CHAPTER-II

REVIEW OF LITERATURE

The literature pertaining to the "Genome and transcriptome sequencing of coriander (Coriandrum sativum L.) to reveal its genome architecture". is presented in this chapter. A brief review of research work related to the different aspects of coriander have been reviewed and highlighted under the following headings. To fill up the gap, wherever it was necessary, the review on other crop is also included.

2.1. NOMENCLATURE

Coriander (Coriandrum sativum L.) is an annual herbaceous plant belonging to the highly valued spice and medicinally important family, Apiaceae.

2.1.1 Scientific classification

Scientific classification of Coriander with its order. Family, Genus and botanical name is given as below (Table 2.1).

<table>
<thead>
<tr>
<th>Crop</th>
<th>Coriander</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kingdom</td>
<td>Plantae</td>
</tr>
<tr>
<td>Order</td>
<td>Apiales</td>
</tr>
<tr>
<td>Family</td>
<td>Apiaceae</td>
</tr>
<tr>
<td>Genus</td>
<td>Coriandrum</td>
</tr>
<tr>
<td>Binominal name</td>
<td>Coriandrum sativum L.</td>
</tr>
</tbody>
</table>

2.2. GEOGRAPHIC ORIGIN

Coriander is a native to the Mediterranean and Middle Eastern region. In India, it is cultivated in Madhya Pradesh, Gujarat, Rajasthan, Maharashtra, Karnataka and West Bengal (Figure 2.1).
Looking to the importance of coriander as a medicinal and spices crop, this research has been undermined to discover the genes responsible for different cellular, biological and molecular aspects through genome sequencing.

**Threats in coriander:**

Coriander is an important spice crop grown in many countries including India which has a major share. The crop is attacked by several pathogens causing diseases like vascular wilt, stem rot, root rot, charcoal rot, seedling blight, stem gall, leaf blight, anthracnose, powdery mildew, seed rot, grain mould, bacterial blight, soft rot, seedling root rot, reniform and root knot nematodes, phyllody and virus diseases. Some new diseases have appeared in some countries. Genome sequencing is one of the options to solve these problems in coriander. Resistant varieties are the best to control the diseases.
2.3 Genome sequencing

DNA sequencing is the act of determining the nucleotide sequence of given DNA molecules — from a short segment of a single molecule, such as a regulatory region or a gene, up to collections of entire genomes. In the early 1970s, the first DNA sequences were obtained through extremely laborious techniques. An example is the sequencing of the two dozen base pairs of the lac operator (Gilbert and Maxam 1973). The first revolution in the DNA sequencing field took place in the second half of the 1970s with methods published by Allan Maxam and Walter Gilbert (Maxam and Gilbert 1977) and Frederick Sanger and colleagues (Sanger et al. 1977). Both techniques greatly increased the throughput of sequencing DNA. The method from Gilbert and Maxam, however, was more complex and involved the use of hazardous chemicals. The Sanger method, on the other hand, offered overall higher efficiency after a series of optimisations, in particular switching from radioactive to dye labelling of nucleotides and using capillary electrophoresis instead of slab gels.

2.3.1 DNA Sequence in Plant System

Although nucleic acid sequencing is a relatively new approach in plant systematic, the power of the technique and the data generated have made it become one of the most utilized of the molecular approaches for inferring phylogenetic history. DNA sequence data are the most informative tool for molecular systematic, and comparative analysis of DNA sequences is becoming increasingly important in plant systematic. There are two major reasons why nucleotide sequencing is becoming so valuable in plant systematic: 1) the characters (nucleotides) are the basic units of information encoded in organisms; 2) the potential sizes of informative data sets are immense. For example, one in 100 nucleotides is polymorphic in the human genome so that there will be about 2 x 10. Furthermore, different genes or parts of the genome might evolve at different rates.

2.3.2 Ion Torrent semiconductor sequencing

Ion Torrent Systems Inc. developed a system based on using standard sequencing chemistry, but with a novel, semiconductor based detection system. This method of sequencing is based on the detection of hydrogen ions that are released during the polymerisation of DNA, as opposed to the optical methods used in other sequencing systems. A microwell containing a template DNA strand to be sequenced is flooded with a single type of nucleotide. If the introduced nucleotide
is complementary to the leading template nucleotide it is incorporated into the growing complementary strand. This causes the release of a hydrogen ion that triggers a hypersensitive ion sensor, which indicates that a reaction has occurred.

2.3.3 Ion Torrent

This newer platform utilizes the adaptor-ligated library followed by sequencing-by-synthesis chemistry of other platforms, but has a unique feature. Instead of detecting fluorescent signals or photons, it detects changes in the pH of the solution in a well when a nucleotide is added and protons are produced. These changes are miniscule, however the Ion Torrent device utilizes technologies developed in the semiconductor industry to achieve detectors of sufficient sensitivity and scales that are useful for nucleic acid sequencing. One limitation that has been pointed out is that homopolymers may be difficult to read as there is no way to stop the addition of only one nucleotide if the same nucleotide is next in the sequence. Ion Torrent can detect a larger change in the pH and uses this measurement to read through polymer regions.

This platform produces overall fewer reads than the others in a single run. For example, 60–80 M reads at 200 bases per read are possible on the Proton instrument in a run producing 10 Gb of data. However, the run time is only 2–4 h instead of 1–2 weeks on other platforms. Since neither modified nucleotides nor optical measurement instrumentation is needed, an advantage of this platform is affordability, both of the instrument and reagents. The machine has a small footprint, can be powered down when not in use and easily brought back to use, and requires minimal maintenance.

2.3.4 Impact of genome sequencing:

The possibility of sequencing large amounts of DNA at considerably low cost has been a tremendous boon to genomics and has lead to the advancement and establishment of several sub-disciplines. The main driver for cheap genome-scale sequencing has been the medical sector. The compilation of detailed catalogues of alleles underlying human diseases will help to dissect their mechanisms and will lead to better treatment, particularly making possible personalised medicine (Hingorani et al. 2010). Traditionally, variant discovery was undertaken by blanket sequencing of a small pool of individuals. These variants were then used to construct assay chips which made economically feasible the genotyping of thousands of additional
individuals. Such variation data were successfully employed in several large-scale case-control studies, e.g. (Wellcome Trust Case Control Consortium 2007, 2010; Barrett et al. 2008). This strategy, however, can only capture common variants and additionally suffers from sampling bias. Crucial for the success of this approach is that untyped causal alleles are in linkage disequilibrium (LD) with typed alleles.

Besides its implication in disease association, sequencing of entire genomes of individuals is also greatly benefiting population genetics studies. Data from the 1000 Genomes project have already been used to infer mutation rates (Conrad et al. 2011) and coalescent times (Li and Durbin 2011) in humans.

2.3.5 Sequencing application

NGS technologies have contributed a series of genetic improvements on plant breeding and biotechnology. In contrast to the first-generation sequencing, the second- and third-generation technologies produce an enormous volume of sequence data at a much lower cost, making the system versatile for plenty of applications (Metzker, 2009; Llaca, 2012). Today, the second generation sequencing is extensively used in discovery of genetic markers, gene expression profiling through mRNA sequencing and comparative and evolutionary studies to answer a diverse set of biological questions (Wang et al., 2009; Jia et al., 2013; Nystedt et al., 2013; Dohm et al., 2014; Sierro et al., 2014). Even more promising for the near future is the third generation sequencing, being mostly in active development nowadays.

2.4 Genome size estimation and DNA isolation

Nath et al. (2014) improved method of genome size estimation by flow cytometry in five mucilaginous species of hyacinthaceae. They report on genome size estimation by flow cytometric analysis of hyacinthaceae family plants are rare due to the presence of mucilage, which causes problem with nuclei sample preparation. To overcome this problem five reported nuclei isolation buffers were tested in Drimia indica of which Galbraith’s buffer gave comparatively better results and was further modified by increasing pH, detergent concentration, and replacing sodium citrate by citric acid. The modified buffer enabled better sample preparation with increased yield, lesser debris, and improved DNA peak CV.

Ved et al. (2013) performed genetic diversity in capsicum germplasm based on microsatellite and random amplified microsatellite polymorphism markers DNA isolation and quantification of leaf samples from five random plants of each genotype were collected and total genomic DNA was extracted following Doyle and Doyle
(1987) with some modifications. Total genomic DNA was quantified using nanodrop. Genomic simple sequence repeats (SSRs) and random amplified microsatellite polymorphism (RAMPO) markers were used to analyse diversity and relationships among 48 pepper (Capsicum spp.) genotypes originating from nine countries.

For assessment of genetic variability for three coriander (Coriandrum sativum L.), El-Nasr et al. (2013) isolated DNA from young and fresh leaves samples were collected separately from each cultivar.

2.5 Whole genome sequencing

The broadest application of the new sequencing approaches to plant species may be whole genome sequencing (WGS) to reveal the full sequence and genetic structure of genomes. In WGS projects such as strawberry (Shulaev et al. 2011) and wheat (Brenchley et al. 2012) whole genomic DNA content was first randomly cut into fragments of different sizes. Then, BAC-end sequencing was carried out and the obtained reads were assembled using powerful bioinformatics tools. The WGS approach can be accomplished not only for resequencing, but also for De novo projects.

