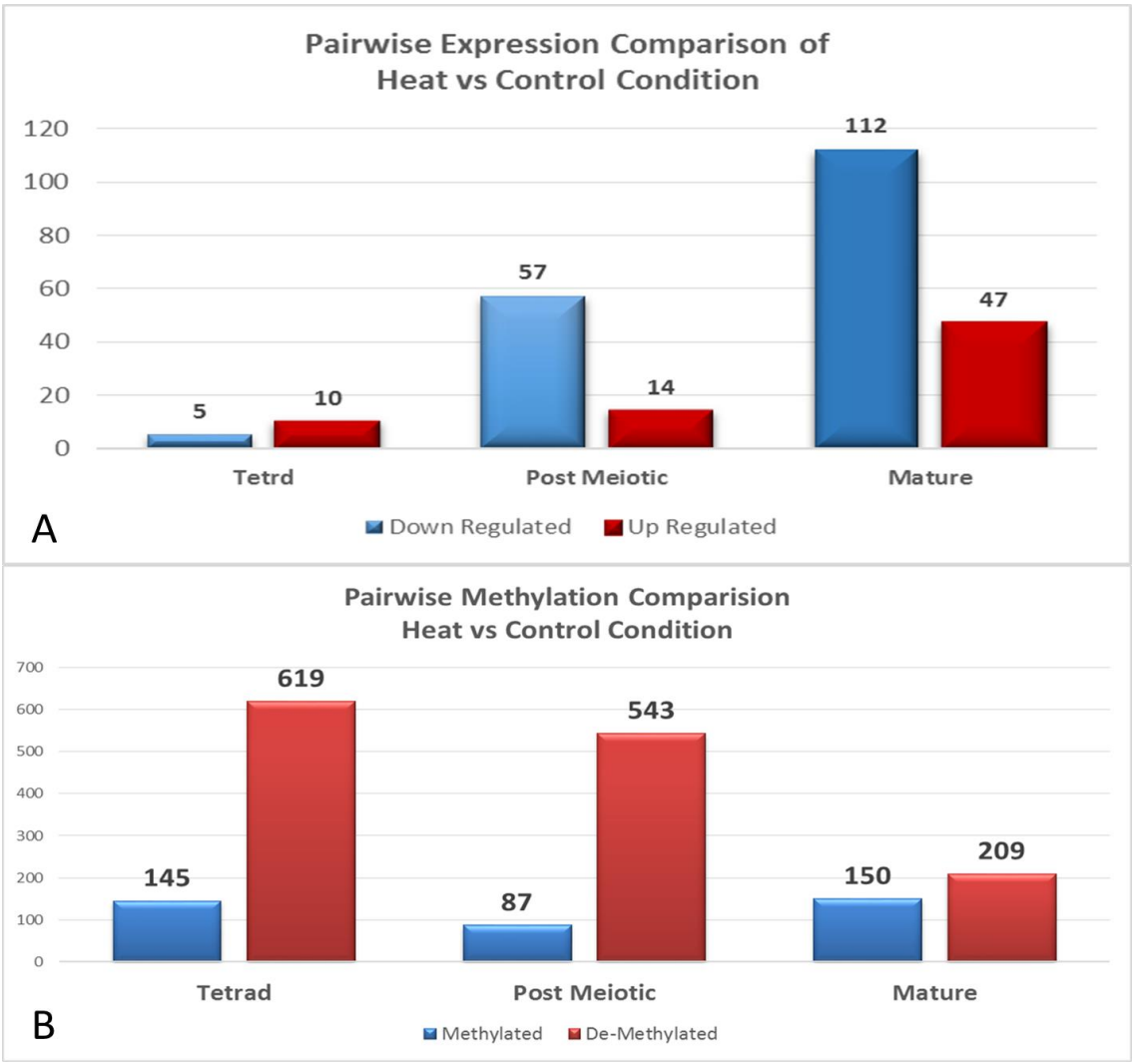


Figure 56: A) Genes differentially expressed and suppressed, and B) Methylated and de-methylated CCGG sites (Opening of chromatin) in response to heat shock during the developmental stages of Tetrad, Post-Meiotic and Mature Pollen.

The general trend of expression and methylation, comparing the heat shock and physiological conditions for each pollen stage, across the developmental processes can be overviewed. In terms of the expression (Figure 56.A), an increase of down-regulation and up-regulation of the genes across the developmental stages is observed. However, except the tetrad stage, a higher down-regulation comparing to the up-regulation of the genes in the same stage for the other two stages is presented. Interestingly, the de-methylation event

preceded and accompanied the expression in the corresponding stages with a significant difference comparing to the methylation process (619 vs 145 in Tetrad, 543 vs 87 in Post-meiotic, and 209 vs 150 in Mature). Although the de-methylation is decreasing in the pair wise comparison of each developmental stage, the de-methylated regions comparing to the methylated regions along the pollen development under heat shock is increasing significantly. This can be the result of tomato plant response to the stress during the development. A chromosomal distribution of this process for each step during the pollen development is presented in figure below

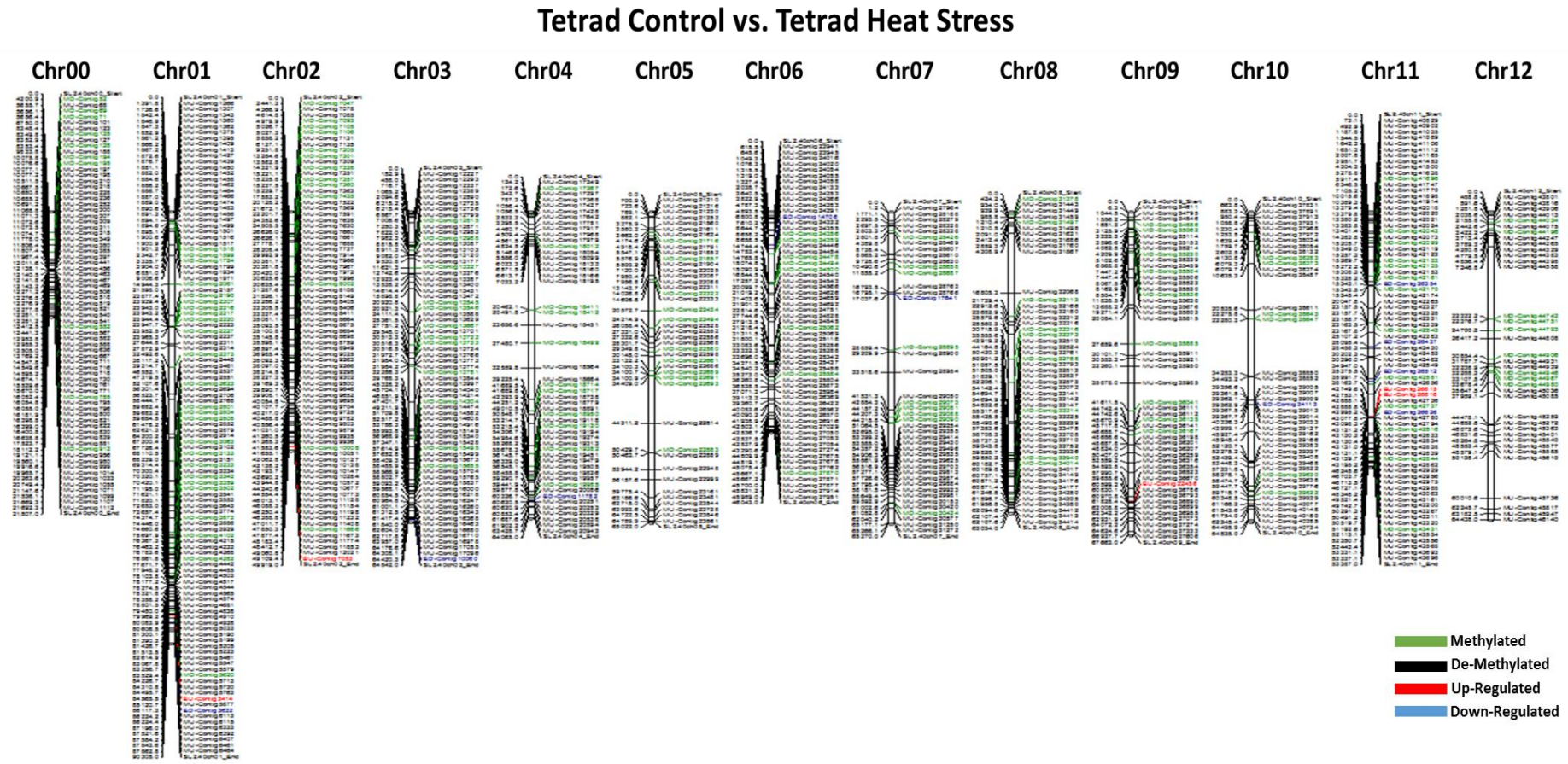
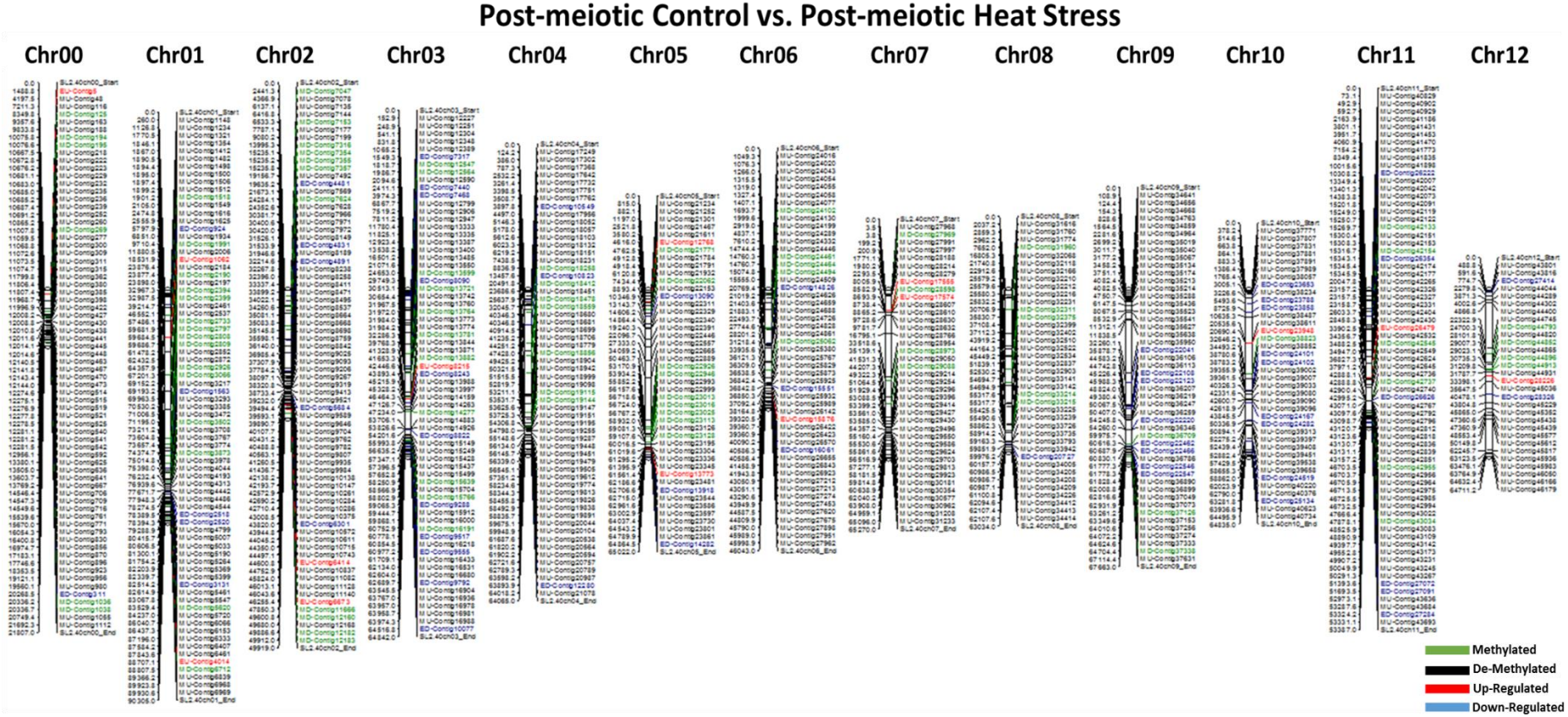
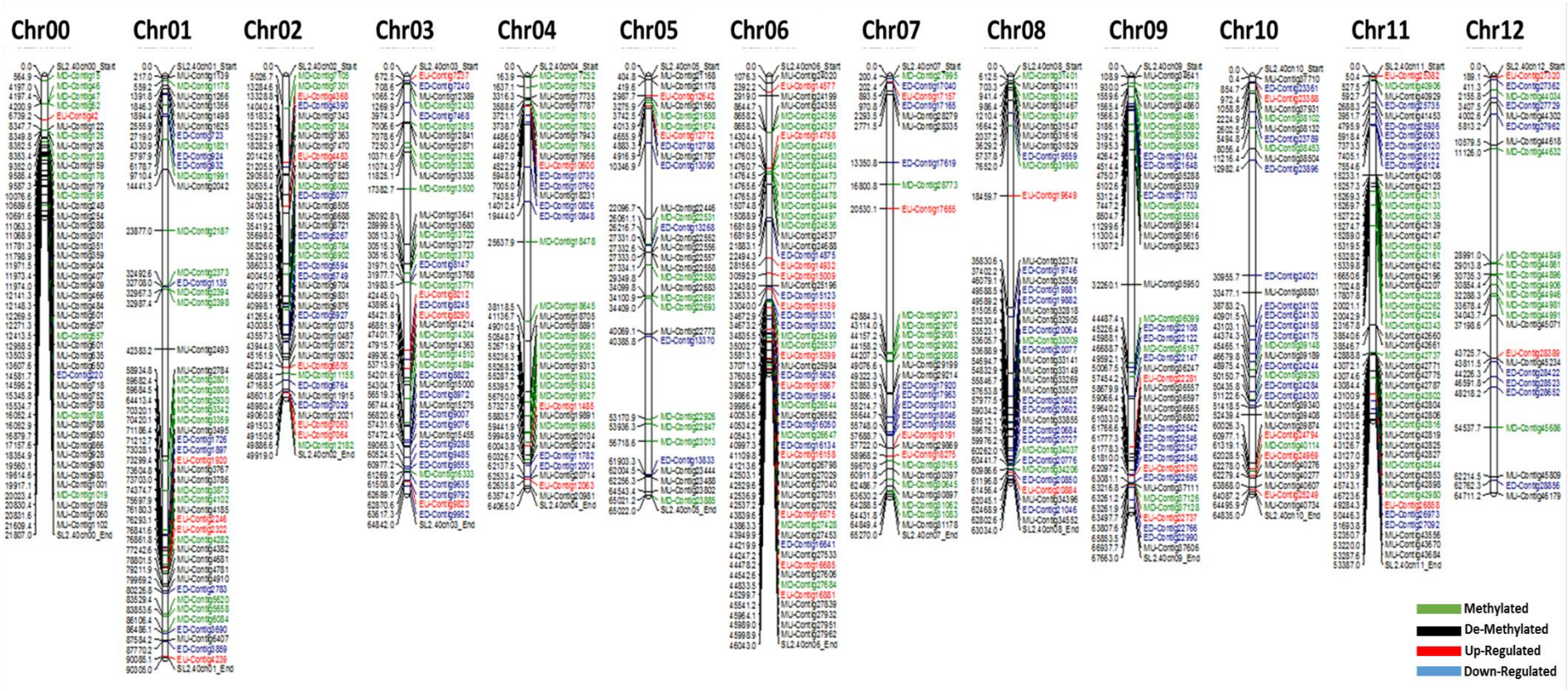


Figure 57: A) Changes of Methylation and expression in tomato genome and genes during the post-meiotic pollen developmental stages in response to the heat shock stress


Post-meiotic Control vs. Post-meiotic Heat Stress
Chr00
Chr01
Chr02
Chr03
Chr04
Chr05
Chr06
Chr07
Chr08
Chr09
Chr10
Chr11
Chr12
Methylated
De-Methylated
Up-Regulated
Down-Regulated

B


Mature Control vs. Mature Heat Stress
Chr00
Chr01
Chr02
Chr03
Chr04
Chr05
Chr06
Chr07
Chr08
Chr09
Chr10
Chr11
Chr12
Methylated
De-Methylated
Up-Regulated
Down-Regulated

C.

Figure 57.A shows the differentially expressed/suppressed and methylated/de-methylated sites in the Tetrad stage comparing the heat shock versus the control. A large number of de-methylation together with some genes up-regulation are observed. Figure 57.B is the demonstration of methylation and expression status of the genome and genes when comparing the heat shock versus physiological condition in the Post-meiotic stage. It is shown that the de-methylation of the genome is evident. Also the up- and down regulation of the genes are increasing comparing the previous stage (Tetrad). Looking at the mature stage of pollen comparing the heat shock versus the physiological condition (Figure 57.C), it is observed that the de-methylation of the genome is still ongoing but to a less extent comparing to the previous stages of the pollen (Tetrad and Post-meiotic). The GO Enrichment analyses for the up- and down-regulated genes detected in our approach considering all the developmental stages also suggests biological processes such as *response to the endoplasmic reticulum stress*, *regulation of pH*, *methylation*, *cell wall modification* and *mitochondrial organization*, and *developmental vegetative growth* are enriched during the pollen developmental under heat shock.

Figure 58: Representation of Gene Set Enrichment Analyses for the genes Differentially Expressed or Suppressed in HSR during the developmental stages of Tetrad, Post-Meiotic and Mature Pollen.

These results can provide a comprehensive overview of the phenomena implied in tomato pollen development under heat stress, helpful to underlie the mechanism involved and the general biological process. On the other hand, due to the lack of a complete annotation for tomato, a large number of Hot Regions are not functionally identified. This requires further experimental and bioinformatics minings to characterize the differentially expressed/suppressed sites (hot regions) where they are not fallen in a predicted gene locus.

3.5.5 The Role of TE-Derived Small Interfering RNAs in Tomato Pollen Development

A genome wide analyses on the small-RNA collection (see 2.2.1) provided by *GenXPro* (Frankfurt, Germany) with respect to their interaction with the TE elements was conducted. Specifically, since novelties from miRNAs were already provided by *GenXPro* (Frankfurt, germany), we focused on a different aspect.

Out of 74,469,092 sequencing reads generated after removing low quality reads, we mapped 94.3% of the reads (which were ranging from 11-38nt. We obtained the percentage of reads mapped per each size class over the total mapped reads for the respected condition. We calculated the fraction of reads per each library for each size class (Figure 59).

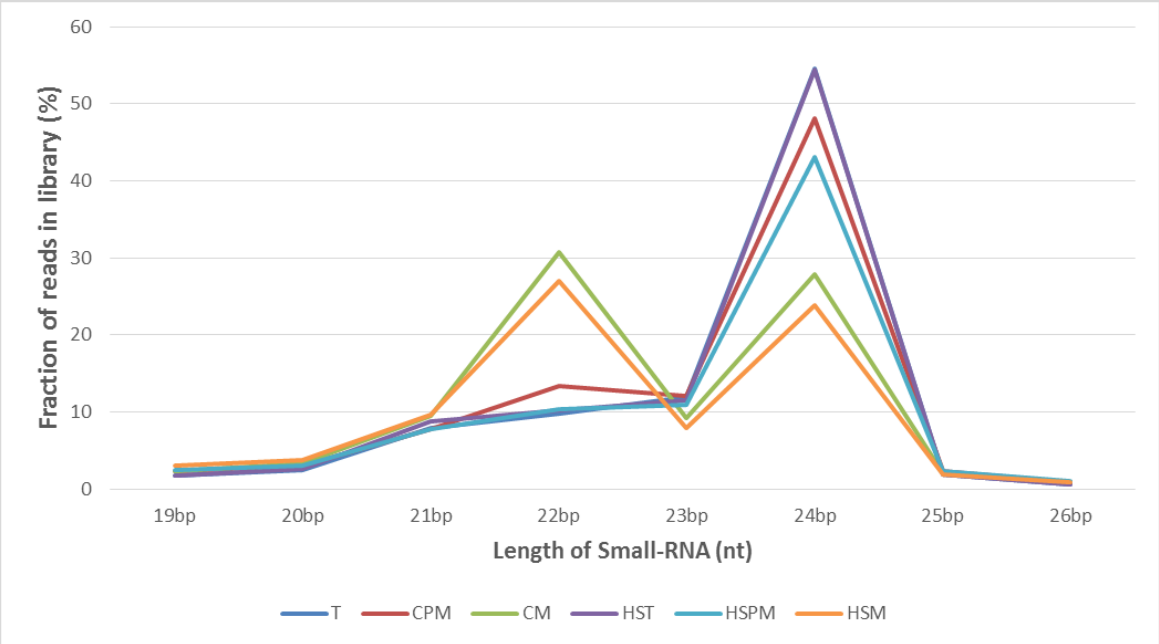


Figure 59: Fraction of reads in library for each of size class per each tissue/stage in pollen developmental stages

As it is demonstrated in the Figure 59, no significant changes are observed between the similar stages under control and heat stress. In contrast, a significant change is observed in the 21.22 and 24 nt classes between the developmental stages.

Notably, we found that the abundance of 21-22nt class and 24nt classes were significantly switched during pollen development (Figure 60).
Basing our analyses on the 2 classes of 21, 22 and 24 nt *RNA fragments*, we considered the type of mapping into unique and multiple mapped on the genome.

In the tetrads and post-meiotic stages, the 24nt *Small-RNA* predominated the *Small-RNA* reads (52% and 46%); however, at mature stage (binucleate pollen), the 21-22nt *Small-RNA*s became dominant (38%), while the abundance of 24nt *Small-RNA*s was drastically reduced (to 27%). Within the 21-22nt class, the relative abundance of 22nt *Small-RNAs* over 21nt *Small-RNAs* was significantly increased from PM stage (63% $\pm$2.5%) to M stage (75%$\pm$1.2%) (P=0.010).

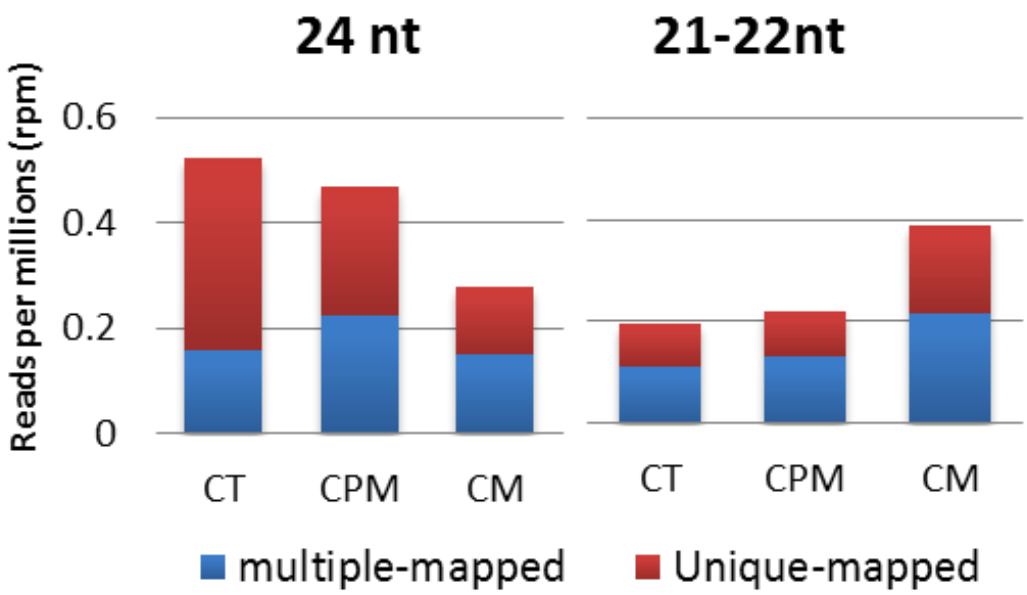


Figure 60: Fraction of reads for each 21, 22 and 24 nt size class of Small-RNA normalized by reads per million (RPM) normalized for each of the stages in Pollen.

Among *Small-RNA* reads that mapped to repeat region of tomato genome, both uniquely mapped and multiple mapped reads were significantly altered from PM to M stages ($P<0.05$) (Figure 60).

Table 29: Number of Small-RNA clusters generated for the 21, 22 and 24 nt classes and their repeats overlapping status

	21-22nt	**24nt**
Total Cluster generated	29638	85755
On repeats	24882 (83.9%)	71588 (83.5%)
On gene-coding region	762 (2.5%)	1716 (2.0%)

As presented in Table 29, a total cluster of 29638 and 85755 *Small-RNAs* for the 21, 22 and 24 nt classes were generated respectively (see 2.6.3). Over all the clusters, 83.9% for the 21, 22 nt and 83.5% for the 24 nt classes were located on the repeated regions. The clusters generated for the 21-22nt were ranging in the length from 25bp to 9063bp (Median: 971bp) while for the 24 nt were ranging from 100bp to 31259bp (Median: 2658bp).

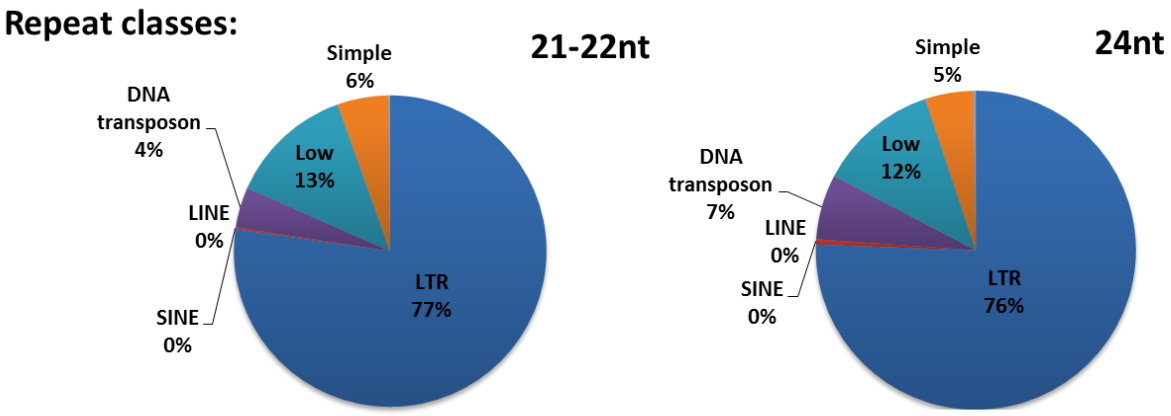


Figure 61: The intersection of the 21, 22nt and 24 nt Small-RNA classes with the iTAG 2.3 repeats aggressive.

As presented in Figure 61, the majority of the *Small-RNAs* are located in the LTR repeated regions (over 75%) while the other repeat classes such as low and simple overlap over 12 and 6 percent of the clusters respectively. During pollen development, we identified 4,488 and 55,458 of 21-22nt and 24nt differential expressed *Small-RNA* clusters (DEC= Differentially Expressed Clusters) on repeat region respectively. Specifically, 596 (13.3%) of 21-22nt clusters and 284 (0.5%) of 24nt clusters were significantly altered from tetrad to post-meiotic stage; 3,892 (86.7%) of 21-22nt and 55,174 (99.5%) of 24nt *Small-RNA* clusters were significantly altered from post-meiotic to mature stages. We further annotated the DECs to repeat regions on the tomato genome, including both class I retrotransposons (LTR, LINE, and SINE) and class II DNA

transposons (Figure 61). The majority of DECs (80-85%) are located at the Long-terminal-repeat (LTR) retrotransposons.

Unlike miRNAs, which function mainly in post transcriptional gene silencing (PTGS) mechanism, there is a class of *Small-RNA*s (siRNAs) that can function in both PTGS and transcriptional gene silencing (TGS). In PTGS, siRNAs target transcripts specifically by sequence complementary, similar to the action of miRNAs; while in TGS, siRNAs rather mediate DNA and histone modification events to surrounding genome regions, thereby influencing transcription ability [161]. To determine whether the 21-22nt and 24nt *Small-RNA*s could be involved in the same mechanism during tomato pollen development, we associated the 21-22nt and 24nt differential expressed clusters (DECs) to their mapping patterns on the genome respectively.

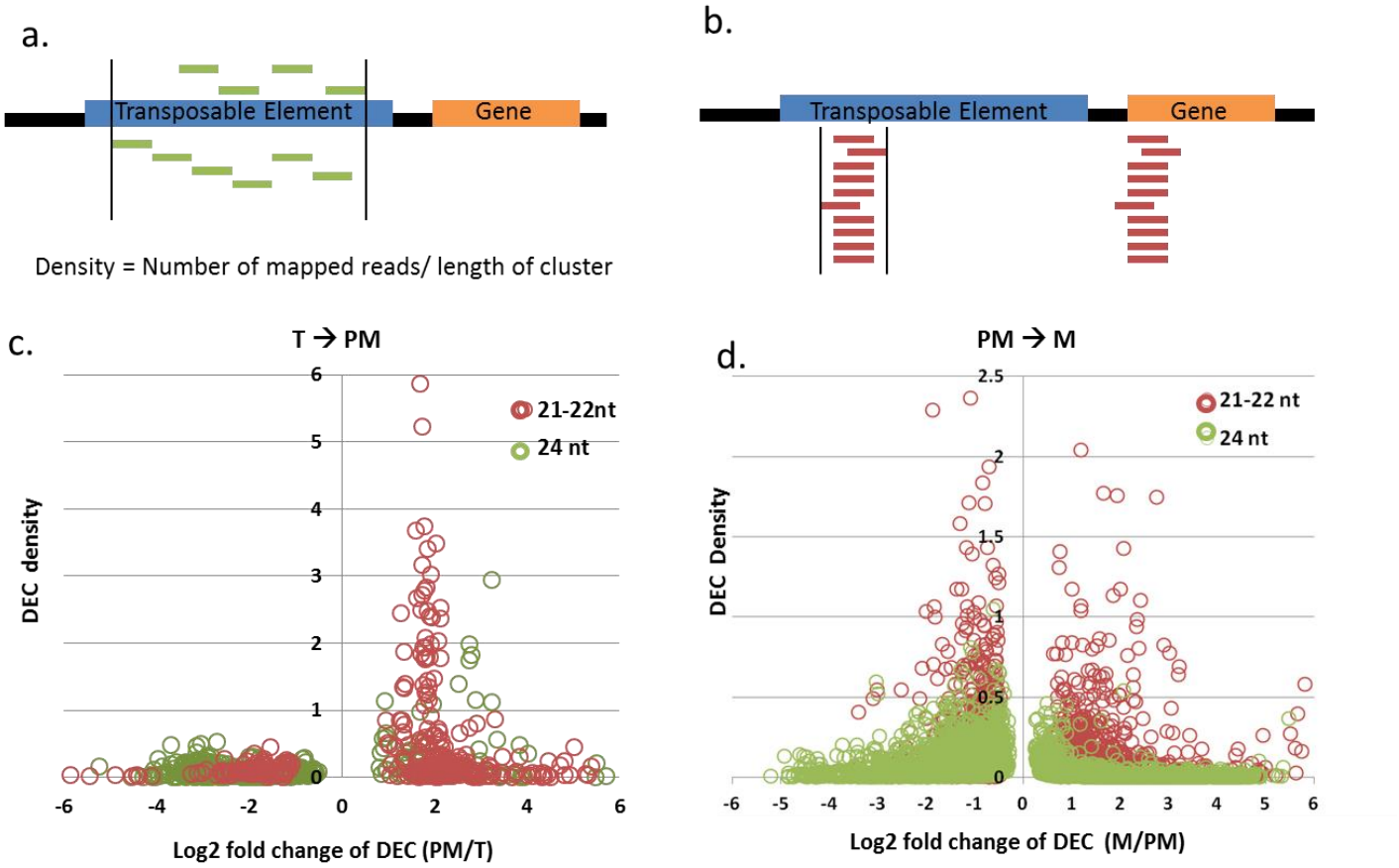


Figure 62: a) and b) are schematic representation of two types of mapping patterns in Small-RNA libraries. c) Dot plot showing the differential mapping patterns of 21-22nt and 24nt siRNAs at DECs. X-axis: siRNA expression changes, represented as log2 fold change between development stages. Y-axis: Density of according DECs, where density=Number of mapped reads / length of clusters.

We were able to identify differential mapping patterns for 21-22nt DECs and 24nt DECs at both developmental stage transitions: from tetrads to post-meiotic and from post-meiotic to mature. The 21-22nt *Small-RNA*s were mostly densely mapped to specific loci with high abundance, likely involved in PTGS; while 24nt *Small-RNA*s mapped with low density, but covering broader genome regions, are likely involved in the TGS mechanism.