Although it takes more time, the De novo sequencing of whole DNA or mRNA is useful for producing draft genomes when the plant genome of interest is unknown. For instance, draft genomes of several crop species such as einkorn (Ling et al., 2013), as well as wheat and A.tauschii (Jia et al. 2013) were produced using the WGS approach. Apart from this, resequencing is mostly used in transcriptome profiling and SNP discovery for marker development (Llaca, 2012). Thus, a high-quality reference genome of potato was revealed utilizing WGS approach and SNP identification was performed to compare a homozygous doubled-monoploid line with its heterozygous diploid line (Xu et al. 2011). More recently, several accessions of watermelon were resequenced and compared with each other. Thus, a total of 6,784,860 SNP were identified, representing the genetic diversity of the crop species (Guo et al. 2013).

2.5.1 Whole genome sequenced of plants

Genome of the pitcher plant Cephalotus reveals genetic changes associated with carnivory (Fukushima et al. 2017). Carnivorous plants exploit animals as a nutritional source and have inspired long-standing questions about the origin and evolution of carnivory-related traits. To investigate the molecular bases of carnivory, they sequenced the genome of the heterophyllous pitcher plant Cephalotus
follicularis, in which we succeeded in regulating the developmental switch between carnivorous and non-carnivorous leaves. Transcriptome comparison of the two leaf types and gene repertoire analysis identified genetic changes associated with prey attraction, capture, digestion and nutrient absorption. Analysis of digestive fluid proteins from C. follicularis and three other carnivorous plants with independent carnivorous origins revealed repeated co-options of stress-responsive protein lineages coupled with convergent amino acid substitutions to acquire digestive physiology. These results imply constraints on the available routes to evolve plant carnivory.

Genome sizes of plants widely varies and covers at least in four orders of magnitude ranging from 125 Mbp (Arabidopsis), 430 Mbp (Rice), 2670 Mbp (Maize), 10,000 Mbp (Wheat), 125, 000 Mbp (Fritillaria assyriaca), 250, 000 Mbp (Pislotum nudum) (Obermayer et al. 2002). An increase in plant genome size has been driven in partly by polyploidy and partly by integration and multiplication of retrotransposon repeat sequences. Retrotransposon sequences are problematic within the context of complete genome sequencing and thousands of these large repeat sequences yield no information on gene content and generate statistical issues that prevent the assembly of individual sequence reads into meaningful contigs in whole genome assembly (Hoskins et al. 2002; Kent and Haussler, 2001).

The large plant genomes, including those of barley, maize, wheat and pearl millet are mostly comprised of 5-20 kb blocks of genes intermixed with repetitive DNA blocks that range in size from a few kilo bases up to more than 100 (SanMiguel et al., 1996; Panstruga et al., 1998; Tikhonov et al., 1999; Dubcovsky et al., 2005). The repetitive DNAs found in the intermixed repeat blocks are usually nested insertions of a class of mobile DNAs called long terminal repeat- (LTR-) retrotransposons (Panstruga et al. 1998; Kumar and Bennetzen 1999). These nested LTR retrotransposons can make up well over 50% of total genomic DNA, with most of this DNA belonging to only a small number of different element families that have copy numbers of several thousand per nuclear genome (Kumar and Bennetzen, 1999). The LTR-retrotransposons are relatively large in size (usually greater than 5 kb) and can have numerous copies that are <99% identical within the same genome. Hence, shotgun sequencing of complex grass genomes would not yield significant useful information that can be converted into long contiguous sequences (Bennetzen et al. 2005).
In order to understand the biology of *R. communis*, the genome by whole genome shotgun sequencing of its ~400 Mbp genome, were draft assembled which shows approximately 4X sequence coverage of the genome. In addition to 40,000 ESTs from different tissues that was essential for gene discovery and annotation. One of the immediate uses of the sequence was to develop markers for genotyping (Agnes *et al.*, 2006).

The castor bean genome, which is distributed across ten chromosomes, is estimated by flow cytometry to be ~320 Mb. Agnes *et al.* (2010) set out to generate a draft sequence of the castor bean genome by producing ~2.1 million high-quality sequence reads from plasmid and fosmid libraries (Online Methods), and then using the Celera assembler to build consensus sequences or contigs that were linked to form 25,800 scaffolds using the two end-sequences from individual clones (mate-paired reads). The assembly covered the genome ~4.6X coverage, spanning 350 Mb, which is consistent with previous genome size estimations. In this study, only the 3,500 scaffolds larger than 2 kb are considered.

Eman *et al.* (2011) assembled a draft genome for a Khalas variety female date palm. The ~380 Mb sequence, spanning mainly gene-rich regions, includes >25,000 gene models and is predicted to cover ~90% of genes and ~60% of the genome. Sequencing of eight other cultivars including females of the Deglet Noor and Medjool varieties and their backcrossed males. The identified >3.5 million polymorphic sites including >10,000 genic copy number variations. A small subset of these polymorphisms can distinguish multiple varieties. They identified a region of the genome linked to gender and found evidence that date palm employs an XY system of gender inheritance.

Gengyun *et al.* (2012) studied the draft genome sequence of foxtail millet, using shotgun strategy combined with next-generation sequencing technology. Foxtail millet has a small genome (~490 Mb) anchored onto nine chromosomes and annotated 38,801 genes, which is rich in repeats. The genome assembly covered ~86% of the estimated genome size, and the unassembled part largely composed of repeat elements.
The papaya genome was sequenced using a whole-genome shotgun approach by the traditional Sanger method to approximately 3x coverage. Cytogenetic studies suggest that the genome is about 65-70% euchromatic and 35-30% heterochromatic. Various measures were used to assess the coverage of the draft genome, such as the percentage of unique genes (unigenes) and genetic markers matching the assembly. Fusheng et al. (2008) estimate that approximately 90% of the euchromatin covered, 92.1% of the unigenes and 92.4% of the genetic markers. Automated annotation of the genome combined with the genome coverage led the gene content of 24,746 genes. Compared with the other four sequenced plant genomes, this gene count is 11-20% less than that of Arabidopsis.

Xu et al. (2014 b) performed first whole-genome sequences of DC-27+ carrot were De novo assembled and analyzed by using downloaded transcriptomic sequences of 14 carrot genotypes from the Sequence Read Archive (SRA) database of NCBI. Before assembly they mapped to the whole-genome sequence. Based on these data sets, they developed the first Web-based genomic and transcriptomic database for Daucus carota (CarrotDB). CarrotDB offers the tools of Genome Map and Basic Local Alignment Search Tool. Users can download De novo assembled whole-genome sequences, putative gene sequences and putative protein sequences of DC-27. A total of 2826 transcription factor (TF) genes classified into 57 families were identified in the entire genome sequences.

Wang et al. (2014) studied De novo assembly of sesame. After reads filtering, 54.5 Gb of high-quality data from sesame genotype ‘Zhongzhi No. 13’ were obtained using the Illumina Hiseq2000 platform, representing approximately 152.7-fold coverage of the predicted sesame genome. SOAP denovo was used to assemble the genome, which resulted in a draft genome of 274 Mb with contig N50 of 52.2 kb and scaffold N50 of 2.1 Mb, which are approximately 2.7 and 92.9 fold longer, respectively. They have anchored 150 large scaffolds (117 oriented) into 16 pseudomolecules, which harbored 85.3% of the genome assembly and 91.7% of the predicted genes.

Qin et al. (2014) sequenced whole-genome of cultivated and wild peppers which provides insights into Capsicum domestication and specialization. To gain a better understanding of Capsicum evolution, domestication, and specialization, they carried out the genome sequence of the cultivated pepper Zunla-1 (C. annuum L.) and its wild progenitor Chiltepín (C. annuum var. glabriusculum). Results shown that the
pepper genome expanded ~0.3 Mya by a rapid amplification of retrotransposons elements, resulting in a genome comprised of 81% repetitive sequences. Approximately 79% of 3.48-Gb scaffolds containing 34,476 protein-coding genes were anchored to chromosomes by a high-density genetic map.

Genome sequencing of the hot pepper provides insights into the evolution of pungency in Capsicum species. Kim et al. (2014) reported whole-genome sequencing and assembly of the hot pepper (Mexican landrace of Capsicum annuum cv. CM334) at 186.6x coverage. They suggested integrative genomic and transcriptomic analyses that change in gene expression and neofunctionalization of capsaicin synthase have shaped capsaicinoid biosynthesis. They found differential molecular patterns of ripening regulators and ethylene synthesis in hot pepper and tomato.

Upadhyay et al. (2015) studied key genes behind strong medicinal properties of herb Tulsi (Ocimum tenuiflorum) by genome sequencing. The pathways leading to the production of medicinally-important specialized metabolites have been studied in detail, in relation to similar pathways in Arabidopsis thaliana and other plants. Expression levels of anthocyanin biosynthesis-related genes in leaf samples of Krishna Tulsi were observed to be relatively high, explaining the purple colouration of Krishna Tulsi leaves. The expression of six important genes identified from genome data were validated by performing q-RT-PCR in different tissues of five different species.

Whole-genome sequencing of Oryza brachyantha reveals mechanisms underlying Oryza genome evolution analyzed by Chen et al. 2013. The 261-Mb De novo assembled genome sequence of Oryza brachyantha. Low activity of long-terminal repeat retrotransposons and massive internal deletions of ancient long-terminal repeat elements lead to the compact genome of Oryza brachyantha. They model 32,038 protein-coding genes in the Oryza brachyantha genome, of which only 70% are located in collinear positions in comparison with the rice genome.
Wang et al. (2014 a) performed genome sequencing of the high oil crop sesame which provides insight into oil biosynthesis. They reported a high-quality genome sequence of sesame assembled De novo with a contig N50 of 52.2kb and a scaffold N50 of 2.1 Mb, containing an estimated 27,148 genes. Candidate genes and oil biosynthetic pathways contributing to high oil content were discovered by comparative genomic and transcriptomic analyses. Also revealed the expansion of type1 lipid transfer genes by tandem duplication, the contraction of lipid degradation genes, and the differential expression of essential genes in the triacylglycerol biosynthesis pathway, particularly in the early stage of seed development.

Analysis of the bread wheat genome using whole-genome shotgun sequencing Brenchley et al. (2012). They reported the sequencing of its large, 17-gigabase-pair, hexaploid genome using 454 pyrosequencing, and comparison of this with the sequences of diploid ancestral and progenitor genomes. They identified between 94,000 and 96,000 genes, and assigned two-thirds to the three component genomes (A, B and D) of hexaploid wheat. High-resolution synteny maps identified many small disruptions to conserved gene order. They had shown that the hexaploid genome is highly dynamic, with significant loss of gene family members on polyploidization and domestication, and an abundance of gene fragments.

2.6 Reads and coverage

Short read NGS technologies nowadays provide terabyte-sized data files, so coverage does not seem to be an issue and previously inflexible plant genomes (for example pine genomes, which are 7 to 10-fold longer than the human genome and probably contain >95% repetitive sequences) are now feasible, at least in theory. Variation in coverage is introduced by chance by variation of the copy number within DNA (i.e., repeats) and by the technology. But when coverage is homogeneous along the genome, local biases can be interpreted as follows: Gaps are a consequence of very low coverage and high-coverage is a diagnosis of an over collapsed repeat. Unfortunately, coverage variability is the rule and undermines the coverage based diagnostics. It can be speculated that the sequencing itself needs to be improved to reduce the biases, for example from GC composition and PCR, so that the coverage along the genome will be uniform and complete (Aird et al., 2011).

The overwhelming throughput of NGS raises a collateral issue related to data overload on a laboratory, institutional and community scale. In fact, the infrastructure costs for data storage, processing and handling are becoming more worrying than the
costs of generating the reads. Since sequencing throughput is expected to increase in coming years, data storage and handling are becoming a real concern (Hamilton et al., 2012).

The comparison of each read with others required by the overlap-layout-consensus algorithms as well as the resolution of the Eulerian paths for de Bruijn graphs are the most time-consuming part of the assembling process. Therefore, the task could become never-ending or result in a faulty execution when temporary data do not fit in available RAM. The situation could arise that the right data and the right algorithm are available, but the right computer or software to hold calculations and memory are not. The most recent assemblers are focused on distributing among CPUs the processing load that cannot be managed with current serial algorithms. The de Bruijn graphs methods for assembly have the advantage of avoiding the all comparisons, but their use is limited when there are too many errors or there is too low coverage, since they lead to infinite loops in the Eulerian paths that produce erroneous superstrings (Li et al., 2012).

Genome assembly is a challenging problem that requires time, resources and expertise. Before engaging in a genome sequencing project, it should thus be carefully considered whether a genome reference sequence is strictly necessary for the purpose in question. Genome sequences are merely a resource and in many cases will contribute very little per se to a problem in conservation biology. In case a genome draft is judged to be of significant value to address the problem at hand, it needs to be considered whether sufficient financial and computational resources are available to produce a genome of satisfactory quality. The De novo assembly of high-throughput sequencing reads into high quality reference sequences will increase our knowledge of important organisms and yield important advantages in many genome analysis tasks. The number of De novo short read genome assembly tools has been increasing steadily. Assemblies of important model organisms are subject to continuing finishing processes.

2.7 Transcriptome

RNA-seq, also called whole-transcriptome shotgun sequencing, refers to the use of high-throughput sequencing technologies for characterizing the RNA content and composition of a given sample. The great advantage of RNA-seq data over other next generation-sequencing applications is that it allows users to investigate differences in gene expression patterns between populations, for example in the context of speciation (Wolf et al. 2010) or eco-type-specific adaptation (Lenz et al. 2013).
Gene regulatory variation need not be confined to gene expression levels. Simultaneous information on sequence variation at individuals’ genomes and transcriptomes allows inferring patterns of allele-specific expression that can be relevant to environmental response and adaptation (Guo et al. 2004; Tirosh et al. 2009) and has yet to be examined in the wild. Due to technological limitations at present, sequence information from transcripts cannot be retrieved as a whole, but is randomly decomposed into short reads of up to several hundred base pairs. In the absence of genome or transcriptome information, transcripts first need to be reconstructed from these reads (or read pairs), which is referred to as De novo assembly. In the case where transcript or genome information is readily available, reads can be directly aligned onto the reference. Further, counting the reads that fall onto a given transcript provides a digital measurement of transcript abundance, which serves as the starting point for biological inference.

Galata et al. (2014) studied transcriptome profiling, and cloning and characterization of the main monoterpene synthases of coriander. To investigate EO metabolism, the transcriptome of coriander mericarps, at three developmental stages (early, mid, late) was sequenced via Illumina technology and a transcript library was produced. To validate the usability of the transcriptome sequences, two terpene synthase candidate genes, CscTRPS and CsLINS, encoding 558 and 562 amino acid proteins were expressed in bacteria, and the recombinant proteins purified by Ni–NTA affinity chromatography. The 65.16 (CscTRPS) and 65.91 (CsLINS) kDa recombinant proteins catalyzed the conversion of geranyl diphosphate, the precursor to monoterpenes, to c-terpinene and (S)-linalool.

Transcriptome analysis of Thapsia laciniata rouy (apiaceae) provides insights into terpenoid biosynthesis and diversity in apiaceae carried out by Drew et al. (2013). A transcriptomic analysis utilizing Illumina next-generation sequencing enabled the identification of novel genes involved in the biosynthesis of terpenoids in Thapsia. From 66.78 million HQ paired-end reads obtained from T. laciniata roots, 64.58 million were assembled into 76,565 contigs (N50: 1261 bp). Seventeen contigs were annotated as terpene synthases and five of these were predicted to be sesquiterpene synthases. Three contigs annotated as aldehyde dehydrogenases grouped phylogenetically with the characterized ALDH1 from Artemisia annua and three contigs annotated as alcohol dehydrogenases.
Rong et al. (2014) studied new insights into domestication of carrot from its root transcriptome. They identified 11,369 SNPs. Of these, 622 (out of 1000 tested SNPs) were validated and used to genotype a large set of cultivated carrot, wild 13 carrot and other wild Daucus carota subspecies, primarily of European origin. Genetic diversity was significantly reduced in western cultivars, probably through bottlenecks and selection. They showed that expression patterns differed radically for some genes between cultivated and wild carrot roots which may be related to changes in root traits.

Iorizzo et al. (2011) performed De novo assembly and characterization of the carrot transcriptome which reveals novel genes, new markers, and genetic diversity. A De novo assembly of transcriptome sequence from four genetic backgrounds produced 58,751 contigs and singletons. Over 50% of these assembled sequences were annotated allowing detection of transposable elements and new carrot anthocyanin genes. Presence of multiple genetic backgrounds in their assembly allowed the identification of 114 computationally polymorphic SSRs, and 20,058 SNPs at a depth of coverage of 20× or more. About 90% and 88% of tested SSR and SNP primers amplified a product, of which 70% and 46%, respectively, were of the expected size.

De novo assembly, transcriptome characterization, lignin accumulation, and anatomic characteristics: novel insights into lignin biosynthesis during celery leaf development was done by Jia et al. (2015). The transcriptome profile, lignin distribution, anatomical characteristics, and expression profile of leaves at three stages were analyzed. Celery leaves at three stages were collected, and Illumina paired-end sequencing technology was used to analyze large-scale transcriptome sequences. They Identified the genes that encode lignin biosynthesis-related enzymes accompanied by lignin distribution may help elucidate the regulatory mechanisms of lignin biosynthesis in celery.

De novo transcriptome assembly in chili pepper (Capsicum frutescens) to identify genes involved in the biosynthesis of capsaicinoids was performed by Liu et al. (2013) a. To understand more knowledge in biosynthesis of capsaicinoids, they applied RNA-seq for the mixture of placenta and pericarp of pungent pepper (Capsicum frutescens L.). They obtained a total 54,045 high-quality unigenes (transcripts) using Trinity software. About 92.65% of unigenes showed similarity to the public protein sequences, genome of potato and tomato and pepper (C. annuum) ESTs databases. Their results predicted 3 new structural genes (DHAD, TD, PAT), which filled gaps of the capsaicinoid biosynthetic pathway predicted by Mazourek.
and revealed new candidate genes involved in capsaicinoid biosynthesis based on KEGG (Kyoto Encyclopedia of Genes and Genomes) analysis.