CpG methylation was not affected by siRNA alterations in repeats

We next investigated if the differential expression of siRNAs altered the DNA methylation status at surrounding loci. It has been reported that during plant gametogenesis both asymmetric CHH methylation and symmetric CG methylation went through reprogramming, preferentially in sperm nucleus and vegetative nucleus respectively [67, 162, 163]. Using methylation-sensitive restriction enzyme-assisted DNA methylation deep sequencing (Meth-Seq), we were able to detect genome-wide CG methylation during pollen development. We found that in tomato pollen, even though TE-derived siRNAs (both 21-22nt and 24nt) were differentially expressed at both stage transitions (from T to PM, and from PM to M), the CG methylation level was not affected (Figure 63).

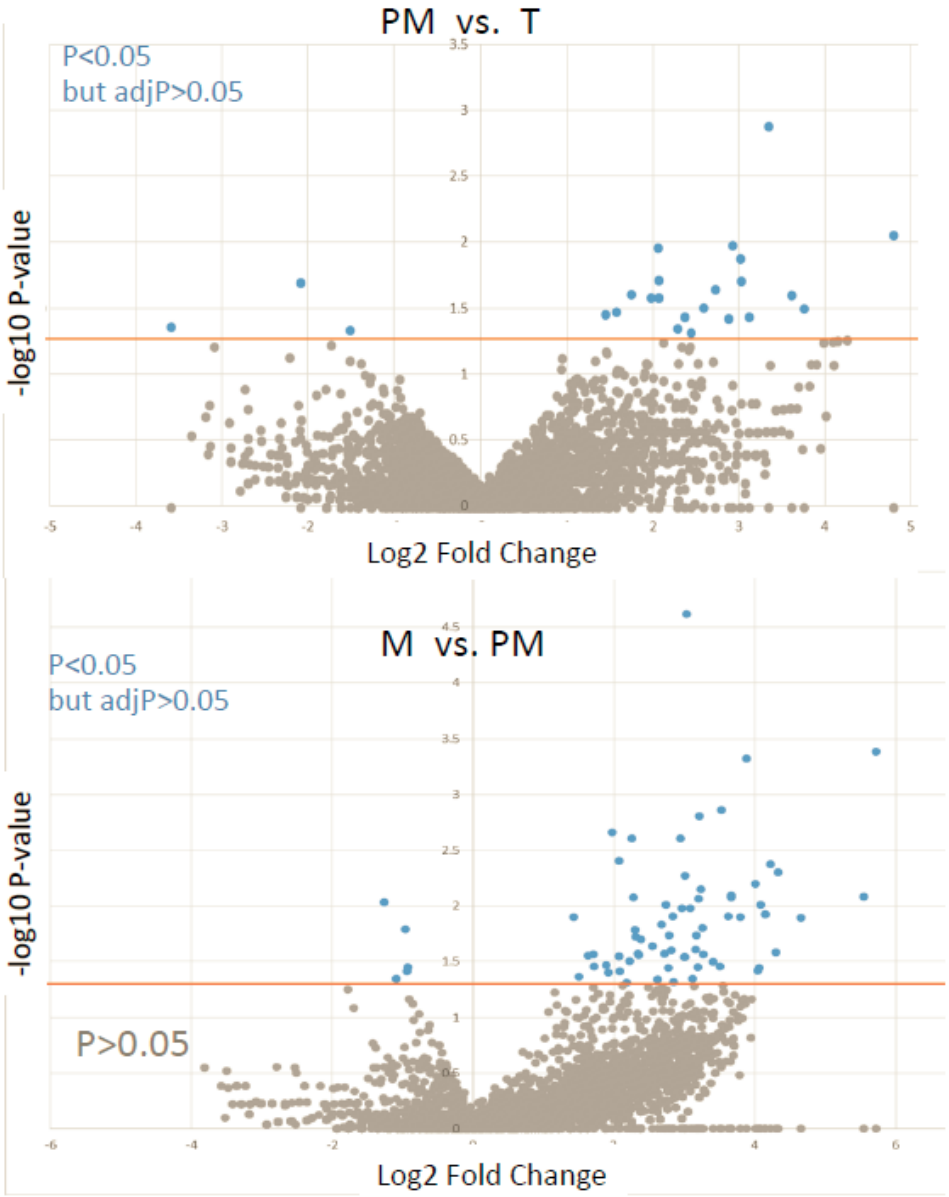


Figure 63: Methylation changes in the developmental stages of pollen in transition from Tetrad to post-meiotic and Post-meiotic to Mature stages.

TE-derived siRNA affected the expression of adjacent genes

The siRNAs have been shown to affect TE activity, which can further influence transcription ability of neighboring genes [164]. Therefore, we further investigated if alteration in siRNAs expression (or targeting) affected nearby

gene expression in tomato pollen development. To assess the relationship between siRNA abundance and nearby gene expression, we measured the distance from any TE-mapped, differential expressed siRNA clusters (DECs) (either 21-22nt or 24nt) to its nearest neighboring gene, including 5kb upstream of TSS (Transcription Start Site) and 5kb downstream of TTS (Transcriptional Termination Site). The effect of siRNA alteration on gene expression was further separated as to whether the siRNA clusters were up-regulated or down-regulated during development (from T to PM, and from PM to M). The genome-wide gene expression data were generated from Massive Analysis of c*DNA* Ends (*MACE*).

Overall, we identified 310 and 1,005 genes potentially affected by 21-22nt siRNA targeting or 24nt siRNA targeting respectively. We found that, for the 21-22nt class, gene expression level was most strongly influenced by DECs located close to the TSS (Figure 64.a). The most proximally located genes (to the up-regulated DECs) showed an average 4 fold up-regulation in gene expression (log2 fold change of 2); as the distance increased to 2kb from DECs, the influence on gene expression became trivial (log2 fold change of 0); when the distance was about 5kb apart, gene expression become negatively correlated to the siRNA changes. Similarly, genes proximal to the down-regulated DECs were averagely down-regulated (log2 fold change of -2 to 0). However, for the 24nt class siRNAs, the effect of DEC proximity on gene expression is not detectable (Figure 64.b). These results showed a differential influence of 21-22nt or 24nt siRNAs (targeting to TEs) on their nearby gene expression, possibly indicating their differential involvements in cellular mechanisms and functions.

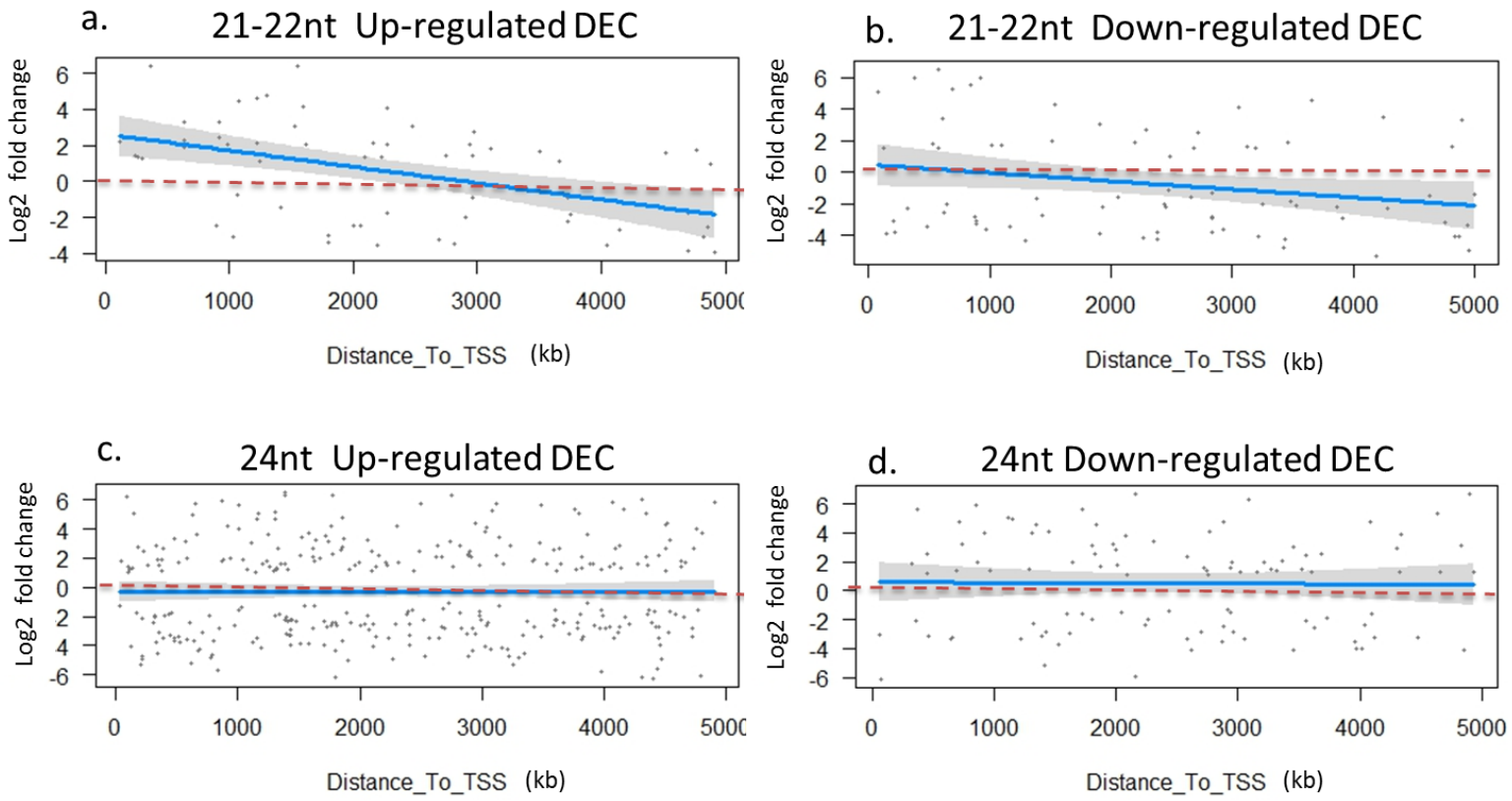


Figure 64: DEC and neighboring gene expression during pollen development. Linear regression models showing the relationship between the distance (kilobase) of DEC to TSS and the change of the corresponding gene expression (log2 fold change) for 21-22nt siRNAs (a and b), and 24nt siRNAs (c and d). Grey area: 95% confidence interval (CI) for the linear fit. Red dotted line: log2 fold change=0.

We further annotated those genes that are differentially expressed by siRNA targeting. GO enrichment analyses revealed that genes involved in metabolic and biosynthetic processes, actin filament bundle and nucleosome assembly, cell cycle and embryonic development, as well as cellular defense functions were significantly enriched (Figure 65).

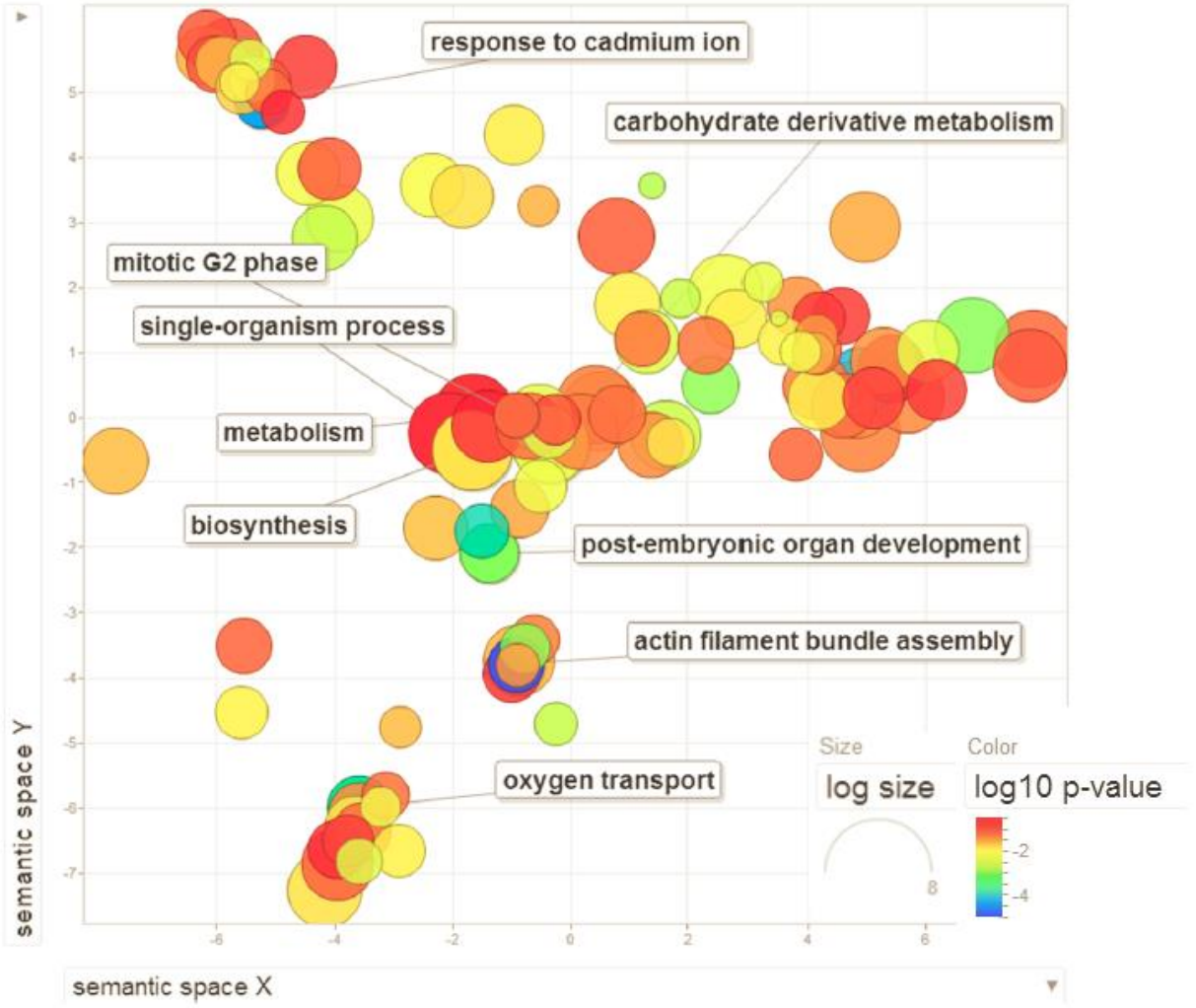


Figure 65: The GO Enrichment analyses for the genes adjacent to the SiRNA targeting on the genome affected in the expression level.

The spatial distance between the Go terms indicates the similarity and dissimilarity of the GO terms, while the ratio of the size for each circle is related

to the GO frequency in the subset. The color indicates the enrichment adjusted p-value for the fisher exact test.

siRNAs Pathways were developmentally regulated

Cascades of genes are involved in both the production and the targeting of 21-22nt and 24nt siRNAs specifically in plants [165]. To determine how these genes are regulated during tomato pollen development, we performed pathway analyses using gene expression data generated from *MACE*. In the 24nt siRNA pathway, genes involved in siRNA processing and targeting (e.g. *HEN1*, *AGO4/6/9*), as well as RdDM (e.g. *SUVH2/9*, *DMS3*) were significantly up-regulated at the post-meiotic and mature stages; genes involved in chromatin modification (e.g. *MET1*, *DDM1*, *LDL1/2*) were highly expression throughout development stages (Figure 66).