Wang et al. (2015 b) analyzed novel insights into cytokinin accumulation during carrot growth and development for morphological characteristics, anatomical structure, and gene expression. They collected five stages of carrot plants, and morphological and anatomical characteristics and expression profiles of cytokinin-related genes were determined. During carrot growth and development, cytokinin levels were the highest at the second stage in the roots, whereas relatively stable levels were observed in the petioles and leaves. DcCYP735A2 showed high expression at stage 2 in the roots, which may contribute largely to the higher cytokinin level at this stage.

Wang et al. (2015 a) checked novel insights into gibberellin biosynthesis and perception during carrot growth and development for morphological characteristics, anatomical structure, and gene expression. They identified genes involved in gibberellin biosynthesis and signaling from the carrotDB, and their expression was analyzed. All of the genes were evidently responsive to carrot growth and development, and some of them showed tissue-specific expression. They observed that relative transcription levels of gibberellin pathway-related genes may be the main cause of the different bioactive GAs levels.

Transcript profiling of structural genes involved in cyanidin-based anthocyanin biosynthesis between purple and non-purple carrot (Daucus carota L.) cultivars reveals distinct patterns was studied by Xu et al. (2014 a). Six novel structural genes, CA4H1, CA4H2, 4CL1, 4CL2, CHI1, and F3′H1 were isolated from purple carrots. The expression profiles of these genes, together with other structural genes known to be involved in anthocyanins biosynthesis, were analyzed in three purple and six non-purple carrot cultivars. PAL3/PAL4, CA4H1, and 4CL1 expression levels were higher in purple than in non-purple carrot cultivars. Expression of CHS1, CHI1, F3H1, F3′H1, DFR1, and LDOX1/LDOX2 was highly correlated with the presence of anthocyanin as these genes were highly expressed in purple carrot taproots but not or scarcely expressed in non-purple carrot taproots.

For in silico discovery of SNPs, SSRs and candidate genes Ashrafi et al. (2012) performed De novo assembly of the pepper transcriptome (Capsicum annuum). A total of 2,489 simple sequence repeats (SSRs) were identified from the assembly, and primers were designed for the SSRs. Annotation of contigs using 15 Blast2GO
software resulted in information for 60% of the unigenes in the assembly. The second transcriptome assembly was constructed from more than 200 million Illumina Genome Analyzer II reads (80–120 nt) using a combination of Velvet, CLC workbench and CAP3 software packages. Using the MISA software, 10,398 SSR markers were also identified within the Illumina transcriptome assembly and primers were designed for the identified markers. The assembly was annotated by Blast2GO and 14,740 (12%) of annotated contigs were associated with functional proteins.

Hu et al. (2015) performed De novo assembly and characterization of fruit transcriptome in black pepper (Piper nigrum). They describe the fruit transcriptome of black pepper by sequencing on Illumina HiSeq 2000 platform. A total of 56,281,710 raw reads were obtained and assembled. From these raw reads, 44,061 unigenes with an average length of 1,345 nt were generated. During functional annotation, 40,537 unigenes were annotated in Gene Ontology categories, Kyoto Encyclopedia of Genes and Genomes pathways, Swiss-Prot database, and Nucleotide Collection (NR/NT) database. In addition, 8,196 simple sequence repeats (SSRs) were detected. In a detailed analysis of the transcriptome, housekeeping genes for quantitative polymerase chain reaction internal control, polymorphic SSRs, and lysine/ornithine metabolism-related genes were identified.

Transcriptome analyses of early cucumber fruit growth identifies distinct gene modules associated with phases of development was performed by Ando et al. (2012). The fruit samples produced 1.13 million ESTs which were assembled into 27,859 contigs with a mean length of 834 base pairs and a mean of 67 reads per contig. All contigs were mapped to the cucumber genome. Principal component analysis separated the fruit ages into three groups corresponding with cell division/preexponential growth (0 and 4 days post pollination (dpp), peak exponential expansion (8dpp), and late/postexponential expansion stages of growth (12 and 16 dpp). Transcripts predominantly expressed at 0 and 4 dpp included homologs of histones, cyclins, and plastid and photosynthesis related genes. The group of genes with peak transcript levels at 8dpp included cytoskeleton, cell wall, lipid metabolism and phloem related proteins. They showed that suggest that the interval between expansive growth and ripening includes further developmental differentiation with an emphasis on defense.

De novo Transcriptome Assembly (NGS) of Curcuma longa L. Rhizome Reveals Novel Transcripts Related to Anticancer and Antimalarial Terpenoids
Annadura et al. (2013). They sequenced the transcriptome of the rhizome of the 3 varieties of Curcuma longa L. using Illumina reversible dye terminator sequencing followed by De novo transcriptome assembly. Multiple databases were used to obtain a comprehensive annotation and the transcripts were functionally classified using GO, KOG and PlantCyc. Special emphasis was given for annotating the secondary metabolite pathways and terpenoid biosynthesis pathways. Their data not only provides molecular signatures for several terpenoids but also a comprehensive molecular resource for facilitating deeper insights into the transcriptome of C. longa.

Transcriptome Sequencing and Expression Analysis of Terpenoid Biosynthesis Genes in Litsea cubeba Han et al. (2013). Using Illumina paired-end sequencing, approximately 23.5 million high-quality reads were generated. De novo assembly yielded 68,648 unigenes with an average length of 834 bp. A total of 38,439 (56%) unigenes were annotated for their functions, and 35,732 and 25,806 unigenes could be aligned to the GO and COG database, respectively. By searching against the Kyoto Encyclopedia of Genes and Genomes Pathway database (KEGG), 16,130 unigenes were assigned to 297 KEGG pathways, and 61 unigenes, which contained the mevalonate and 2-Cmethyl-D-erythritol 4-phosphate pathways, could be related to terpenoid backbone biosynthesis. Of the 12,963 unigenes, 285 were annotated to the terpenoid pathways using the PlantCyc database. The expression patterns of the 16 genes related to terpenoid biosynthesis were analyzed by RT-qPCR to explore their putative functions.

2.8 Gene expression

In conventional PCR, the amplified product, or amplicon, is detected by an end-point analysis, by running DNA on an agarose gel after the reaction has finished. In contrast, real-time PCR allows the accumulation of amplified product to be detected and measured as the reaction progresses, that is, in “real time”. Real-time detection of PCR products is made possible by including in the reaction a fluorescent molecule that reports an increase in the amount of DNA with a proportional increase in fluorescent signal. The fluorescent chemistries employed for this purpose include DNA-binding dyes and fluorescently labeled sequence specific primers or probes. Specialized thermal cyclers equipped with fluorescence detection modules are used to monitor the fluorescence as amplification occurs.

The measured fluorescence reflects the amount of amplified product in each cycle. Gene expression profile has been widely used to address the relationship
between ecologically influenced or disease phenotypes and the cellular expression patterns. PCR–based detection technologies utilizing species specific primers are proving indispensable as research tools providing enhanced information on biology of plant/microbe interactions with special regard to the ecology, aetiology and epidemiology of plant pathogenic microorganisms.

The main advantage of real-time PCR over conventional PCR is that real-time PCR allows you to determine the starting template copy number with accuracy and high sensitivity over a wide dynamic range. Real-time PCR results can either be qualitative (presence or absence of a sequence) or quantitative (number of copies of DNA). Real-time PCR that is quantitative is also known as qPCR. In contrast, conventional PCR is at best semi-quantitative. Additionally, real-time PCR data can be evaluated without gel electrophoresis, resulting in reduced experiment time and increased throughput. Finally, because reactions are run and data are evaluated in a closed-tube system, opportunities for contamination are reduced and the need for post amplification manipulation is eliminated.

The RT PCR allows quantitative genotyping and detection of single nucleotide polymorphisms and allelic discrimination as well as genetic variations when only a small proportion of the sample carrying the mutation. The use of multiplex PCR systems using combined probes and primes targeted to sequences specific to counter partners of plant/microbe associations is becoming more important than standard PCR, which is proving to be insufficient for such living systems.

Gene expression and genetic analysis during higher plants embryogenesis studied by Abid et al. (2010). During embryo development, different posttranscriptional regulation mechanisms appear to be required to control the genetic programs induced at each stage. They identified and characterized embryo-defective mutants have been shown to be potentially useful in dissecting events in embryo development as a genetic approach. They analyzed many genes involved in diverse processes of embryo development have already been characterized through a wide range of techniques such as differential screening approaches or mutational screening. Seed development proceeds through a series of spatially and temporally regulated gene expression steps.
Digital gene expression analysis of gene expression differences within Brassica diploids and allopolyploids studied by Jiang et al. (2015). They examined the gene expression patterns of three diploids (Brassica rapa, B. nigra, and B. oleracea) and three amphidiploids (B. napus, B. juncea, and B. carinata) via digital gene expression analysis. In total, the libraries generated between 5.7 and 6.1 million raw reads, and the clean tags of each library were mapped to 18547–21995 genes of B. rapa genome. Gene ontological analysis was performed to functionally categorize these DEGs into different classes. The Kyoto Encyclopedia of Genes and Genomes analysis was performed to assign these DEGs into approximately 120 pathways, among which the metabolic pathway, biosynthesis of secondary metabolites, and peroxisomal pathway were enriched.

Cloning and analysis of expression patterns and transcriptional regulation of RghBNG in response to plant growth regulators and abiotic stresses in Rehmannia glutinosa performed by Zhou et al. (2015). RghBNG, a gene of unknown function, was cloned from Rehmannia glutinosa by reverse transcription PCR and rapid amplification of cDNA ends. The full-length cDNA of RghBNG was 548 bp with a 282-bp open reading frame. It encoded a polypeptide of 93 amino acids with a predicted molecular weight of 10.5 kDa and a theoretical isoelectric point of 9.25. Bioinformatics analysis indicated that RghBNG had no homology to any known plant genes, whereas the RghBNG polypeptide was highly similar to other plant proteins and possessed one conserved B12D protein family functional domain. RghBNG transcripts were detected in roots, stems, leaves, petals, receptacles, stamens and pistils with the highest and lowest levels respectively observed in petals and leaves of mature plants.

Wu et al. (2001) analyzed and checked expression of the a-expansin and b-expansin gene families in maize1 carried out. Expansins comprise a multigene family of proteins in maize (Zea mays). They isolated and characterized 13 different maize expansin cDNAs, five of which are a-expansins and eight of which are b-expansins. They analysed 13 expansins, as well as an expression analysis by northern blotting with materials from young and mature maize plants. Some expansins were expressed in restricted regions, such as the b-expansins ExpB1 (specifically expressed in maize pollen) and ExpB4 (expressed principally in young husks). Other expansins such as a expansin, Exp1 and b-expansin ExpB2
were expressed in several organs. The expression of yet a third group was not detected in the selected organs and tissues.

Gene expression analysis of flax seed development Venglat *et al* (2011). They described a large-scale generation and analysis of expressed sequences in various tissues. Collectively, the 13 libraries we have used provide a broad representation of genes active in developing embryos (globular, heart, torpedo, cotyledon and mature stages) seed coats (globular and torpedo stages) and endosperm (pooled globular to torpedo stages) and genes expressed in flowers, etiolated seedlings, leaves, and stem tissue. A total of 261,272 expressed sequence tags (EST) (GenBank accessions LIBEST_026995 to LIBEST_027011) were generated. These EST libraries included transcription factor genes that are typically expressed at low levels, indicating that the depth is adequate for in silico expression analysis. Assembly of the ESTs resulted in 30,640 unigenes and 82% of these could be identified on the basis of homology to known and hypothetical genes from other plants. When compared with fully sequenced plant genomes, the flax unigenes resembled poplar and castor bean more than grape, sorghum, rice or Arabidopsis. Digital analyses revealed gene expression dynamics for the biosynthesis of a number of important seed constituents during seed development.

Patterns of gene expression during Arabidopsis flower development from the time of initiation to maturation Ryan *et al*. (2015). Using a floral induction system, we collected floral buds at 14 different stages from the time of initiation until maturation. Using whole-genome microarray analysis, we identified 7,405 genes that exhibit rapid expression changes during flower development. These genes comprise many known floral regulators and we found that the expression profiles for these regulators match their known expression patterns, thus validating the dataset. We analyzed groups of co-expressed genes for over-represented cellular and developmental functions through Gene Ontology analysis and found that they could be assigned specific patterns of activities, which are in agreement with the progression of flower development. Furthermore, by mapping binding sites of floral organ identity factors onto our dataset, we were able to identify gene groups that are likely predominantly under control of these transcriptional regulators. They found that the distribution of paralogs among groups of co-expressed genes varies considerably, with genes expressed predominantly at early and intermediate stages of flower development showing the highest proportion of such genes.
Identification of new gene expression regulators specifically expressed during plant seed maturation Gutierrez et al. (2006). A cDNA-AFLP approach on *Linum usitatissimum* (flax) was used to identify genes specifically expressed during the seed maturation process. Among the 20 000 cDNA-AFLP tags produced, 486 were selected for their seed-specific expression during maturation. The results obtained confirmed the accuracy of the approach as numerous genes previously described as being expressed exclusively in plant seeds were identified in this screen. Using a real-time RT-PCR approach, seed-specific expression kinetics were confirmed for 13 of these regulators that were never characterized for being expressed during seed maturation. Among these, a flax gene of the non-LEC1-like HAP3 family and a flax MYB factor were shown to be expressed in specialized tissues of flax embryo using an in situ hybridization approach.

Differential Gene Expression in Soybean Leaf Tissues at Late Developmental Stages under Drought Stress Revealed by Genome-Wide Transcriptome Analysis Le et al. (2012). The availability of complete genome sequence of soybean has allowed research community to design the 66 K Affymetrix Soybean Array Gene Chip for genome-wide expression profiling of soybean. In this study, we carried out microarray analysis of leaf tissues of soybean plants, which were subjected to drought stress from late vegetative V6 and from full bloom reproductive R2 stages. They showed that out of 46093 soybean genes, which were predicted with high confidence among approximately 66000 putative genes, 41059 genes could be assigned with a known function. Using the criteria of a ratio change. 2 and a q-value,0.05, we identified 1458 and 1818 upregulated and 1582 and 1688 downregulated genes in drought-stressed V6 and R2 leaves, respectively.

Analysis of Gene Expression Patterns during Seed Coat Development in Arabidopsis Dean et al. (2011). Their data indicated that the developmental programs of the epidermis and the pigmented layer proceed relatively independently. Global expression datasets that can be used for identification of new gene candidates for seed coat development were generated. These dataset provide a comprehensive expression profile for developing seed coats in Arabidopsis, and should provide a useful resource and reference for other seed systems.
Visualizing Plant Development and Gene Expression in Three Dimensions Using Optical Projection Tomography Lee et al. (2006). A deeper understanding of the mechanisms that underlie plant growth and development requires quantitative data on three dimensional (3D) morphology and gene activity at a variety of stages and scales. To address this, they explored the use of optical projection tomography (OPT) as a method for capturing 3D data from plant specimens. They show that OPT can be conveniently applied to a wide variety of plant material at a range of scales, including seedlings, leaves, flowers, roots, seeds, embryos, and meristems. At the highest resolution, large individual cells can be seen in the context of the surrounding plant structure. For naturally semitransparent structures, such as roots, live 3D imaging using OPT is also possible. 3D domains of gene expression can be visualized using either marker genes, such as β-glucuronidase, or more directly by whole-mount in situ hybridization.

Genome-Wide Analysis Reveals Gene Expression and Metabolic Network Dynamics during Embryo Development in Arabidopsis Xiang et al. (2011). To address this gap, we have developed methods to isolate whole live Arabidopsis (Arabidopsis thaliana) embryos as young as zygote and performed genome-wide profiling of gene expression. These studies revealed insights into patterns of gene expression relating to: maternal and paternal contributions to zygote development, chromosomal level clustering of temporal expression in embryogenesis, and embryo-specific functions. Functional analysis of some of the modulated transcription factor encoding genes from our data sets confirmed that they are critical for embryogenesis.

2.9 SSR marker
2.9.1 Genetic Markers

In recent years, the molecular markers especially DNA-based markers, have been extensively used in many areas of research such as gene mapping and tagging (Kliebenstein et al. 2001; Karp and Edwards 1997), characterization of sex (Flachowsky et al. 2001; Martinez et al. 1999), analysis of genetic diversity, genetic relatedness, linkage map construction and marker assisted breeding (Kalia et al. 2011). According to Stansfield (1986), the term MARKER is usually used for “LOCUS MARKER”. Each gene has a particular place along the chromosome called LOCUS. Thus, a molecular marker is defined as a particular segment of DNA that is
representative of the differences at the genomic level. There are three major types of genetic markers: (1) morphological markers, which themselves are phenotypic characters of a trait (2) biochemical markers, which include allelic variants of enzymes called isozymes (3) DNA (or molecular) markers, which reveal sites of variation in DNA (Winter and Kahl 1995, Jones et al. 1997). Brief description of different types of genetic markers is given below.

2.9.2 Morphological markers
Morphological markers are visually characterized as phenotypic characters such as flower color, seed shape, growth habits or pigmentation. In conventional plant breeding programs breeder generally used to select desired plant types based on morphological data and also tried to correlate them to specific morphological markers. These markers are not of much utility in plant breeding programs.

2.9.3 Biochemical marker- allozymes (Isozyme)
In population genetics, protein-based markers (allozymes) were the first markers developed and widely used (Hamrick and Godt 1990). Allozyme was the first true molecular marker established to distinguish protein variants in enzymes (Schlotterer 2004). Isozymes were defined as structurally different molecular forms of an enzyme with, qualitatively, the catalytic function. Isozymes originate through amino acid alterations, which cause changes in net charge or the spatial structure (conformation) of the enzyme molecules and also, therefore, their electrophoretic mobility.

2.9.4 Molecular markers or DNA markers
Molecular markers are the DNA sequence variations that can be readily detected and whose inheritance can be monitored easily. The uses of molecular markers are based on the naturally occurring DNA polymorphism, which forms basis for designing strategies to exploit for applied purposes. A marker must be polymorphic i.e. it must exit in different forms so that chromosome carrying the mutant genes can be distinguished from it. Genetic polymorphism is defined as the simultaneous occurrence of a trait in the same population of two variants or genotypes.