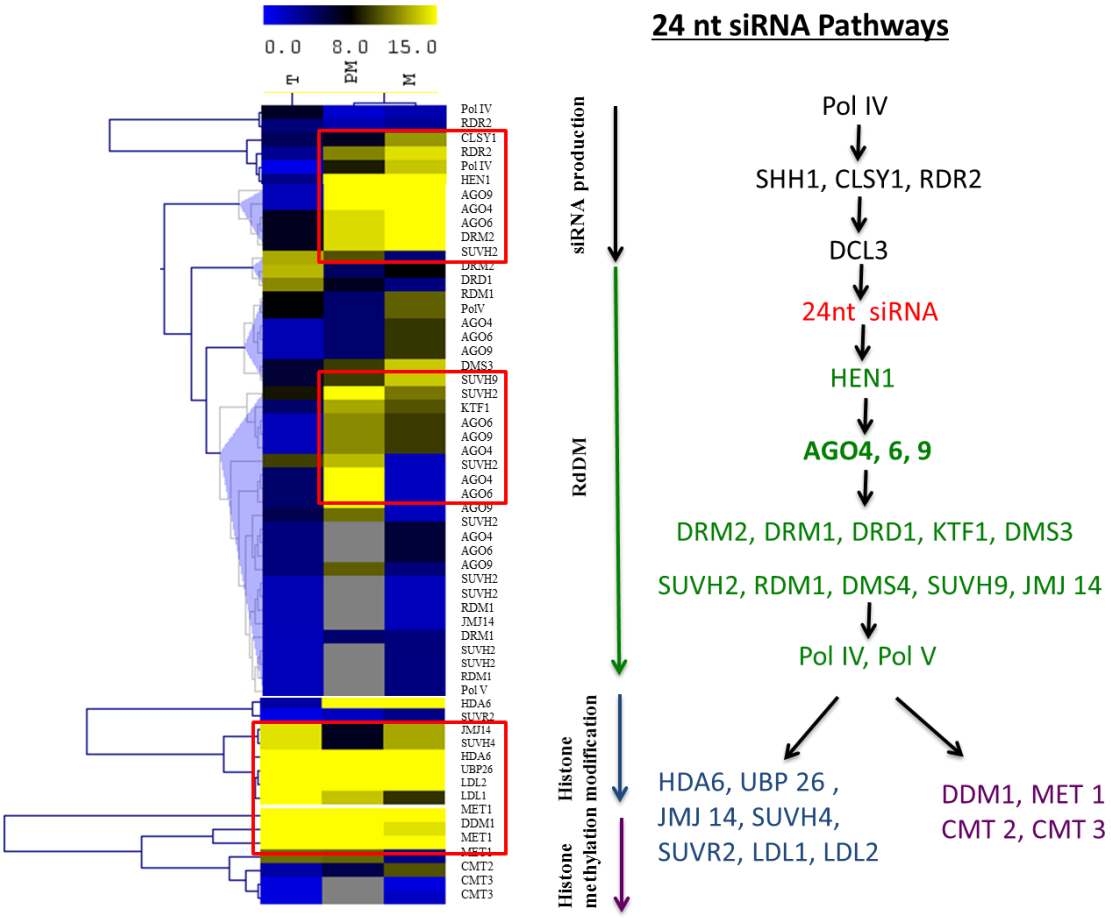


Figure 66: The expression heatmap and clustering of the genes involved in the 24 nt siRNA pathway.

In the 21-22nt siRNA pathway, there was a prominent elevation in the siRNA production genes (e.g. *DCL2/4*) at the PM and M stages, correlating with the significant up-regulation of 21-22nt siRNA abundance (Figure 67). Interestingly, an important gene *NERD* involved in non-canonical RdDM was also up-regulated in the mature pollen (Figure 67).

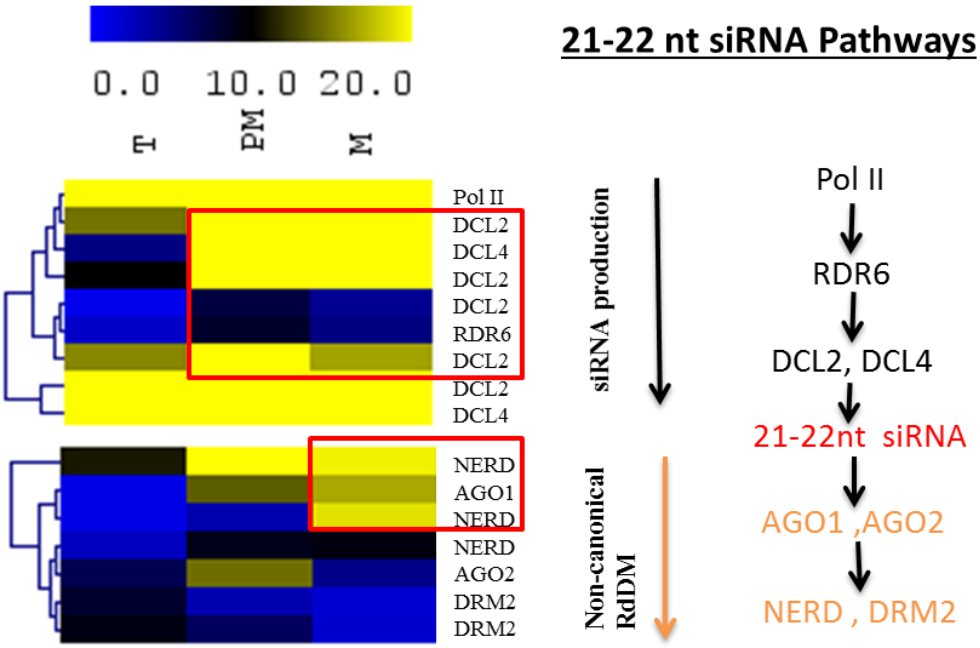


Figure 67: The expression heatmap and clustering of the genes involved in the 21-22 nt siRNA pathway

3.6 Summary

As presented in this chapter, the materials used, and the methods to collect, process, reconcile and analyze them was presented in details. In terms of the bioinformatics platforms developed, the architecture used and the data sources included in each partition was also presented.

4 Conclusions

In the light of the SPOT-ITN project objectives, and to provide a comprehensive bioinformatics infrastructure to support extensive genomics analyses in tomato, we collected, processed and integrated different resources; and organized them into dedicated databases with appropriate query user interfaces. In this thesis, the main efforts undertaken and the analyses conducted on the basis of such resources with the strategies and approaches developed are reported in details.

Deeper investigation on the two available reference genome sequences of the tomato revealed that the newly released genome (SL2.50) is the reorganization of the previous version (SL2.40) based exclusively on added gaps (in the form of "N" insertion) and some genomic sequence's inversions. In other words, both of the genome sequences are the same in the sense of their genomic content which was not immediately derived from the presented paper associated to the second release ref.

Comparing the two available *iTAG* gene annotations for tomato we also revealed that except two genes that were removed in the newest version (*iTAG*2.4) , all the other genes were transferred to the new annotation considering the genomic location shifts caused by the gap insertions in the new genome release (see 2.4).

Deeper investigation on the tomato *iTAG* gene prediction also highlighted several issues in the annotation regarding the miss-annotated and ambiguous genes. We found that many genes from *iTAG* have multiple copies on the genome overlapping other predicted loci. In many cases the genes are also

mapped on regions where no gene were predicted by the consortium. Moreover, 303 genes were predicted completely inside the repeated regions.

We also detected several genes that are predicted as split genes with respect to the possible correct loci (putative split genes). These issues can highly compromise genomic analyses such as gene expression quantification and functional investigations. To give support to this end we contributed a revised version of the *iTAG* gene annotation to highlight, and in some cases correct, these issues.

Due to the availability of the two different annotation pipeline for tomato (*iTAG* and *RefSeq*), we processed them into different alternative annotations described to meet the need of the interested scientific community. We also further analyzed the *Heinz RNAseq* collection on the basis of these annotations, and the results revealed a better coverage for the uniquely mapped reads on the genome for the *iTAG* annotation complemented with *RefSeq* (*iTAG* preferred, see 2.4.2). Deeper analyses are however required to define an updated annotation for tomato.

The effort to organize resources for tomato resulted in the several dedicated platforms. The aim was to allow the exploration and exploitation of the tomato genome space in an integrative way. The platforms are enriched with user friendly interfaces allowing ease of access to the processed collections using query dialog boxes. We setup a unique genome platform including both tomato genome sequences and the unmapped BACs. The availability of querying all the EST, TC and unigene collections mapped on both genomes together with the availability of different annotations are some of the peculiarities of our infrastructure. Indeed, though several reference sites are available for tomato [55, 135, 136], no platform provide access to the whole Solanaceae EST, TC

and universal unigene collections together in as tarcks mapped on a common genome reference.

We also designed and setup an expression platform in which, at the current setting for tomato (*NexGenEx-Tom*), access to different processed NGS atlas collections (*S. lycopersicum* cv. *Heinz*, *S. pimpinellifolium* and *Ailsa Craig*) was made available. The possibility of gene expression profiling and differentially expression analyses in one single click, and the availability of different online toolboxes for the cluster analyses and GO Enrichment are the main peculiarities of this platform. *NexGenEx-Tom* is also enriched with a cross link to the tomato genome platform.

We also implemented an otholog database and the dedicated interfaces enriched with different ortholog collections (see two ortholog collections sections 2.8.3) allowing the extensive comparative genomics between different species.

The different transcriptome collections from ESTs, TCs, and unigenes were processed when necessary (raw ESTs) and integrated in the infrastructure. Thanks to the availability of such transcriptome data, besides of being useful for the comparative genomic and exhaustive genomics analyses, they supported us to better exploit the tomato genome reference and its genomic content. Using these collections, we also investigated the content of the 112 unmapped BAC sequences (those that were not anchored to the tomato pseudomolecules), providing information not considered in the reference tomato genome sequences.

The availability of some public and private NGS data collections allowed us to further investigate the tomato genome space in terms of its expression content. As an example, the availability of the *Heinz* expression data from 11 tissue

stages supported general overviews on gene expression in the different tissues/stages.

We also presented that to properly process, integrate and investigate such data collections, various tools analytical approaches were necessary, with some of them implemented during the thesis work. Examples come from *Tracker*, a tool to conduct annotation free analyses due to the limits in the gene annotation, and *Overlapper*, a tool to intersect different genomic features; and NGS data analyses pipelines that analyze NGS data collections in parallel; etc. (see 3.2) that were implemented to allow genome wide investigations when a preliminary gene annotation is available.

Getting advantage of the developed tools, approaches and the bioinformatics platform we setup, we were able to carry out the integrative analyses on the tomato pollen developmental stages deciphering the role of heat shock on the gene expression and on genome reorganization in terms of methylation of CpGs changes. Our finding suggested that the genome methylation is affecting the gene expression during the developmental stages as the plant response to the stress (see 3.5.4).

We also applied our methodologies and tools for the identification of the TE-derived small-RNAs characterizing the role of the 21-22 and 24 nt *Small-RNA* fragments on the silencing of the Transposon Elements during the developmental stages of tomato pollen in physiological stages and also in comparison with the heat shock conditions. Based on our analyses, no significant change was observed between the similar stages under control and heat stress. On the contrary, a significant change is observed in the 21.22 and 24 nt *Small-RNA* classes between the developmental stages. Notably, we also found that the abundance of 21-22nt class and 24nt class *Small-RNA*s were significantly switched during pollen development. We also intersected these changes with the expression and methylation for the pollen developmental

stages in the physiological condition. Our findings suggests that the differentially regulation of the small-RNAs might have some effects on the adjacent genes expression level while no significant methylation changes in the developmental stages of pollen in the physiological condition were observed (see 3.5.5).

References

1. Müller, H., F. Naumann, and J.-C. Freytag, *Data quality in genome databases*. 2003.

2. Chou, H.-H. and M.H. Holmes, *DNA sequence quality trimming and vector removal*. Bioinformatics, 2001. **17**(12): p. 1093-1104.

3. Li, S. and H.-H. Chou, *LUCY2: an interactive DNA sequence quality trimming and vector removal tool*. Bioinformatics, 2004. **20**(16): p. 2865-2866.

4. Falgueras, J., et al., *SeqTrim: a high-throughput pipeline for pre-processing any type of sequence read*. BMC bioinformatics, 2010. **11**(1): p. 38.

5. Network, U.C.P.B., *and its impact on plant science*. 1998.

6. Morozova, O. and M.A. Marra, *Applications of next-generation sequencing technologies in functional genomics*. Genomics, 2008. **92**(5): p. 255-264.

7. Metzker, M.L., *Sequencing technologies—the next generation*. Nature reviews genetics, 2010. **11**(1): p. 31-46.

8. Benson, D.A., et al., *GenBank*. Nucleic acids research, 2013. **41**(D1): p. D36-D42.

9. Parra, G., K. Bradnam, and I. Korf, *CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes*. Bioinformatics, 2007. **23**(9): p. 1061-1067.

10. Seemann, T., *Prokka: rapid prokaryotic genome annotation*. Bioinformatics, 2014: p. btu153.

11. Sugawara, H., et al. *Microbial genome annotation pipeline (MiGAP) for diverse users*. in *The 20th International Conference on Genome Informatics (GIW2009)*. 2009.

12. Sallet, E., J. Gouzy, and T. Schiex, *EuGene-PP: A Next Generation Automated Annotation Pipeline for Prokaryotic Genomes*. Bioinformatics, 2014: p. btu366.

13. Pruitt, K.D., T. Tatusova, and D.R. Maglott, *NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes,*

transcripts and proteins. Nucleic acids research, 2007. **35**(suppl 1): p. D61-D65.

14. Boguski, M.S., T.M. Lowe, and C.M. Tolstoshev, *dbEST—database for "expressed sequence tags"*. Nature genetics, 1993. **4**(4): p. 332-333.

15. Leinonen, R., H. Sugawara, and M. Shumway, *The sequence read archive*. Nucleic acids research, 2011. **39**(suppl 1): p. D19-D21.

16. D'Agostino, N., M. Aversano, and M.L. Chiusano, *ParPEST: a pipeline for EST data analysis based on parallel computing*. BMC bioinformatics, 2005. **6**(Suppl 4): p. S9.

17. Parkinson, J. and M. Blaxter, *Expressed sequence tags: an overview*, in *Expressed sequence tags (ESTs)*. 2009, Springer. p. 1-12.

18. *EST NCBI FACT SHEET (https://www.cs.duke.edu/courses/cps006g/fall04/class/isis/estfaq.pdf)*. Available from: https://www.cs.duke.edu/courses/cps006g/fall04/class/isis/estfaq.pdf.

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

20. Müller, S., et al., *APADB: a database for alternative polyadenylation and microRNA regulation events*. Database, 2014. **2014**: p. bau076.

21. Mercer, T.R., M.E. Dinger, and J.S. Mattick, *Long non-coding RNAs: insights into functions*. Nature Reviews Genetics, 2009. **10**(3): p. 155-159.

22. Wilusz, J.E., H. Sunwoo, and D.L. Spector, *Long noncoding RNAs: functional surprises from the RNA world*. Genes & development, 2009. **23**(13): p. 1494-1504.

23. He, L. and G.J. Hannon, *MicroRNAs: small RNAs with a big role in gene regulation*. Nature Reviews Genetics, 2004. **5**(7): p. 522-531.

24. Chen, X., *Small RNAs and their roles in plant development*. Annual Review of Cell and Developmental, 2009. **25**: p. 21-44.

25. Malone, C.D. and G.J. Hannon, *Small RNAs as guardians of the genome*. Cell, 2009. **136**(4): p. 656-668.

26. Elbashir, S.M., W. Lendeckel, and T. Tuschl, *RNA interference is mediated by 21-and 22-nucleotide RNAs*. Genes & development, 2001. **15**(2): p. 188-200.

27. Pasquinelli, A.E., et al., *Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA.* Nature, 2000. **408**(6808): p. 86-89.

28. Reinhart, B.J. and D.P. Bartel, *Small RNAs correspond to centromere heterochromatic repeats.* Science, 2002. **297**(5588): p. 1831-1831.

29. Ambros, V., et al., *MicroRNAs and other tiny endogenous RNAs in C. elegans.* Current Biology, 2003. **13**(10): p. 807-818.

30. Lee, R.C. and V. Ambros, *An extensive class of small RNAs in Caenorhabditis elegans.* Science, 2001. **294**(5543): p. 862-864.

31. Lagos-Quintana, M., et al., *Identification of novel genes coding for small expressed RNAs.* Science, 2001. **294**(5543): p. 853-858.

32. Finnegan, E.J. and M.A. Matzke, *The small RNA world.* Journal of cell science, 2003. **116**(23): p. 4689-4693.

33. Mattick, J.S. and I.V. Makunin, *Non-coding RNA.* Human molecular genetics, 2006. **15**(suppl 1): p. R17-R29.

34. Zawada, A.M., et al., *SuperTAG methylation-specific digital karyotyping (SMSDK) reveals uremia induced epigenetic dysregulation of atherosclerosis-related genes.* Circulation: Cardiovascular Genetics, 2012: p. CIRCGENETICS. 112.963207.

35. Guerrero-Bosagna, C., *DNA Methylation Research Methods.* 2014.

36. Bird, A.P., *CpG-rich islands and the function of DNA methylation.* Nature, 1985. **321**(6067): p. 209-213.

37. Razin, A. and H. Cedar, *DNA methylation and gene expression.* Microbiological reviews, 1991. **55**(3): p. 451-458.

38. Baylin, S.B., et al., *Aberrant patterns of DNA methylation, chromatin formation and gene expression in cancer.* Human molecular genetics, 2001. **10**(7): p. 687-692.

39. Razin, A. and A.D. Riggs, *DNA methylation and gene function.* Science, 1980. **210**(4470): p. 604-610.

40. Bedford, M.T. and S. Richard, *Arginine methylation: an emerging regulatorof protein function.* Molecular cell, 2005. **18**(3): p. 263-272.