2.9.5 Classification of DNA markers
The various types of available DNA markers can be classified broadly into following three groups:
2.9.5.1 First generation DNA markers (Hybridization based)
The first generation DNA marker system employed southern blot based markers such as RFLPs (Restriction Fragment Length Polymorphism). The RFLP technique employs molecular hybridization of cDNA or genomic DNA probes with genomic DNA digested with restriction enzymes. RFLP is the most widely used hybridization-based molecular marker. RFLP markers were first used in 1975 to identify DNA sequence polymorphism for genetic mapping of a temperature-sensitive mutation of adeno-virus serotypes (Grodzicker et al. 1975).

2.9.5.2 Second generation (PCR-based markers)
The second generation DNA-based molecular markers were driven by the invention of polymerase chain reaction (PCR) (Mullis et al. 1986). PCR is a molecular biology technique for enzymatically replicating (amplifying) small quantities of DNA and analyzed in many individuals without the requirement for cloning or isolating large amounts of ultra-pure genomic DNA. The various types of PCR based markers such as RAPD, AFLP, ISSR, IRAPs and SSR relies on the use of PCR primers, which binds to multiple or specific sites in the genome.

2.9.5.3 Third generation DNA markers (DNA sequence based)
In recent years, there has been an emphasis on the development of newer and more efficient high throughput molecular marker systems involving inexpensive non gel-based assays with high throughput detection systems i.e. SNPs (Single Nucleotide Polymorphism) (Gupta et al. 2001) and Microarrays (Linman et al. 2009).

2.9.6 Simple sequence repeats (SSRs) or microsatellite markers
The first widespread markers to avail full advantage of PCR technology was microsatellites (Litt and Luty 1989; Tautz 1989; Weber and May 1989). The genomes of higher organisms contain three types of multiple copies of simple repetitive DNA sequences (satellite DNAs, minisatellites and microsatellites) arranged in arrays of vastly differing size (Armour et al. 1999; Hancock 1999). Microsatellites, variously known as short tandem repeats (STR), Simple Sequence Repeats (SSRs) or Simple Sequence Length Polymorphism (SSLPs) are tandem repeats occurs in the form of iterations of repeat units of almost anything from a single base pair to thousands of base pairs (Litt and Luty 1989). Some researchers (e.g. Armour et al. 1999) define microsatellites as 2-8 bp repeats, others (e.g., Goldstein and Pollock 1997) as 1-6 or even 1-5 bp repeats (Schlotterer 1998). Mono-, di-, tri- and tetrancleotide repeats are
the main types of microsatellite, but repeats of five (penta-) or six (hexa-) nucleotides are usually classified as microsatellites as well.

Weber and May (1989) demonstrated that SSR polymorphisms (SSRPs) could be easily detected by PCR, using two flanking primers, which prompted the development of SSRs in various mammalian species and their subsequent assignment to specific chromosomes. In plants, the presence of SSRs was first demonstrated by the hybridization of oligonucleotide probes of poly (GT) and poly (AG) on the phage libraries of tropical tree genomes.

Microsatellites are highly polymorphic, abundant and fairly evenly distributed throughout the euchromatic part of the genomes. SSRs are non-randomly distributed within expressed sequence tags (ESTs), UTR regions, introns and coding regions. SSR variations within these regions can cause frame-shift, alteration in gene expression, inactivation of gene, change of function (Li et al. 2004). The significant part of SSR structure are functionally important for gene transcription, translation, chromatin organization, recombination, DNA replication, DNA MMR system, cell cycle.

2.9.6.1 Development of SSRs from EST sequences (genic or EST-SSRs)

A wealth of sequence data of ESTs has been generated as a result of sequencing projects for gene discovery from several plant species, giving scientists the flexibility to access many fulllength cDNA clones and characterized genes. These sequences are usually available in online database in public domain and can be downloaded and scanned for identification of SSRs. The expressed sequence tag-simple sequence repeats (EST-SSRs), genic SSRs or genederived SSRs, found in complementary DNA (cDNA) or ESTs sequence, between 1.1% and 4.8% have EST-SSR tandems (Saha et al. 2004). These informative markers are becoming Extracted genomic DNA from young leaves Digested Genomic DNA with restriction enzymes Size fractionated from 300 to 1500bp Adaptor ligation One way PCR amplification using adaptor primers Hybridization with SSR motif biotin probe and Streptavidin coated magnetic beads in amplicons. The development of EST-SSRs is relatively easy and inexpensive because they are produced form ESTs that are publicly available. However, the generation of EST-SSRs is largely limited to those species or close relatives for which the ESTs sequences are available. The EST-SSRs have some intrinsic advantages over genomic SSRs as they have higher transferability rate because the primers are designed from the more conserved coding regions of the
Review of Literature

Several search modules or programmes have been extensively used to identify SSRs such as:
MISA (MIcroSAtellite, http://pgrc.ipk-gatersleben.de/misa),
Sputnik (http://abajian.nert/sputnik/index.html),
SSRIT (SSR Identification Tool, http://www.gramene.org/db/searches/ssrtool),
TRF (Tandem Repeat Finder, http://tandem.bu.edu/trf/trf.html) etc.

Yadav et al. (2011) developed and deployed 50 EST-SSRs over 25 accessions of J. curcas collected from different geographical regions of India for genetic diversity analysis.

Cubry et al. (2014) developed 226 EST-SSR markers and tested over 19 Hevea accessions to assess the polymorphism and diversity analysis.

Jain et al. (2014) downloaded 13,513 ESTs from NCBI and obtained 7552 unigenes from these ESTs and developed 377 ESTs-SSR in Jatropha curcas. SSRs have been isolated for a number of plant species using this strategy. Greater DNA sequence conservation in transcribed regions, however, leads to lower polymorphism in EST-SSRs making them less efficient compared to genomic SSRs for distinguishing the closely related genotypes. Therefore, genomic SSRs are superior over EST-SSRs for fingerprinting or varietal identification studies.

Advantage SSR analysis

The nature of SSRs gives them a number of advantages over other molecular markers: (i) multiple SSR alleles may be detected at a single locus using a simple PCR-based screen, (ii) SSRs are evenly distributed all over the genome, (iii) they are co-dominant, (iv) very small quantities of DNA are required for screening and (v) analysis may be semi-automated.

1) High reproducibility: The high reproducibility would be the most important genetic analysis. While reproducibility of the SSR profile is as robust as it is with RFLPs, experimental procedures for SSR analysis are much simpler and require only a small amount of the template DNA. 2) Hyper-variability: The hyper-variable nature of SSRs produces very high allelic variations even among very closely related varieties. A literature survey showed that number of alleles varied from 1 to 37 with diversity indices of 0.29-0.95 in major crops pecies (Powell et al. 1996).The level of genetic variation detected by SSRs analysis was all most two times higher than that
detected by RFLPs, with 61 soybean lines (Morgante et al. 1994). In a comparative study of the utility of RFLP, RAPD, AFLP and SSR marker systems for germplasm analysis, SSRs showed the highest expected heterozygosity, while AFLP had the highest effective multiplex ratio. 3) **Co-dominancy:** The third advantage has to do with the co-dominant nature of SSR polymorphisms. Although homoplasicous bands can be misleading in scoring SSR profiles, the SSR bands produced from the same set of primers are intuitively orthologous. The multiple bands generated from RAPD and AFLP analysis do not permit their designation as allelic or orthologous bands until they are converted into STS markers after sequencing. The co-dominant nature of SSRs is suitable for genetical analysis in segregating F2 population or parentage analysis in hybrids (Scott et al. 2000; Slavov et al. 2005). 4) **Abundance and distribution:** The SSRs are shown to be highly abundant and distributed throughout the genomes (Wang et al. 1994; Toth et al. 2000, Varshney et al. 2005).

### 2.9.5 Application of SSR from genome

Microsatellites have become a marker of choice for an array of applications in plants due to hypervariable nature and extensive genome coverage. There are many applications of microsatellites in plants, a few can be categorized as:

- (1) genome mapping
- (2) cultivar identification and markers-assisted selection
- (3) genetic diversity analysis and phylogenetic relationship
- (4) population and evolutionary

#### 2.7.6.3.1 Genome mapping

Genome mapping consists of (a) genetic mapping (b) comparative mapping and (c) physical mapping.

**a) Genetic mapping**

Genetic mapping with microsatellite markers in plants was first reported in tropical trees (Condit and Hubbell 1991) and then reported in soybean (Akkaya et al. 1992) and rice (Wu and Tanksley 1993; Zhao and Kochert 1993).

**b) Comparative mapping**

Comparative mapping is the alignment of chromosome or chromosomal fragment of related species based on genetic mapping of common DNA markers and can trace the history of chromosome rearrangements during the evolution of plants, animals and insects. Microsatellite markers have been used for comparative mapping between *Quercus robur* (L.) and *Castanea sativa* (Mill.) (Barreneche et al. 2004). EST-SSRs were used in comparative mapping in wheat, barley, rye and rice (Stein et al. 2007).
(c) Physical mapping
SSR markers can be used as anchor markers for joining large pieces of overlapped DNA fragments such as bacterial artificial clones (BACs). Physical maps give us a real physical distance between markers or genes in bp (base pair) or kbp. SSR markers have been used to construct a whole genome physical map of model crop species.

2.9.6.3.2 Cultivar identification
SSR marker-assisted selection (MAS) can also greatly enhance the efficiency of plant breeding programs. SSR markers used for selection can be classified into flanking SSR markers (closely linked to the locus for a trait) and targeted gene SSR markers (developed within the targeted gene itself). In tomato, a set of 65 SSR markers has been selected for distinguishing 19 diverse tomato cultivars (He et al. 2003).

2.9.6.3.3 Genetic diversity and phylogenetic relationships
Genetic diversity refers to any variation in nucleotides, genes, chromosomes or whole genomes of organism. Genetic diversity can be assessed at different levels within a species or among species. SSR markers often are a powerful system for revealing interspecific or intraspecific phylogenetic relationship.

2.9.6.3.4 Population and evolutionary studies
Studies of plant evolution were traditionally based on taxonomic and phenotypic data (such as morphological and karyotype). Microsatellite markers can be used to determine the population structure within and among the natural population and identify the potential progenitors. The development of organelle specific SSR markers (i.e. cpSSR and mtSSR) had a great impact on the determination of structure and variation within a natural population as well as phylogenetic relationships.

Simple sequence repeats (SSRs) or microsatellites have been shown to be one of the most powerful genetic markers in biology. Defined as runs of tandem repeated DNA, they exhibit a high degree of polymorphism due to the mutation affecting the number of repeat units (Tautz, 1989). This hyper variability among related organisms makes them excellent markers for genotype identification, analysis of genetic diversity, phenotype mapping and marker assisted selection of crop plants.

Microsatellites are extremely informative markers that can be used for a variety of population genetics studies. Microsatellites are also considered ideal markers for genetic mapping studies (Hearne et al. 1992; Morgante and Olivieri, 1993) and germplasm evolution.
Li et al. (2014) identified SSRs and differentially expressed genes in two cultivars of celery (*Apium graveolens* L.) by deep transcriptome sequencing. They obtained 15,893,516 and 19,818,161 high-quality sequences by RNA-seq from two celery varieties ‘Ventura’ and ‘Jinnan Shiqin’, respectively. The obtained reads were assembled into 39,584 and 41,740 unigenes with mean lengths of 683 bp and 690 bp, respectively. A total of 1939 simple sequence repeat (SSR) markers were identified in Ventura’ and 2004 SSRs in Jinnan Shiqin’. Di-nucleotide repeats were the most common repeat motif, accounting for 55.49% and 54.84% in Ventura’ and Jinnan Shiqin’, respectively. A comparison of expressed genes between the two libraries, identified 338 differentially expressed genes (DEGs). Three hundred and three of the DEGs were annotated based on a sequence similarity search utilizing eight public databases.

Microsatellite isolation and marker development in carrot studied by Cavagnaro et al. (2011) for genomic distribution, linkage mapping, genetic diversity
analysis and marker transferability across Apiaceae. Overall, 196 SSRs (65.1%) were polymorphic in at least one mapping population, and the percentage of polymorphic SSRs across F2 populations ranged from 17.8 to 24.7. Polymorphic markers in one family were evaluated in the entire F2, allowing the genetic mapping of 55 SSRs (38 codominant) onto the carrot reference map. The SSR loci were distributed throughout all 9 carrot linkage groups (LGs), with 2 to 9 SSRs/LG.

Development and validation of EST-SSR markers in adzuki bean (Vigna angularis) performed by Chen et al. (2015) utilized the transcriptome sequencing. They developed EST-SSR markers for the adzuki bean through next-generation sequencing. More than 112 million high-quality cDNA sequence reads were obtained from adzuki bean using Illumina paired-end sequencing technology, and the sequences were De novo assembled into 65,950 unigenes. The average length of the unigenes was 1,213 bp. Among the unigenes, 14,547 sequences contained a unique simple sequence repeat (SSR) and 3,350 sequences contained more than one SSR. A total of 7,947 EST-SSRs were identified as potential molecular markers, with mononucleotide A/T repeats (99.0%) as the most abundant motif class, followed by AG/CT (68.4%), AAG/CTT (30.0%), AAAG/CTTT (26.2%), AAAAG/CTTTT (16.1%), and AACGGG/CCCGTT (6.0%).

In Bupleurum chinense Sui et al. (2009) studied DC development of genomic SSR and potential EST-SSR markers. Nineteen genomic SSR markers were developed using inter-simple sequence repeat (ISSR) suppression PCR technique in Bupleurum chinense DC, a widely used Chinese medicinal plant. A total of 126 alleles were detected across 22 individual plants of B. chinense DC. f. octoradiatum (Bunge) with an average of 3-13 alleles per locus. The species transferability of these genomic SSR markers was also detected in seven other Bupleurum species. Eight SSR markers were successfully amplified in all tested species. In addition, forty four EST-SSRs which can be amplified with expected sizes were identified from a B. chinense root cDNA library.

De novo Assembly, Gene Annotation and Marker Development Using Illumina Paired-End Transcriptome Sequences in Celery (Apium graveolens L.) Fu et al. (2013). Celery transcriptomes from four tissues were sequenced using Illumina paired-end sequencing technology. De novo assembling was performed to generate a collection of 42,280 unigenes (average length of 502.6 bp) that represent the first transcriptome of the species. 78.43% and 48.93% of the unigenes had significant

Review of Literature
similarity with proteins in the National Center for Biotechnology Information (NCBI) non-redundant protein database (Nr) and Swiss-Prot database respectively, and 10,473 (24.77%) unigenes were assigned to Clusters of Orthologous Groups (COG). 21,126 (49.97%) unigenes harboring Interpro domains were annotated, in which 15,409 (36.45%) were assigned to Gene Ontology (GO) categories. Additionally, 7,478 unigenes were mapped onto 228 pathways using the Kyoto Encyclopedia of Genes and Genomes Pathway database (KEGG). Large numbers of simple sequence repeats (SSRs) were indentified, and then the rate of successful amplification and polymorphism were investigated among 31 celery accessions.

A comprehensive characterization of simple sequence repeats in pepper genomes provides valuable resources for marker development in Capsicum performed by Cheng et al. (2015). They identified an average of 868,047.50, 45.50 and 30.00 SSR loci in the nuclear, mitochondrial and chloroplast genomes of pepper, respectively. Subsequently, systematic comparisons of various species, genome types, motif lengths, repeat numbers and classified types were executed and discussed. In addition, a local database composed of 113,500 in silico unique SSR primer pairs was built using a homemade bioinformatics workflow. As a pilot study, 65 polymorphic markers were validated among a wide collection of 21 Capsicum genotypes with allele number and polymorphic information content value per marker raging from 2 to 6 and 0.05 to 0.64, respectively.

De novo assembly and Characterisation of the Transcriptome during seed development, and generation of genic-SSR markers in Peanut (Arachis hypogaea L.) Zhang et al. (2012 b). A total of 26.1-27.2 million paired-end reads with lengths of 100 bp were generated from the three varieties and 59,077 unigenes were assembled with N50 of 823 bp. Based on sequence similarity search with known proteins, a total of 40,100 genes were identified. Among these unigenes, only 8,252 unigenes were annotated with 42 gene ontology (GO) functional categories. In addition, 3,919 microsatellite markers were developed in the unigene library, and 160 PCR primers of SSR loci were used for validation of the amplification and the polymorphism. They developed four thousand SSR primers.

Development and characterization of simple sequence repeat (SSR) markers based on RNA-sequencing of medicago sativa and in silico mapping onto the M. truncatula genome Wang et al. (2014 b). A total of 54,278 alfalfa non-redundant unigenes were assembled through the Illumina HiSeqTM 2000 sequencing
technology. Based on 3,903 unigene sequences, 4,493 SSRs were identified. Tri-nucleotide repeats (56.71%) were the most abundant motif class while AG/CT (21.7%), AGG/CCT (19.8%), AAC/GTT (10.3%), ATC/ATG (8.8%), and ACC/GGT 20 (6.3%) were the subsequent top five nucleotide repeat motifs. Eight hundred and thirty-seven EST-SSR primer pairs were successfully designed. Of these, 527 (63%) primer pairs yielded clear and scored PCR products and 372 (70.6%) exhibited polymorphisms.

Zhang et al. (2013) Developed EST-SSR markers to study genetic diversity in hyacinth bean (Lablab purpureus L.). A total of 459 hyacinth bean ESTs from the National Center for Biotechnology Information (NCBI) database were downloaded and analyzed to search for SSRs with a minimum of 12 repeating nucleotides. Finally, 22 microsatellites were identified in 420 unigenes, indicating that merely 4.79% of the sequences contained SSRs. EST-SSR loci were subsequently screened on 24 hyacinth bean accessions collected from both China and Africa. Among 22 EST-SSRs, 11 loci showed polymorphism and revealed two to four alleles per locus. Tri-nucleotide motifs were the most common type of repeats, accounting for 54.55% of the total, followed by di-nucleotide repeats. The polymorphic information content (PIC) values ranged from 0.0767 to 0.4864, with a mean of 0.286. Furthermore, both principal coordinate analysis (PCA) and phylogenetic tree analysis indicated that all accessions were clustered into two main groups, and that all 19 Chinese accessions were clustered into the single group.