41. Boyes, J. and A. Bird, *DNA methylation inhibits transcription indirectly via a methyl-CpG binding protein.* Cell, 1991. **64**(6): p. 1123-1134.

42. Mello, C.C. and D. Conte, *Revealing the world of RNA interference.* Nature, 2004. **431**(7006): p. 338-342.

43. Selker, E.U., *Gene silencing: repeats that count.* Cell, 1999. **97**(2): p. 157-160.

44. Li, Y., et al., *The inheritance pattern of 24 nt siRNA clusters in Arabidopsis hybrids is influenced by proximity to transposable elements.* 2012.

45. Smit, A.F., R. Hubley, and P. Green, *RepeatMasker Open-3.0*. 1996.

46. Andrews, S., *FastQC: A quality control tool for high throughput sequence data*. Reference Source, 2010.

47. Felix, K. *Trim Galore*. Available from: http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.

48. Bolger, A.M., M. Lohse, and B. Usadel, *Trimmomatic: a flexible trimmer for Illumina sequence data*. Bioinformatics, 2014: p. btu170.

49. Gremme, G., *GenomeThreader Gene Prediction Software*. 2014.

50. Huang, X. and A. Madan, *CAP3: A DNA sequence assembly program.* Genome research, 1999. **9**(9): p. 868-877.

51. Haas, B.J., et al., *De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis.* Nature protocols, 2013. **8**(8): p. 1494-1512.

52. Kim, D., et al., *TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions*. Genome Biol, 2013. **14**(4): p. R36.

53. Chiusano, M.L., et al., *ISOL@: an Italian SOLAnaceae genomics resource*. BMC bioinformatics, 2008. **9**(Suppl 2): p. S7.

54. Bostan, H. and M.L. Chiusano, *NexGenEx-Tom: a gene expression platform to investigate the functionalities of the tomato genome*. BMC Plant Biology, 2015. **15**(1): p. 48.

55. Mueller, L.A., et al., *The SOL Genomics Network. A comparative resource for Solanaceae biology and beyond.* Plant physiology, 2005. **138**(3): p. 1310-1317.

56. Duvick, J., et al., *PlantGDB: a resource for comparative plant genomics.* Nucleic acids research, 2008. **36**(suppl 1): p. D959-D965.

57. Antonescu, C., et al., *Using the DFCI gene index databases for biological discovery*. Current Protocols in Bioinformatics, 2010: p. 1.6. 1-1.6. 36.

58. Zerbino, D.R. and E. Birney, *Velvet: algorithms for de novo short read assembly using de Bruijn graphs*. Genome research, 2008. **18**(5): p. 821-829.

59. Xie, Y., et al., *SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads*. Bioinformatics, 2014. **30**(12): p. 1660-1666.

60. Trapnell, C., et al., *Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks*. Nature protocols, 2012. **7**(3): p. 562-578.

61. Warren, R.L., et al., *Assembling millions of short DNA sequences using SSAKE*. Bioinformatics, 2007. **23**(4): p. 500-501.

62. Hsu, S.-D., et al., *miRTarBase: a database curates experimentally validated microRNA–target interactions*. Nucleic acids research, 2010: p. gkq1107.

63. Grabherr, M.G., et al., *Full-length transcriptome assembly from RNA-Seq data without a reference genome*. Nature biotechnology, 2011. **29**(7): p. 644-652.

64. Trapnell, C., et al., *Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation*. Nature biotechnology, 2010. **28**(5): p. 511-515.

65. MacQueen, J. *Some methods for classification and analysis of multivariate observations*. in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. 1967. Oakland, CA, USA.

66. Hollister, J.D., et al., *Transposable elements and small RNAs contribute to gene expression divergence between Arabidopsis thaliana and Arabidopsis lyrata*. Proceedings of the National Academy of Sciences, 2011. **108**(6): p. 2322-2327.

67. Calarco, J.P., et al., *Reprogramming of DNA methylation in pollen guides epigenetic inheritance via small RNA*. Cell, 2012. **151**(1): p. 194-205.

68. Anders, S., P.T. Pyl, and W. Huber, *HTSeq–A Python framework to work with high-throughput sequencing data*. bioRxiv, 2014.

69. Liao, Y., G.K. Smyth, and W. Shi, *featureCounts: an efficient general purpose program for assigning sequence reads to genomic features.* Bioinformatics, 2014. **30**(7): p. 923-930.

70. Hansen, K.D., R.A. Irizarry, and W. Zhijin, *Removing technical variability in RNA-seq data using conditional quantile normalization.* Biostatistics, 2012. **13**(2): p. 204-216.

71. Chuai-Aree, S., et al., *Fuzzy c-mean: A statistical feature classification of text and image segmentation method.* International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2001. **9**(06): p. 661-671.

72. Törönen, P., et al., *Analysis of gene expression data using self-organizing maps.* FEBS letters, 1999. **451**(2): p. 142-146.

73. Oshlack, A. and M.J. Wakefield, *Transcript length bias in RNA-seq data confounds systems biology.* Biol Direct, 2009. **4**(1): p. 14.

74. Sturn, A., J. Quackenbush, and Z. Trajanoski, *Genesis: cluster analysis of microarray data.* Bioinformatics, 2002. **18**(1): p. 207-208.

75. Sherlock, G., *Analysis of large-scale gene expression data.* Current opinion in immunology, 2000. **12**(2): p. 201-205.

76. Robinson, M.D. and A. Oshlack, *A scaling normalization method for differential expression analysis of RNA-seq data.* Genome Biol, 2010. **11**(3): p. R25.

77. Mortazavi, A., et al., *Mapping and quantifying mammalian transcriptomes by RNA-Seq.* Nature methods, 2008. **5**(7): p. 621-628.

78. Bullard, J.H., et al., *Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments.* BMC bioinformatics, 2010. **11**(1): p. 94.

79. Anders, S., *Analysing RNA-Seq data with the DESeq package.* Molecular biology, 2010: p. 1-17.

80. Nikkilä, J., et al., *Analysis and visualization of gene expression data using self-organizing maps.* Neural networks, 2002. **15**(8): p. 953-966.

81. Meinicke, P., et al., *Metabolite-based clustering and visualization of mass spectrometry data using one-dimensional self-organizing maps.* Algorithms for Molecular Biology, 2008. **3**(9): p. P13.

82. Yang, Z.R. and K.-C. Chou, *Mining biological data using self-organizing map*. Journal of chemical information and computer sciences, 2003. **43**(6): p. 1748-1753.

83. Mardia, K.V., J.T. Kent, and J.M. Bibby, *Multivariate analysis*. 1979: Academic press.

84. Ketchen, D.J. and C.L. Shook, *The application of cluster analysis in strategic management research: an analysis and critique*. Strategic management journal, 1996. **17**(6): p. 441-458.

85. Griffiths-Jones, S., et al., *miRBase: microRNA sequences, targets and gene nomenclature*. Nucleic acids research, 2006. **34**(suppl 1): p. D140-D144.

86. Griffiths-Jones, S., et al., *miRBase: tools for microRNA genomics*. Nucleic acids research, 2008. **36**(suppl 1): p. D154-D158.

87. Barabasi, A.-L. and Z.N. Oltvai, *Network biology: understanding the cell's functional organization*. Nature reviews genetics, 2004. **5**(2): p. 101-113.

88. Zhang, B. and S. Horvath, *A general framework for weighted gene co-expression network analysis*. Statistical applications in genetics and molecular biology, 2005. **4**(1).

89. Berkhin, P., *A survey of clustering data mining techniques*, in *Grouping multidimensional data*. 2006, Springer. p. 25-71.

90. Bao, R., et al., *Review of current methods, applications, and data management for the bioinformatics analysis of whole exome sequencing.* Cancer informatics, 2014. **13**(Suppl 2): p. 67.

91. Cao, Y. *A simple but fast tool for K-means clustering*. 2008.

92. *A Tutorial on Clustering Algorithms*. Available from: http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/cmeans.html.

93. Kohonen, T., *The self-organizing map*. Neurocomputing, 1998. **21**(1): p. 1-6.

94. Vesanto, J. and E. Alhoniemi, *Clustering of the self-organizing map*. Neural Networks, IEEE Transactions on, 2000. **11**(3): p. 586-600.

95. Tamayo, P., et al., *Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic*

differentiation. Proceedings of the National Academy of Sciences, 1999. **96**(6): p. 2907-2912.

96. Jae-Wook Ahn, S.Y.S. *Self-Organizing Maps*. 2005.

97. Maksim. *Determining the number of clusters in a data set*. 2006.

98. Jiang, D., C. Tang, and A. Zhang, *Cluster analysis for gene expression data: A survey*. Knowledge and Data Engineering, IEEE Transactions on, 2004. **16**(11): p. 1370-1386.

99. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles*. Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(43): p. 15545-15550.

100. Huang, D.W., B.T. Sherman, and R.A. Lempicki, *Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists*. Nucleic acids research, 2009. **37**(1): p. 1-13.

101. Consortium, G.O., *Gene ontology consortium: going forward*. Nucleic acids research, 2015. **43**(D1): p. D1049-D1056.

102. Kinsella, R.J., et al., *Ensembl BioMarts: a hub for data retrieval across taxonomic space*. Database, 2011. **2011**: p. bar030.

103. Conesa, A., et al., *Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research*. Bioinformatics, 2005. **21**(18): p. 3674-3676.

104. Du, Z., et al., *agriGO: a GO analysis toolkit for the agricultural community*. Nucleic acids research, 2010: p. gkq310.

105. Eden, E., et al., *GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists*. BMC bioinformatics, 2009. **10**(1): p. 48.

106. Zhou, X. and Z. Su, *EasyGO: Gene Ontology-based annotation and functional enrichment analysis tool for agronomical species*. BMC genomics, 2007. **8**(1): p. 246.

107. Bauer, S., et al., *Ontologizer 2.0—a multifunctional tool for GO term enrichment analysis and data exploration*. Bioinformatics, 2008. **24**(14): p. 1650-1651.

108. Zheng, Q. and X.-J. Wang, *GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis*. Nucleic acids research, 2008. **36**(suppl 2): p. W358-W363.

109. D'Agostino, N., et al., *SolEST database: a*. BMC plant biology, 2009. **9**(1): p. 142.

110. Altenhoff, A.M., et al., *OMA 2011: orthology inference among 1000 complete genomes*. Nucleic Acids Res, 2011. **39**(Database issue): p. D289-94.

111. Chen, F., et al., *OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups*. Nucleic acids research, 2006. **34**(suppl 1): p. D363-D368.

112. Dessimoz, C., et al., *OMA, A Comprehensive, Automated Project for the Identification of Orthologs from Complete Genome Data: Introduction and First Achievements*, in *Comparative Genomics*, A. McLysaght and D. Huson, Editors. 2005, Springer Berlin Heidelberg. p. 61-72.

113. Dessimoz, C., et al., *Toward community standards in the quest for orthologs*. Bioinformatics, 2012. **28**(6): p. 900-4.

114. Flicek, P., et al., *Ensembl 2013*. Nucleic Acids Res, 2013. **41**(Database issue): p. D48-55.

115. Huerta-Cepas, J., et al., *PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome*. Nucleic Acids Res, 2014. **42**(Database issue): p. D897-902.

116. Hulsen, T., et al., *Benchmarking ortholog identification methods using functional genomics data*. Genome Biol, 2006. **7**(4): p. R31.

117. Kuzniar, A., et al., *The quest for orthologs: finding the corresponding gene across genomes*. Trends Genet, 2008. **24**(11): p. 539-51.

118. O'Brien, K.P., M. Remm, and E.L. Sonnhammer, *Inparanoid: a comprehensive database of eukaryotic orthologs*. Nucleic acids research, 2005. **33**(suppl 1): p. D476-D480.

119. Powell, S., et al., *eggNOG v4.0: nested orthology inference across 3686 organisms*. Nucleic Acids Res, 2014. **42**(Database issue): p. D231-9.

120. Proost, S., et al., *PLAZA: a comparative genomics resource to study gene and genome evolution in plants*. The Plant Cell, 2009. **21**(12): p. 3718-3731.

121. Rouard, M., et al., *GreenPhylDB v2.0: comparative and functional genomics in plants*. Nucleic Acids Res, 2011. **39**(Database issue): p. D1095-102.

122. Schreiber, F., et al., *TreeFam v9: a new website, more species and orthology-on-the-fly*. Nucleic Acids Res, 2014. **42**(Database issue): p. D922-5.

123. Waterhouse, R.M., et al., *OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs*. Nucleic Acids Res, 2013. **41**(Database issue): p. D358-65.

124. Altenhoff, A.M. and C. Dessimoz, *Phylogenetic and functional assessment of orthologs inference projects and methods*. PLoS Comput Biol, 2009. **5**(1): p. e1000262.

125. Dolinski, K. and D. Botstein, *Orthology and functional conservation in eukaryotes*. Annu Rev Genet, 2007. **41**: p. 465-507.

126. Koonin, E.V., *Orthologs, paralogs, and evolutionary genomics*. Annu Rev Genet, 2005. **39**: p. 309-38.

127. Sonnhammer, E.L. and E.V. Koonin, *Orthology, paralogy and proposed classification for paralog subtypes*. Trends Genet, 2002. **18**(12): p. 619-20.

128. Quinlan, A.R. and I.M. Hall, *BEDTools: a flexible suite of utilities for comparing genomic features*. Bioinformatics, 2010. **26**(6): p. 841-842.

129. Johnson, C., et al., *Clusters and superclusters of phased small RNAs in the developing inflorescence of rice*. Genome research, 2009. **19**(8): p. 1429-1440.

130. Quackenbush, J., et al., *The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species*. Nucleic Acids Research, 2001. **29**(1): p. 159-164.

131. Luscombe, N.M., D. Greenbaum, and M. Gerstein, *What is bioinformatics? A proposed definition and overview of the field*. Methods of information in medicine, 2001. **40**(4): p. 346-358.

132. D'Agostino, N., et al., *TomatEST database: in silico exploitation of EST data to explore expression patterns in tomato species*. Nucleic Acids Research, 2007. **35**(suppl 1): p. D901-D905.

133. Yano, K., et al. *KaFTom: Full-length cDNA Database of a miniature tomato cultivar Micro-Tom*. in *PLANT AND CELL PHYSIOLOGY*.

2007. OXFORD UNIV PRESS GREAT CLARENDON ST, OXFORD OX2 6DP, ENGLAND.

134. Yano, K., et al., *MiBASE: a database of a miniature tomato cultivar Micro-Tom.* Plant Biotechnology, 2006. **23**(2): p. 195-198.

135. Fei, Z., et al., *Tomato Functional Genomics Database: a comprehensive resource and analysis package for tomato functional genomics.* Nucleic acids research, 2011. **39**(suppl 1): p. D1156-D1163.

136. Suresh, B.V., et al., *Tomato genomic resources database: an integrated repository of useful tomato genomic information for basic and applied research.* PloS one, 2014. **9**(1).

137. Consortium, T.G., *The tomato genome sequence provides insights into fleshy fruit evolution.* Nature, 2012. **485**(7400): p. 635-641.

138. Hummel, J., et al., *ProMEX: a mass spectral reference database for proteins and protein phosphorylation sites.* BMC bioinformatics, 2007. **8**(1): p. 216.

139. Kanehisa, M. and S. Goto, *KEGG: kyoto encyclopedia of genes and genomes.* Nucleic acids research, 2000. **28**(1): p. 27-30.

140. Lu, S., *Plant Metabolic Network*, in *Plant Metabolomics*. 2015, Springer. p. 195-211.

141. Kent, W.J., et al., *The human genome browser at UCSC.* Genome research, 2002. **12**(6): p. 996-1006.

142. Thorvaldsdóttir, H., J.T. Robinson, and J.P. Mesirov, *Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration.* Briefings in bioinformatics, 2012: p. bbs017.

143. Sharma, S.K., et al., *Construction of reference chromosome-scale pseudomolecules for potato: integrating the potato genome with genetic and physical maps.* G3: Genes| Genomes| Genetics, 2013. **3**(11): p. 2031-2047.

144. Nawrocki, E.P., D.L. Kolbe, and S.R. Eddy, *Infernal 1.0: inference of RNA alignments.* Bioinformatics, 2009. **25**(10): p. 1335-1337.

145. Langfelder, P. and S. Horvath, *WGCNA: an R package for weighted correlation network analysis.* BMC bioinformatics, 2008. **9**(1): p. 559.

146. Caporaso, J.G., et al., *Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms.* The ISME journal, 2012. **6**(8): p. 1621-1624.

147. Zhong, S., et al., *Single-base resolution methylomes of tomato fruit development reveal epigenome modifications associated with ripening.* Nature biotechnology, 2013. **31**(2): p. 154-159.

148. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nature methods, 2012. **9**(4): p. 357-359.

149. Voorrips, R., *MapChart: software for the graphical presentation of linkage maps and QTLs.* Journal of Heredity, 2002. **93**(1): p. 77-78.

150. Bokszczanin, K.L., et al., *Identification of novel small ncRNAs in pollen of tomato.* BMC Genomics, 2015. **16**(1): p. 714.

151. Fruhwirt, P., et al. *Innodb database forensics.* in *Advanced Information Networking and Applications (AINA), 2010 24th IEEE International Conference on.* 2010. IEEE.

152. Bayer, R., *The universal B-tree for multidimensional indexing: General concepts,* in *Worldwide Computing and Its Applications.* 1997, Springer. p. 198-209.

153. Goodstein, D.M., et al., *Phytozome: a comparative platform for green plant genomics.* Nucleic acids research, 2012. **40**(D1): p. D1178-D1186.

154. Li, H., et al., *The sequence alignment/map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-2079.

155. Martin, M., *Cutadapt removes adapter sequences from high-throughput sequencing reads.* EMBnet. journal, 2011. **17**(1): p. pp. 10-12.

156. Soneson, C. and M. Delorenzi, *A comparison of methods for differential expression analysis of RNA-seq data.* BMC bioinformatics, 2013. **14**(1): p. 91.

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

158. Pruitt, K.D., T. Tatusova, and D.R. Maglott, *NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.* Nucleic acids research, 2005. **33**(suppl 1): p. D501-D504.

159. Seo, S., *A review and comparison of methods for detecting outliers in univariate data sets*. 2006, University of Pittsburgh.

160. Carbon, S., et al., *AmiGO: online access to ontology and annotation data*. Bioinformatics, 2009. **25**(2): p. 288-289.

161. Holoch, D. and D. Moazed, *RNA-mediated epigenetic regulation of gene expression*. Nature Reviews Genetics, 2015.

162. Nuthikattu, S., et al., *The initiation of epigenetic silencing of active transposable elements is triggered by RDR6 and 21-22 nucleotide small interfering RNAs*. Plant physiology, 2013. **162**(1): p. 116-131.

163. Haag, J.R. and C.S. Pikaard, *Multisubunit RNA polymerases IV and V: purveyors of non-coding RNA for plant gene silencing*. Nature reviews Molecular cell biology, 2011. **12**(8): p. 483-492.

164. Girard, L. and M. Freeling, *Regulatory changes as a consequence of transposon insertion*. Developmental genetics, 1999. **25**(4): p. 291-296.

165. Matzke, M.A. and R.A. Mosher, *RNA-directed DNA methylation: an epigenetic pathway of increasing complexity*. Nature Reviews Genetics, 2014. **15**(6): p. 394-408.

166. Kao, D. *Concept: Pearson vs Spearman correlation for RNA-seq comparisons*. 2012; Available from: blog.nextgenetics.net/?e=46.

5 ANNEX I: Bioinformatics Tools

5.1.1 Bulk-Sorter

Motivation

Sorting is one of the routine parsing events which may be required very often when working with the text files. With the advent of high-throughput technologies and the amount of data they offer, management of these data files is a big challenge to overcome. Searching for a feature or record, intersecting 2 files to find the overlaps and indexing of the records in a flat file for easier and faster random access are of those processes that can be done easier and more efficiently on a sorted file. However, the memory resources on the computing machines are limited, and management of bulky files in memory can be impossible in most cases. Hence, the availability of such a tools to easily sort large files in short with low memory consumption is essential.

Description

Getting advantage of the Merge Sort approach [164], we developed a simple merge sorter (so called Bulk-Sorter) that can manage to sort text files with any size on any memory resource. Depending on the memory size the user specifies for the software, the tool splits the file into sub fragments in which each file is sorted independently. The merging process will then be done on the fragmented file parts considering the sorting of the incoming records from each fragmented part. Eventually the whole files is sorted into one merged file and the sub fragments are removed (refer to the Merge Sort Method Description). The tools is also implemented as an external module in the sorting section of the *Tracker* pipeline. The tool can be run under both *Windows* and *UNIX* environments.

5.1.2 Small-RNA Analyses Pipeline

Motivation

Several tools and pipelines exists for the *Small-RNA* target analyses [67, 96, 165]. The need for a Differentially Sites detection of *Small-RNAs* with the possibility of customizing its steps due to the needs and requirements of our analyses resulted to the development of a *Small-RNA* pipeline in which, different tuning of settings and thresholds can be orchestrated.

Description

The *Small-RNA* pipeline designed is a general Differentially Sites of *Small-RNAs* detection in which, 1) detection of *Small-RNA* classes with significant changes (e.g. 21+22 nt or 24 nt sequences), 2) categorization and classification of different *Small-RNA* classes to be subjected to the downstream analyses (keeping only the 24 nt sequences and discarding all the others), 3) selection of *Small-RNAs* on specific genomic regions with to address some dedicated biological questions (e.g. those only overlapping the repeated regions on the genome), and 4) clustering of adjacent *Small-RNA* sequences to create customized or universal reference genomic features for the counting is made possible in a convenient and efficient way. The sequence size selection can be conducted before or after the mapping process.

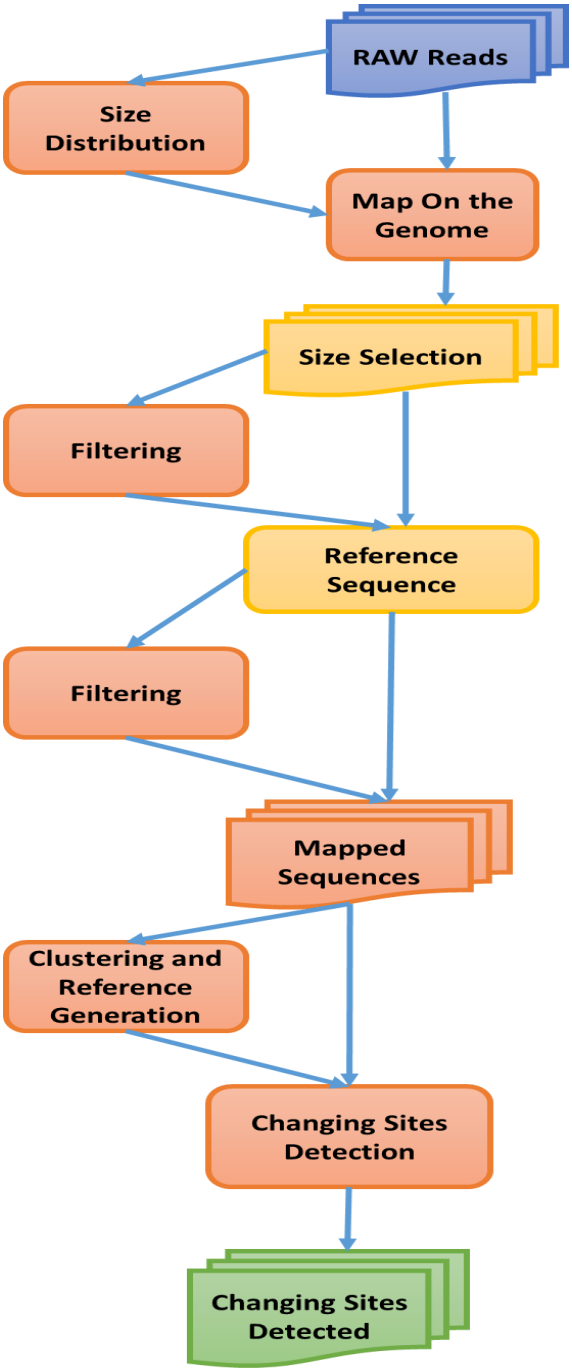


Figure 68: A general Small-RNA analyses workflow and pipeline schema

The detected Changing Clusters or Sites can be then intersected (using *Overlapper* software) with other genomic regions (coding or non-coding features) for the downstream analyses. The tool can be run under both *Windows* and *UNIX* environments.

5.1.3 Correlationer (maybe remove)

Motivation

Correlation analyses is the basis of many approaches in the field of bioinformatics. Generation of gene networks and reasoning on the relation of molecular components are often made by considering the level of correlation the 2 objects (genes or compounds) possess across different conditions. However the concept is quite simple and several packages implement the correlation coefficient calculation with a simple function (R environment, Matlab etc.), calculation of the correlation level among a list with thousands or millions of records is a challenge is not simply possible. In most cases such as or Matlab, an error indicating lack of memory is produced, or the result will be very hard to manage.

Description

With regards to the importance of correlation coefficient calculation as the basis of many approaches used, and due to the challenges and limits the correlation coefficient calculation of thousands of genes, all versus all, may introduce, we developed the tool "*Correlationer*" to memory efficiently calculate this value for all the components. The *Correlationer* calculates the Pearson and Spearman correlation coefficient of all the elements versus all the others (with any dimension) allowing to specify a threshold of correlation to discard those not passing that value. Hence, only the genes with a specific threshold specified as the accepted level of correlation (positive, negative or both) will appear in the output. Since the correlation coefficients are being calculated once at the time, the memory consumption is very low, but it also increase the time of processing. The tool can be run under both *Windows* and *UNIX* environments.

5.1.4 K-means calculator and Analyzer

Motivation

K-means clustering is one of the most common and popular clustering algorithms used for the omics data. It has been implemented in several packages (R environment and Matlab) which allows the users to cluster the list of elements specifying a cluster number (k). The possibility of rescaling or normalizing input data is also possible with some scripting in the respective environments. However, such tools exist, having a simple interactive tool efficiently working on bulk datasets, and producing final outputs with some stats (frequency of clusters and their distribution) is an advantage to obtain. Moreover, running the clustering with multiple number of clusters (different K values) is a good way to obtain the best cluster number disjointing the dataset groups.

Description

Here we developed a console application for the *k-means* cluster analyses and a supplementary package to test different K values to obtain the best cluster number for the analyses. The number of iterations, scaling by min-max scaling on rows or all the set can be easily specified as the input parameters. *K-means* calculator a fast and efficient package in which the outliers for each cluster and the frequency and distribution of the clusters are reported in separate output files. The tool can be run under both *Windows* and *UNIX* environments.

5.1.5 FastaToBatchMapper

Motivation

Bioinfromaticians are routinely facing the sequence mapping on the reference sequences. Often happens that thousands or millions of transcripts or protein sequences in the form of multi-fatsa files are supposed to be mapped versus a genome or reference sequence to identify its genomic origin or mapping location. Due to the advancement of computational technologies and the availability of parallel processing approaches, splitting of big jobs into smaller jobs and distributing them on different nodes/processors or even threads (if multi-threading available) are a common and useful methodology to be considered. The job management software applications simply manage your jobs and eventually the output files are produced. Normally the time consumed are divided by the number of sub-jobs you have ran in parallel.

Description

Here we present the *FastaToBatchMapper* as a simple tool in which the fasta sequences inside the multi-fasta file(s) are split into cluster of fasta sequences (the number of sequences in each file is specified by user), and the mapping of these sequences, specifying the mapping parameters depending on the mapping software, is parallelized on the available nodes and cores of the high-level computing machine. The software at the moment is designed to work with the *GenomeThreader* [49] mapping tool and *TORQUE* job management system (http://www.adaptivecomputing.com/products/open-source/torque/). The tool organizes the analyses of each fastaf ile in a separate folder collecting the mapping outputs in the result directory. The distributed mapping outputs are then combined, indexed, and parsed into a valid GFF3 file format.

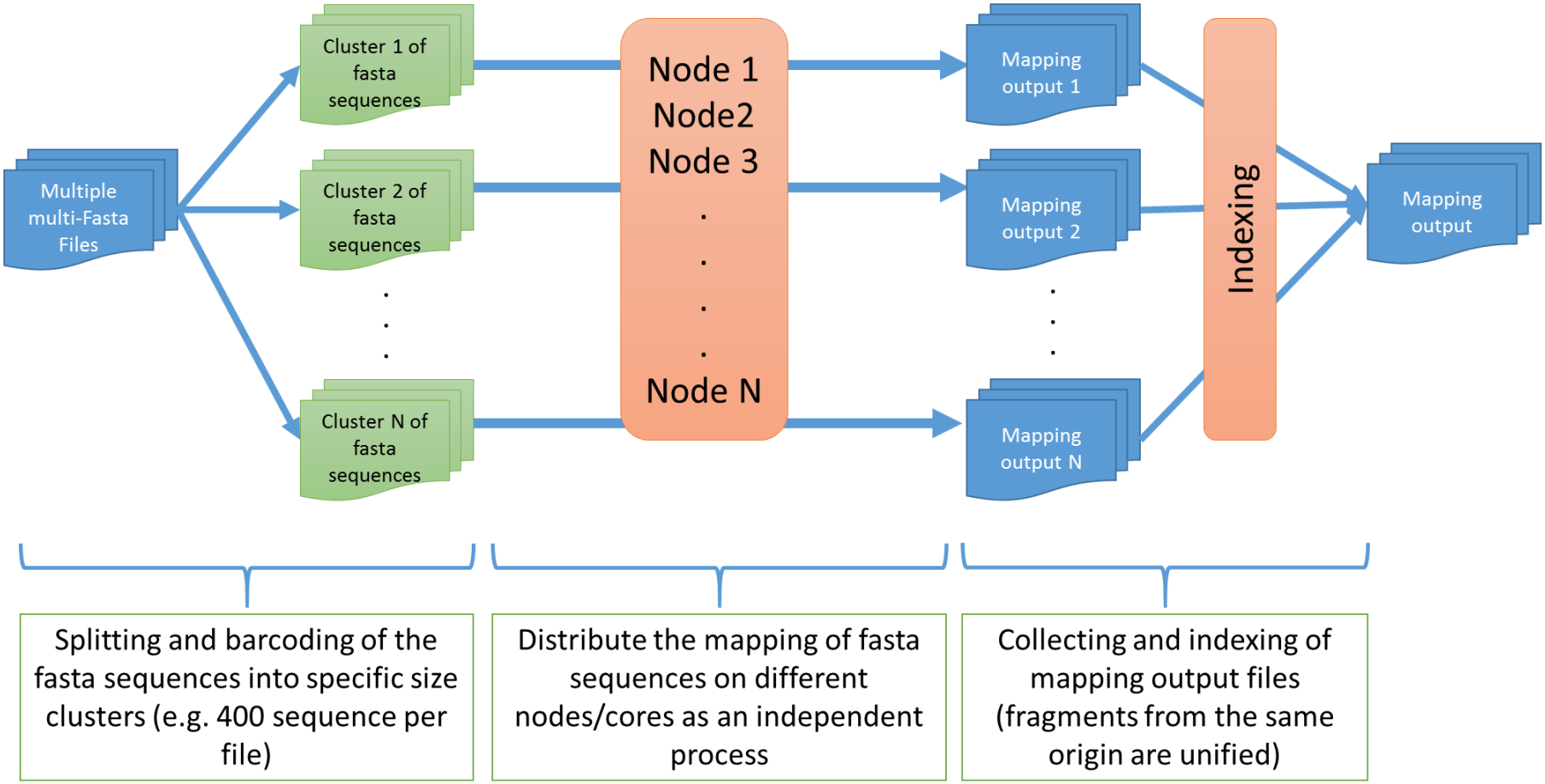


Figure 69: the schema of the distributed and parallelized sequence mapping pipeline

The tool can be run under both *Windows* and *UNIX* environments.

5.1.6 Genome Scanner

Motivation

In the genomics area, it may happen that the general knowledge concerning the dense, overlapping and noncoding regions of the reference genome is required. Several visualization tools such as IGV [], GenomeBrowser [] and JBrowse [] provide the possibility of browsing the genome reference (limited to a specific view) for this observation. Nonetheless, scanning of the whole genome to gain the detailed knowledge of each genomic region with customized annotation files is not easily possible.

Description

We developed a genome scanner tool in which the coverage of each base regarding its overlapping with any genomic region (specified as one of the inputs) is reported in details. Possibility of reporting the bases with specific coverage is also possible in the genome scanner. Genome Scanner is not very memory efficient but provides in-depth and detailed information regarding each base. The tool can be run under both *Windows* and *UNIX* environments.

5.1.7 Sequence Length Classifier

Motivation

In many genomic analyses (e.g. small-RNA or micro-RNA analyses), classification of sequences into respective length classes, before or after the mapping, is an important issue to be addressed. As an example, for the micro-RNA analyses targeting the coding sequences, filtering out all the sequences

longer or shorter than the accepted micro-RNA sequence length is essential. Besides of being able to have a robust idea on the mapped or not mapped sequences, the time for the mapping or analyzing of the non-relevance sequences are also reduced.

Description

Here we present a tool for size classification of sequences before (fastq file format) or after mapping (sam file format), in which a specific or multiple relevant size classes can be combined together. The output file format is identical with the input file introduced to the software, excluded from the non-relevant sequences. The tool can be run under both *Windows* and *UNIX* environments.

5.1.8 Sequence Length Distributioner

Motivation

Understanding of the length distribution of the available sequences in a sample can be important for several reasoning purposes. As an example, in a *Small-RNA* fastq file, understanding of the frequency of sequences with specific lengths can result to better evaluation of samples and more proper sequences size class extraction. Development of a tool parsing and calculating the fastq file in a distribution table is a handy tool to be available.

Description

Sequence Length *Distributioner* is a simple parser summarizing all the available lengths distribution in a/multiple fastq sequence files. The tool produces a tab

delimited distribution table for each sequence size. The tool can be run under both *Windows* and *UNIX* environments.

5.1.9 SequencePatternDetector

Motivation

Detection of all the specific sequence patterns on the reference sequence (e.g. Heat Shock Element binding sequences [163] detection or *CpG* sites for methylation analyses [7] etc.) can be often faced in the reference based data analyses. To the best of our knowledge, *IGV* provides rapid investigation of a sequence on the genome which allows the browsing of its genomic region on the genome reference. But still a tool to list all the existing patterns with their genomic locations can be very useful for the reference annotation production and changing sites detection.

Description

SequencePatternDetector is a tool to detect all the genomic or reference sequence locations matching a specific sequence pattern. For instance by providing the CCGG as the matching pattern sequence, all the possible sited that can be cut by *HpaII* will be extracted and their chromosomal locations will be listed in the output file. By checking the cut and not cut sites by the restriction enzyme for all the detected regions, the sites differentially methylated can be detected easily. The tool can be run under both *Windows* and *UNIX* environments.

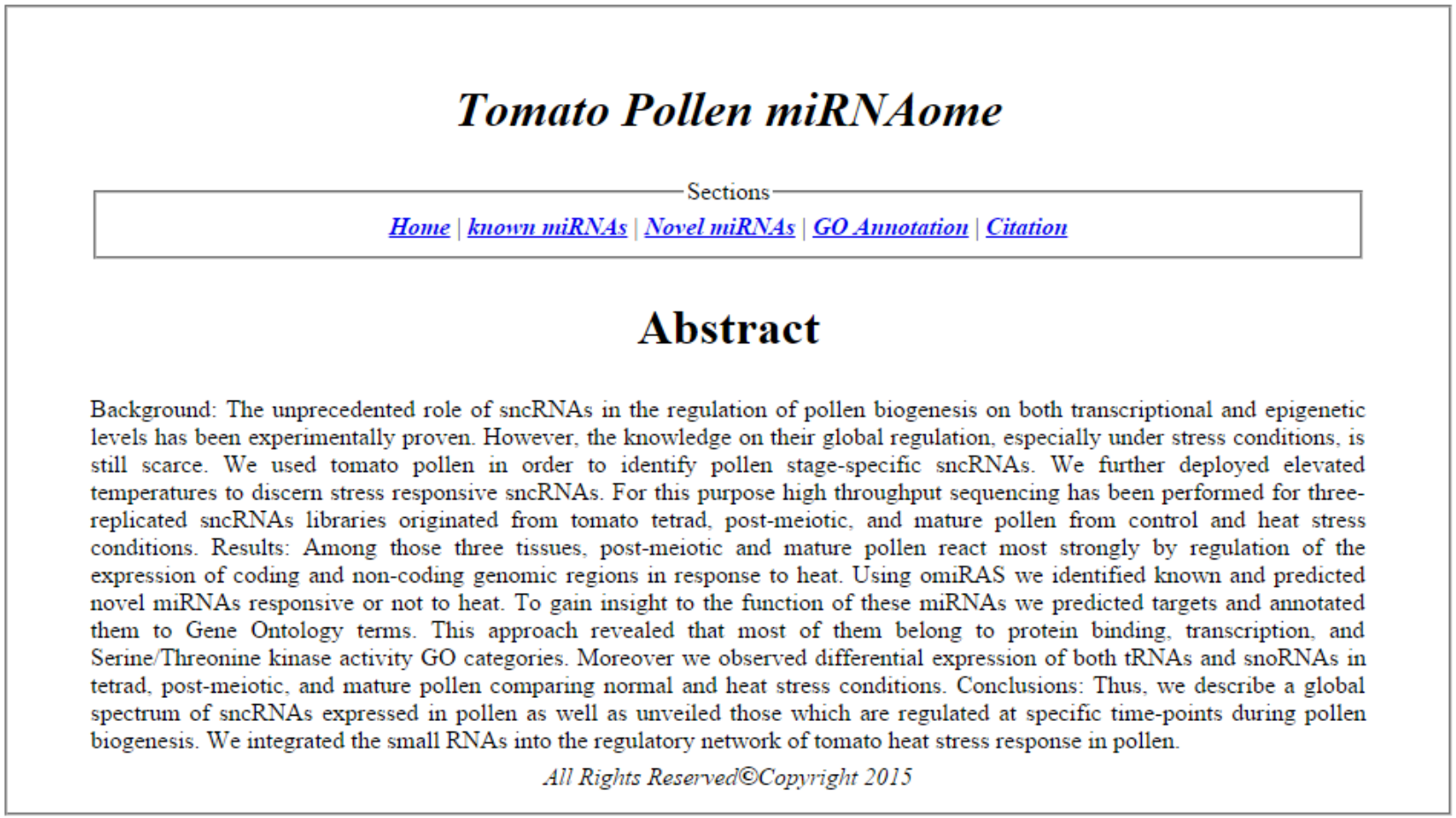
6 ANNEX II: Bioinformatics Platforms and Databases

6.1 Tomato Pollen *miRNAome*

I contribute the organization of a dedicate website organizing the collection of novel miRNAs independently detected from the partner *GenXPro*, Germany in the framework of the *SPOT-ITN* project (http://cab.unina.it/mirna-pollen/) [166].

6.1.1 User Interface and database access

As the default page (Figure 70), the website provides the navigation pane allowing to move to different result views and query pages implemented in the *Tomato Pollen miRNAome* database. The website is enriched with a cross-navigation to the genome browser implemented in the *Tomato Genome Platform* presented before (see 2.8.1).



Tomato Pollen miRNAome

Sections

Home | *known miRNAs* | *Novel miRNAs* | *GO Annotation* | *Citation*

Abstract

Background: The unprecedented role of sncRNAs in the regulation of pollen biogenesis on both transcriptional and epigenetic levels has been experimentally proven. However, the knowledge on their global regulation, especially under stress conditions, is still scarce. We used tomato pollen in order to identify pollen stage-specific sncRNAs. We further deployed elevated temperatures to discern stress responsive sncRNAs. For this purpose high throughput sequencing has been performed for three-replicated sncRNAs libraries originated from tomato tetrad, post-meiotic, and mature pollen from control and heat stress conditions. Results: Among those three tissues, post-meiotic and mature pollen react most strongly by regulation of the expression of coding and non-coding genomic regions in response to heat. Using omiRAS we identified known and predicted novel miRNAs responsive or not to heat. To gain insight to the function of these miRNAs we predicted targets and annotated them to Gene Ontology terms. This approach revealed that most of them belong to protein binding, transcription, and Serine/Threonine kinase activity GO categories. Moreover we observed differential expression of both tRNAs and snoRNAs in tetrad, post-meiotic, and mature pollen comparing normal and heat stress conditions. Conclusions: Thus, we describe a global spectrum of sncRNAs expressed in pollen as well as unveiled those which are regulated at specific time-points during pollen biogenesis. We integrated the small RNAs into the regulatory network of tomato heat stress response in pollen.

Figure 70: snapshot of the miRNA-Pollen webpage available at http://cab.unina.it/mirna-pollen/

Based on the procedure described in [166], the already known identified miRNAs listing their abundant changes between the three stages of pollen development (*Tetrad, post-meiotic, and Mature*) are presented in the "*Known miRNA*" section of the database. The adjusted p-value for the pairwise comparison of the miRNA abundance is also presented in the FDR column. Each of the columns can also be sorted by clicking on its header title (Figure 71).

Sections

Home | known miRNAs | Novel miRNAs | GO Annotation | Citation

ID	Ref.	norm.CT	norma.HST	log2fc_CT_HST	fdr_CT_HST	norm.CPM	norm.HSPM	log2fc_CPM_HSPM	fdr_CPM_HSPM	norm.CM	norm.HSM	log2fc_CM_HSM	fdr_CM_HSM
sly-miR-0210-tgdb	108	231.992	276.208	0.251677	0.889901	404.524	348.993	-0.213028	1	244.83	465.632	0.927409	0.272086
sly-miR-0211-tgdb	108	28.1823	37.2427	0.402168	0.889901	65.5089	29.6252	-1.14487	1	81.6194	181.019	1.14916	0.336385
sly-miR-022-tgdb	108	330.084	438.114	0.408475	0.889901	256.789	202.517	-0.342542	1	148.814	203.073	0.448482	0.379023
sly-miR-023-tgdb	108	332.415	441.309	0.408803	0.889901	259.37	204.629	-0.342002	1	150.854	204.754	0.440734	0.382306
sly-miR-024-tgdb	108	249.149	339.308	0.445587	0.889901	322.916	230.089	-0.488966	1	214.468	296.901	0.469219	0.462161
sly-miR-025-tgdb	108	179.005	232.656	0.378196	0.889901	147.067	104.444	-0.493748	1	115.102	155.931	0.437996	0.462161
sly-miR-027-tgdb	108	81.1818	96.4406	0.248485	0.889901	94.332	64.3835	-0.551056	1	96.5318	133.313	0.46574	0.462161
sly-miR-028-tgdb	108	81.1818	96.4406	0.248485	0.889901	94.332	64.3835	-0.551056	1	96.5318	133.313	0.46574	0.462161
sly-miR-0310-tgdb	108	81.0058	95.9744	0.244624	0.889901	94.2495	64.0529	-0.557221	1	96.5318	132.433	0.456183	0.462161
sly-miR-0311-tgdb	108	82.7078	100.928	0.287238	0.889901	96.781	65.4047	-0.565331	1	103.61	139.161	0.425585	0.470998
sly-miR-0312-	108	82.7078	100.928	0.287238	0.889901	96.781	65.4047	-0.565331	1	103.61	139.161	0.425585	0.470998

Figure 71: snapshot of known miRNAs from the platform

Based on the methodology and procedure described for the novel miRNAs detection in [150], the list of novel miRNAs identified in the collection including several accessory information such as genomic region (chromosome, start, end, and strand), the energy and sequences are provided for each of the stages (see 2.2) (Figure 72).

Sections

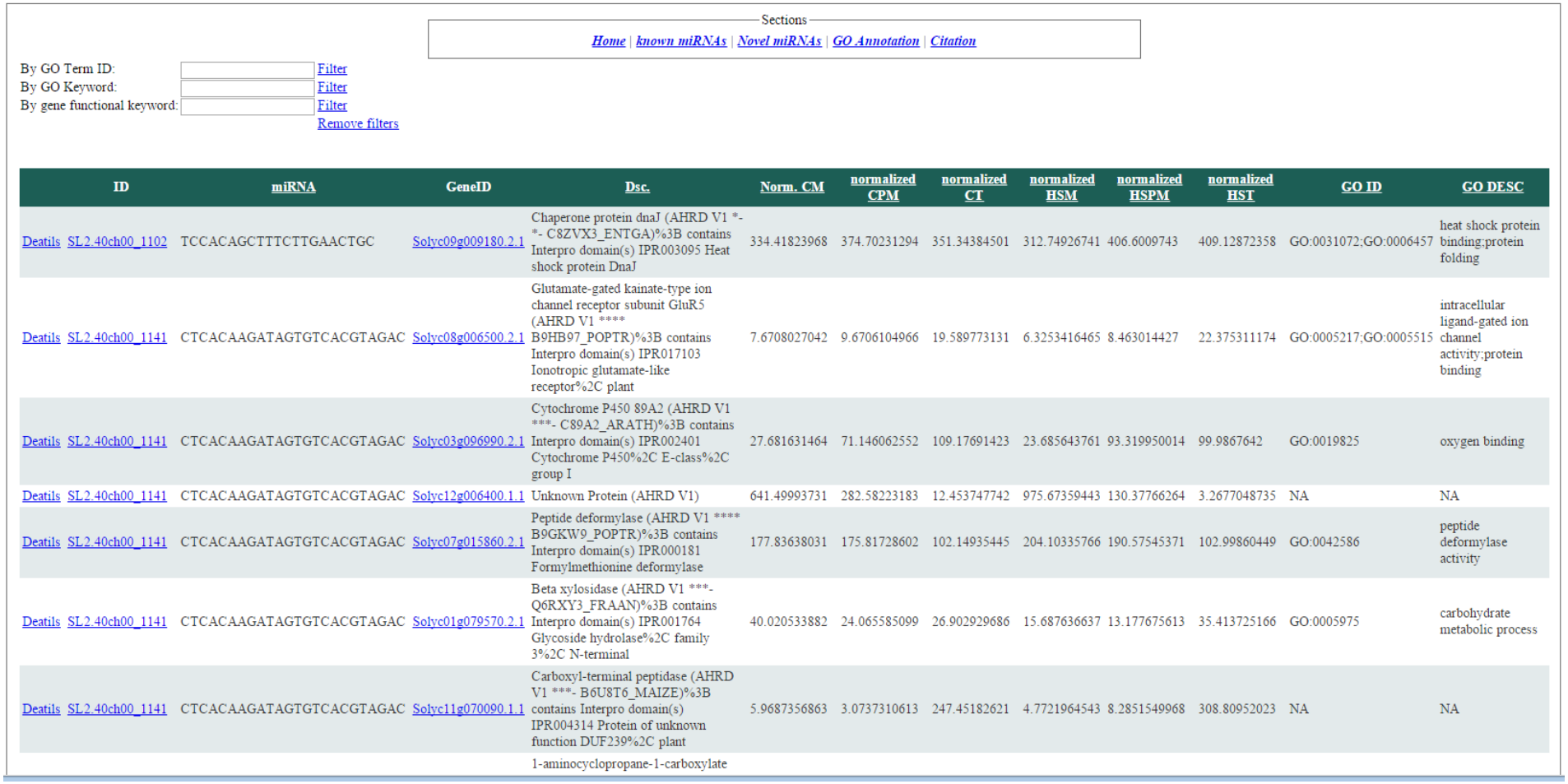
Home | known miRNAs | Novel miRNAs | GO Annotation | Citation

View by: All

stage_name	ID	chr	st	en	str	energy	star_sequence	mature_sequence
Post-meiotic Pollen	SL2.40ch12_12524	SL2.40ch12	63094877	63095144	-	-133.3	AGGTTCTAATGTCAACCATGT	ATTGTTGACATAAGTACCTGT
Post-meiotic Pollen	SL2.40ch09_6937	SL2.40ch09	51705112	51705381	+	-83.73	GTTGCCCTGCAGGGAGATGAA	CATCTCCCTACAAGGCAAGTA
Post-meiotic Pollen	SL2.40ch00_1102	SL2.40ch00	12537499	12537768	+	-81	GGTTCAATAAAGCTGTGGGAA	TCCACAGCTTTCTTGAACTGC
Post-meiotic Pollen	SL2.40ch00_1141	SL2.40ch00	12631154	12631423	+	-56.99	CTTTGTGACACTAGTTTGAAAAAA	CTCACAAGATAGTGTCACGTAGAC
Post-meiotic Pollen	SL2.40ch00_141	SL2.40ch00	2016837	2017106	+	-53.62	CTTTTTGAGGATTTTTGAGATTC	TTCTCAAACAATTTTCAATTTTAC
Post-meiotic Pollen	SL2.40ch00_3421	SL2.40ch00	20038625	20038892	-	-52.12	TGTCTACAAAGTCCTTATTTT	CATGACAGCTTTGACATGACGACG
Post-meiotic Pollen	SL2.40ch00_3791	SL2.40ch00	18265604	18265871	-	-74.37	TTGTGAAAGTTGGAGGTCAAAGT	TTATGCTCTTAAACTTTGGATGTG
Post-meiotic Pollen	SL2.40ch00_4396	SL2.40ch00	15325404	15325671	-	-51.9	GAAGGTTCAATTGGCGTTTCTATA	TTTAATAATGCCCGAACTCTTTC
Post-meiotic Pollen	SL2.40ch00_5330	SL2.40ch00	8770264	8770531	-	-55.2	ATCTCGTTTTGAGAATCAAGATA	TAACACGTTATCAACACGAGACTC
Post-meiotic Pollen	SL2.40ch00_5585	SL2.40ch00	6581480	6581747	-	-129.67	AAAATAAGTTCAGGGGGGTAA	ACCCCTCTGAACTTATTTCAT
Post-meiotic Pollen	SL2.40ch01_1009	SL2.40ch01	2575786	2576055	+	-140.1	GTCCTAAAATACTCTAATTCAAAC	TTGAATTAGAACATTTTAGGACTA
Post-meiotic Pollen	SL2.40ch01_10390	SL2.40ch01	70879593	70879862	+	-79.84	ACGTTTGTGCGTGAATCTAAC	TAGATTCACGCACAAGCTCGT
Post-meiotic Pollen	SL2.40ch01_10590	SL2.40ch01	71389466	71389735	+	-47.3	AACTCAATTATATATGATCTC	GATTTCGGGTATAGATTAAGGAGG
Post-meiotic Pollen	SL2.40ch01_1088	SL2.40ch01	2722512	2722781	+	-72	ATTCAGGGCTATCGATA	TCGATACGCACCTGAATCT
Post-meiotic Pollen	SL2.40ch01_11375	SL2.40ch01	73665788	73666057	+	-90.5	AAACACTAGTATATTGTGTTTTTT	AAAACACAATATACTAGTGATTTC
Post-meiotic Pollen	SL2.40ch01_118	SL2.40ch01	331164	331433	+	-83.3	CTGAAATTCCAAAAACACACCTTA	AGATGTGTCTCTGAGATTTCAATT
Post-meiotic Pollen	SL2.40ch01_11901	SL2.40ch01	74989995	74990091	+	-19.32	ACTATTATTGGACATCTGAAA	AGAGATGTGTCAAGTCAATAGTGA
Post-meiotic Pollen	SL2.40ch01_12400	SL2.40ch01	76334332	76334601	+	-52.03	GCATGTCAGTGACTATGT	TTGGTTACTGATGGCTA
Post-meiotic Pollen	SL2.40ch01_1245	SL2.40ch01	3116729	3116998	+	-88.3	ATATGGAAGAGGTGATTGGAG	CCAGTCACCTCATCCGTATTT
Post-meiotic Pollen	SL2.40ch01_1253	SL2.40ch01	3120713	3120982	+	-57.09	ATTAGGTGAATATGCTAAGGAGATGG	ATCTTCTCATCATAAGCATCTTTT
Post-meiotic Pollen	SL2.40ch01_1290	SL2.40ch01	3201917	3202186	+	-51.92	GTGCTCCTCATAAGACTTGTTTA	GATTTTGAAGTGTGACGTAGACTT
Post-meiotic Pollen	SL2.40ch01_12978	SL2.40ch01	77887671	77887940	+	-83.21	CACTATTAGATATGATCTGAT	CAGATCATATCTAACAGTGGA
Post-meiotic Pollen	SL2.40ch01_14224	SL2.40ch01	81651855	81652124	+	-88.9	ATTTGATGCTAAGGGCTTAAG	TATGCCCTTACCGTCAAATAC
Post-meiotic Pollen	SL2.40ch01_14243	SL2.40ch01	81730157	81730426	+	-47.6	GTTTTCTATAATCACAAAATGAG	CAAGATTGTGAATATAAATTT
Post-meiotic Pollen	SL2.40ch01_14464	SL2.40ch01	82368194	82368319	+	-63.2	CTTCCAAAGCTGCAGAAATGA	ATTTCTGCAGCTTTGGAATTT
Post-meiotic Pollen	SL2.40ch01_14679	SL2.40ch01	82968318	82968587	+	-62.38	ACAAACATTAATTTTAAAAGTAACGA	GTTCTCCAACTTTGAGTGTGT
Post-meiotic Pollen	SL2.40ch01_15139	SL2.40ch01	84522965	84523234	+	-83.8	GAGGGGGCCAAAGTGCCAAA	TGGCATTCTGTCCACCTCCC
Post-meiotic Pollen	SL2.40ch01_15939	SL2.40ch01	87025279	87025450	+	-116.5	TTGAAGTTGGCACCTTGTCTGAT	CAGACTGTGCCAACTTCAAAT

Figure 72: snapshot of Novel miRNA page from the platform

As for the known miRNAs, this result section can be also sorted by each column allowing to search for similar sequences or the miRNAs in the neighboring genomic regions.



Sections

Home | known miRNAs | Novel miRNAs | GO Annotation | Citation

By GO Term ID: Filter
By GO Keyword: Filter
By gene functional keyword: Filter
Remove filters

ID	miRNA	GeneID	Dsc.	Norm. CM	normalized CPM	normalized CT	normalized HSM	normalized HSPM	normalized HST	GO ID	GO DESC
Deatils SL2.40ch00_1102	TCCACAGCTTTCTTGAACTGC	Solyc09g009180.2.1	Chaperone protein dnaJ (AHRD V1 *-*- C8ZVX3_ENTGA)%3B contains Interpro domain(s) IPR003095 Heat shock protein DnaJ	334.41823968	374.70231294	351.34384501	312.74926741	406.6009743	409.12872358	GO:0031072;GO:0006457	heat shock protein binding;protein folding
Deatils SL2.40ch00_1141	CTCACAAGATAGTGTCACGTAGAC	Solyc08g006500.2.1	Glutamate-gated kainate-type ion channel receptor subunit GluR5 (AHRD V1 **** B9HB97_POPTR)%3B contains Interpro domain(s) IPR017103 Ionotropic glutamate-like receptor%2C plant	7.6708027042	9.6706104966	19.589773131	6.3253416465	8.463014427	22.375311174	GO:0005217;GO:0005515	intracellular ligand-gated ion channel activity;protein binding
Deatils SL2.40ch00_1141	CTCACAAGATAGTGTCACGTAGAC	Solyc03g096990.2.1	Cytochrome P450 89A2 (AHRD V1 ***- C89A2_ARATH)%3B contains Interpro domain(s) IPR002401 Cytochrome P450%2C E-class%2C group I	27.681631464	71.146062552	109.17691423	23.685643761	93.319950014	99.9867642	GO:0019825	oxygen binding
Deatils SL2.40ch00_1141	CTCACAAGATAGTGTCACGTAGAC	Solyc12g006400.1.1	Unknown Protein (AHRD V1)	641.49993731	282.58223183	12.453747742	975.67359443	130.37766264	3.2677048735	NA	NA
Deatils SL2.40ch00_1141	CTCACAAGATAGTGTCACGTAGAC	Solyc07g015860.2.1	Peptide deformylase (AHRD V1 **** B9GKW9_POPTR)%3B contains Interpro domain(s) IPR000181 Formylmethionine deformylase	177.83638031	175.81728602	102.14935445	204.10335766	190.57545371	102.99860449	GO:0042586	peptide deformylase activity
Deatils SL2.40ch00_1141	CTCACAAGATAGTGTCACGTAGAC	Solyc01g079570.2.1	Beta xylosidase (AHRD V1 ***- Q6RXY3_FRAAN)%3B contains Interpro domain(s) IPR001764 Glycoside hydrolase%2C family 3%2C N-terminal	40.020533882	24.065585099	26.902929686	15.687636637	13.177675613	35.413725166	GO:0005975	carbohydrate metabolic process
Deatils SL2.40ch00_1141	CTCACAAGATAGTGTCACGTAGAC	Solyc11g070090.1.1	Carboxyl-terminal peptidase (AHRD V1 ***- B6U8T6_MAIZE)%3B contains Interpro domain(s) IPR004314 Protein of unknown function DUF239%2C plant	5.9687356863	3.0737310613	247.45182621	4.7721964543	8.2851549968	308.80952023	NA	NA
			1-aminocyclopropane-1-carboxylate								

Figure 73: snapshot of the GO Enrichment view of the genes associated or the miRNAs with the possibility of querying by gene ID, GO keyword or functional keyword.

To better understand the role and functionality of the detected miRNAs, the target genes (genes overlapping with the miRNA mapping on the genomic locus), their functional annotation and GO description, and the normalized value of their abundance in each of the stages (see 2.2) are presented in the ”*GO Annotation*” section of the database (Figure 73).

7 ANNEX III

RNA Isolation

RNA was isolated from pollen in two fractions (*Small-RNA* < 200 nt and large RNA > 200 nt) according to manufacturer's protocol.

MACE Library Preparation

MACE libraries were prepared as previous established protocol established by *GenXPro GmbH (Frankfurt, Germany)*. Briefly, the large RNA fractions (>200nt) were reverse transcribed with *SuperScript Double-Stranded cDNA Synthesis Kit (Life Technologies)* using *biotinylated poly (dT) primers*. *cDNA* was fragmented with *Bioruptor (Diagenode)* to an average size of 250 bp. Biotinylated c*DNA* ends were captured by *Dynabeads M-270 Streptavidin Beads (Life Technologies)* and ligated with *T4* DNA *Ligase 1 (NEB)* to modified adapters *(TrueQuant, GenXPro)*. The libraries were amplified by PCR with *KAPA HiFi Hot-Start Polymerase (KAPA Biosystems)*, purified by *Agencourt AMPure XP beads (Beckman Coulter)* and sequenced with *HiSeq2000 (Illumina)*.

DNA Isolation and Meth-Seq Library Preparation

DNA was isolated from pollen using the *DNeasy Blood & Tissue Kit (Qiagen, Hilden, Germany)*. Genome-wide analysis of DNA methylation was performed by *MethSeq* at *GenXPro GmbH (Frankfurt, Germany)*. *HpaII* was used as the methylation-sensitive enzyme, recognizing non-CpG-methylated *CCGG* sites. After digestion by *HpaII*, the DNA fragments were ligated to *Illumina's p5* primer for sequencing (*Illumina*).

sncRNA Sequencing Library Preparation

For preparation of *Small-RNA* libraries, *5 µg* RNA (*Small-RNA* fraction) was size-selected (<40 nt) by *polyacrylamide gel electrophoresis (FlashPAGE, Life Technologies)* and precipitated. About 30 ng *Small-RNA* (<40 nt) was successive ligated (*T4* RNA *Ligase 1 and T4* RNA *Ligase 2, NEB*) to modified 3′ and 5′ adapters (*TrueQuant* RNA adapters, *GenXPro*). Adapter-ligated RNA was reverse transcribed (SuperScript III, Life Technologies) and amplified by PCR (*KAPA HiFi Hot-Start Polymerase, KAPA Biosystems*). Amplified libraries were size-selected by *polyacrylamide gel electrophoresis (PAGE)* and sequenced (*HiSeq2000, Illumina*).

8 Publications, presentations and collaborations

8.1 Publications

- **Bostan, H.** and Chiusano, M. L. (2015). NexGenEx-Tom: a gene expression platform to investigate the functionalities of the tomato genome. BMC plant biology, 15(1), 48.
- Bokszczanin K., **Bostan H.**, Bovy A., Chaturvedi P., Chiusano M.L., Firon N., Innacone R., Jegadeesan S., Klaczynski K., Li H., Marina C., Muller F., Paul P., Paupiere M., Pressman E., Rieu I., Scharf K., Schleiff E., Heusden A.W., Virezen W., Weckwerth W., Winter P., And Fragkostefanakis S., "Perspectives on deciphering mechanisms underlying plant heat stress response and thermotolerance", Frontiers in plant science 4 (2013).
- Bokszczanin K., Krezdorn N., Fragkostefanakis S., Müller, S., Rycak, L., Chen Y.Y., Hoffmeier K., Kreutz J., Paupière M., Chaturvedi P., Iannacone R., Müller F., **Bostan H.**, Chiusano M.L., Scharf K.D., Rotter B., Schleiff E. and Winter P. (2015). Identification of novel small ncRNAs in pollen of tomato. BMC Genomics.
- **Bostan H.** †, Ambrosino L. †, Di Salle P., Mara Sangiovanni2, Vigilante A. and Chiusano M.L. (2015). pATsi: paralogs and singleton genes from Arabidopsis thaliana. BMC Bioinformatics (**Accepted**).

8.2 Manuscripts under review

- Ruggieri V., **Bostan H.,** Barone A., Frusciante L. and Chiusano M.L. (2015). Integrated Bioinformatics: To Decipher the Ascorbic Acid Metabolic Network in Tomato. Plant Molecular Biology.

8.3 Manuscripts in preparation

- Revised Annotation: A guidance for the Tomato *Solanum lycopersicum* cv. *Heinz* Gene Annotation,
- New Tomato *Solanum lycopersicum* cv. *Heinz* gene Annotation: Flourishing of the genome using all the official gene annotations, validating and cross comparing the elements structure,
- The Role of TE-Derived Small Interfering RNAs in Tomato Pollen Development,
- *Exonate* Pipeline: An annotation independent pipeline creating customized reference annotations (replicate, tissue/stage or condition based) for Differentially Expression and Methylation analyses,
- Exploring the Tomato based on integrated data sources: the gene hunting season is open,
- *NexGenEx-Pot:* a gene expression platform to investigate the functionalities of the potato genome,
- Tomato Housekeeping genes revised: An NGS Methodology for Identification of housekeeping genes in tomato,
- *SPOT-ITN* Bioinformatics Platform: A reference platform for tomato pollen thermos-tolerance (note: platform design finished, pending for consortium decision: http://www.unina.it/spot-itn-bioinfo).

8.4 Presentations and Conferences

- Bostan, H. and Chiusano, M.L., 2015. "Reconciliation and Integration: an essential step towards the modelling of biological systems starting from omics data", Sorrento, Italy, 18-22 March (Oral presentation).
- Bostan, H. and Chiusano, M.L., 2014. "SPOT-ITN Data Sharing and Bioinformatics Platform". Goethe University of Frankfurt, Germany 8th December (Oral presentation).
- Bostan, H. and Chiusano, M.L., 2014. "A tutorial to the SPOT-ITN Data Sharing and Bioinformatics Platform". University of Vienna, Austria, 4th November (Oral presentation).
- Bostan, H., Ambrosino, L., Ruggieri, V., Chiusano, M.L., 2014. "Characterization of Derivative Relationship between Tomato and Grapevine: A Key Step to Investigate Fruit Development in the Two Species". 3rd Annual Conference of the COST ACTION FA1106 on Fleshy fruit research, Chania, Crete, 21-24 September (Oral presentation).
- Ruggieri, V., Bostan, H., Barone, A., Frusciante, L., Chiusano, M.L., 2014. "Integrating omics For Tomato Ascorbic Acid Pathway". Proceedings Of the 58th Italian Society of Agricultural Genetics Annual Congress Alghero, Italy – 15-18 September, 2014 ISBN 978-88-904570-4-3 (poster).
- Bostan, H., Colontuono, C., Chiusano, M.L., 2014. "Tomato Genome Annotation: Genome peculiarities or miss-annotation". BITs Annual Meeting, Rome, Italy, 23 February (Poster).
- Bostan, H. and Chiusano, M.L., 2013. "Development of a bioinformatics platform for gene expression analysis in tomato: A first step to investigate pollen peculiarities". 2nd SPOT-ITN conference, Arnhem, Netherlands, 2nd November (oral presentation and poster).

- Ruggieri, V., Bostan, H., Chiusano, M.L., 2013. "Integrated Bioinformatics: A key step towards the annotation of metabolic pathways. An example for ascorbic acid in tomato". COST ACTION FA1106 Quality Fruit, Crete, Greece (Poster).
- Bostan, H. and Chiusano, M.L., 2013. "Development of a bioinformatics platform for gene expression analysis in tomato: A first step to investigate pollen peculiarities". Computational Biology and Bioinformatics, Avelino, Italy (Oral presentation).

8.5 Collaborations

- *De novo* Genome Assembly, Functional and Expression analyses of a Fungi in University of Naples "Federico II", Naples, Italy,
- *De novo* Genome Assembly and characterization of the identified transcripts in the collection for a group at stazione zoological, Naples, Italy,
- A deep characterization of the response to water stress and rehydration in tomato (External collaboration with the group of Dr. Grilo, university of Naples "Federico II", Italy, co-author)
- Expression analyses in tomato *San marzano* in collaboration with the group of Prof. Rosa Rao, university of Naples "Federico II", Italy,
- mRNA and *Small-RNA* Expression analyses in *Red setter* in collaboration with the group of Prof. Rosa Rao, university of Naples "Federico II", Italy,
- *Genopom* Bioinformatics Platform (I was in charge of setting up the platform for the whole project, re analyses the data with a common procedure and organize the data in the platform,

- *Epitom* Bioinformatics Platform (I was in charge of setting up the platform for the whole project, re analyses the data with a common procedure and organize the data in the platform)