Development and characterization of genomic SSR markers in Cynodon transvaalensis Burtt-Davy Tan et al. (2014). Genomic DNA of C. transvaalensis ‘4200TN 24-2’ from an Oklahoma State University (OSU) turf nursery was extracted for construction of four SSR genomic libraries enriched with [CA]n, [GA]n, [AAG]n, and [AAT]n as core repeat motifs. A total of 3,064 clones were sequenced at the OSU core facility. The sequences were categorized into singletons and contiguous sequences to exclude redundancy. From the two sequence categories, 1,795 SSR loci were identified. After excluding duplicate SSRs by comparison with previously developed SSR markers using a nucleotide basic local alignment tool, 1,426 unique primer pairs (PPs) were designed. The SSR markers developed in the study are the first large set of co-dominant markers in African bermuda grass and should be highly valuable for molecular and traditional breeding research.
2.9.6.2 Development of SSRs through search of genome sequences

Another approach for the isolation of SSRs involves a computational search of the genome databases.

Wu et al. (2017), developed novel SSR markers for flax (*Linum usitatissimum* L.) using reduced-representation genome sequencing. They screened 1574 microsatellites from *Linum usitatissimum* L. using reduced representation genome sequencing (RRGS) to systematically identify SSR markers. The resulting set of microsatellites consisted mainly of tri nucleotide (56.10%) and dinucleotide (35.23%) repeats, with each motif consisting of 5–8 repeats. They evaluated marker sensitivity and specificity based on samples of 48 flax isolates obtained from northeastern China. Using the new SSR panel, the results demonstrated that fiber flax and oilseed flax varieties clustered into two well separated groups. Then novel SSR markers developed in this study show potential value for selection of varieties for use in flax breeding programs.

Weber and May (1989) reported abundance of the (CA)n SSRs in the human genome through search of human genome sequence database. In plants, the first report of (CA)n repeats in soybean sequence was the one from a computer search of the Gene Bank databases (Akkaya et al. 1992). Morgante and Olivieri (1993) showed an abundance of (AT)n repeats in 30 different plant genomes and demonstrated that the analysis of repeat number variations by PCR was highly informative in genome analyses.

Wang et al. (2014) developed 1129 EST-SSRs from two 454 sequencing cDNA libraries of *Gossypium barbadense* and 311 polymorphic loci were integrated into interspecific BC1 genetic linkage map.

Wang et al. (1994) surveyed mono-, di-, tri- and tetra-nucleotide repeats which were all present in non-coding regions, but 57% of the tri-nucleotide repeats, containing G-C base pairs, reside in the coding region. The abundance of tri-nucleotide repeats in coding region was attributed to the other types being eliminated from the coding sequences because of their ability to cause frame-shift mutations. However, in species lacking genomic sequences, a computer search would not be useful in developing large scale SSR markers. Construction of SSR-enriched libraries leading to development of large-scale sequences would be a plausible way to develop SSR markers.
By utilized genome survey, Wei et al. (2014) developed Simple Sequence Repeat (SSR) Markers of Sesame (*Sesamum indicum*). They reported for the first time a whole genome survey used to develop simple sequence repeat (SSR) markers and to detect the genetic diversity of sesame germplasm. From the initial assembled sesame genome, 23,438 SSRs (≥5 repeats) were identified. The most common repeat motif was dinucleotide with a frequency of 84.24%, followed by 13.53% trinucleotide, 1.65% tetranucleotide, 0.3% pentanucleotide and 0.28% hexanucleotide motifs. From 1500 designed and synthesised primer pairs, 218 polymorphic SSRs were developed and used to screen 31 sesame accessions that from 12 countries.

2.10 Metabolomics

Metabolic fingerprinting is a rapid and global analysis, which often does not involve chromatographic separation and metabolites are generally not identified (Choi et al., 2004). Metabonomics is a similar strategy to metabolomics, but is often used in toxicology studies with NMR for global metabolite screenings (Nicholson et al., 2002).

The first large-scale plant metabolite profiling was first performed by Roessner et al. 2000 on potato tubers (*Solanum tuberosum*) and Arabidopsis thaliana leaf extracts (Fiehn et al., 2000) using Gas Chromatography-Mass Spectrometry (GC/MS). Both metabolic profiling and metabolomics have been applied to many different plant species. To give some examples, the techniques have been used to study the metabolome of rice leaves (Sato et al., 2004), changes in the metabolome during cold acclimation of Arabidopsis (Cook et al., 2004), in comparisons of metabolic profiles of alfalfa and *M. truncatula* to identify new saponins (Huhman & Sumner, 2002), studies of growth processes in hybrid aspen (*Populus tremula* L. x *tremuloides* Michx.; Wiklund et al., 2005) and characterization of flavonoid glycosides in genetically modified tomato (Le Gall et al., 2003). In addition, metabolomic approaches have been applied in flux analyses to elucidate plant metabolism (reviewed by Fernie, Geigenberger & Stitt, 2005), to determine stresses in genetically modified and nutrition research (German, Roberts & Watkins, 2003).

2.10.1 Instruments used for Metabolomics

Numerous analytical techniques have been used in the field of plant metabolomics to monitor and explore metabolic differences between biological samples GC /MS and Liquid chromatography-mass spectrometry (LC/MS). Other important analytical techniques include liquid chromatography-photodiode array
detection-mass spectrometry (LC/PDA/MS; Huhman & Sumner, 2002), capillary electrophoresismass spectrometry (CE/MS; Soga et al., 2003; Sato et al., 2004), Fourier-transform ion-cyclotron mass spectrometry (FT/MS; Tohge et al., 2005) and Nuclear magnetic resonance (NMR; Ward et al., 2003; Wiklund et al., 2005)

2.10.1.1 Gas Chromatography-Mass Spectrometry (GC/MS)

Hyphenated chromatography-mass spectrometry approaches, such as high performance liquid chromatography (HPLC) and gas chromatography (GC) in combination with MS enables good metabolite identification and quantification compared to many other methods for screening metabolites. GC/MS (Watson, 1997; de Hoffmann & Stroobant, 2002) provides a robust system with excellent separation capacities and high thought-put possibilities, and is therefore the most commonly used analytical technique for routine analyses in the field of plant metabolomics. The separation of the analytes in gas chromatography is dependent on analyte interactions with the stationary face and the boiling point. Only compounds that are volatile can be separated on a GC column, so non-volatile metabolites must be derivatized prior to analysis by GC.

The inlet temperature is often higher than 250°C, at which many metabolites are evaporated. Two different injection methods are most widely used; splitless and split (Watson, 1997). In splitless injections the whole sample is introduced onto the high resolution capillary column, which is preferable to split injections (where only a portion is used) for trace analyses. The length of the capillary column varies between 10 to 60 metres. The polarity of the column can also be varied by changing the phenyl stationary phases, such as DB-5, DB-50 and CPSil-8.

Metabolites eluting from GC are ionized by Electron-impact (EI) or Chemical Ionization. For metabolic analysis EI (Fiehn et al., 2000b) is the most commonly used technique. In EI vaporized metabolites are impacted by a beam of electrons with sufficient energy to 11 fragment and ionize the molecule. The number of fragment ions that is produced of each metabolite is a function of the electron impact energy. The source is designed so that when the ions are formed they are pushed out from the source and into the mass analyzer. EI results in molecular ion fragmentation, which is of great importance for structural interpretation of the metabolites. In comparison with EI, CI is a much softer ionization technique, in which the ions are allowed to collide with reagent gas (often proton-rich) to form abundant adduct ions that contain the intact molecular species (Watson, 1997). This is
advantageous for determining the molecular weight of metabolites. To identify compounds, commercially available databases of molecular ion fragmentation patterns of molecules, such as NIST (http://www.nist.gov/srd/nist1.htm; 19-August-2005), can be used. Unfortunately, the number of derivatized plant metabolites is limited and additional retention index information is incomplete.

2.10.2 Metabolic Pathways

Metabolomics can be used to explain the biochemical function of annotated genes. It can also be used to define phenotypes and to bridge the genotype-to-phenotype gap (Fiehn, 2002). Furthermore, metabolites can also be related by their molecular structure and the fact that that they are built up by other metabolites. This can be visualized in maps or biochemical pathways describing the linkage between metabolite reactions. Known metabolite relationships have been used to compile publicly available reference biochemical reference databases (Mueller, Zhang & Rhee, 2003; Lange & Ghassemian, 2005). These reference biochemical databases contain not only information about biochemical pathways, cellular and molecular processes, but also information about the proteins that catalyse the reactions and the genes that code for them. An important point to remember is that the database information is limited since it does not cover species-specific pathways and the resulting diagrams do not cover all of the side reactions. However, databases can still provide a powerful visualization tool for the biological context of functional information.

2.10.3 Metabolomics study:

Aroma characterization of coriander (Coriandrum sativum L.) oil samples was carried out by Ravi et al. (2006). Coriander (Coriandrum sativum L.) seeds from eight regions of India, labelled as S1 to S8 were examined for their volatile constituents by gas chromatography–mass spectroscopy (GC–MS). GC–olfactometry (GC–O) was carried out for major compounds and odour profiling was done by trained panelists. The GC–MS analysis revealed presence of 30 compounds in coriander oil and around 98% of the compounds were identified in all the samples. Linalool which has floral and pleasant odour notes was the major compound (56.71–75.14%) in the essential oil. Higher α-pinene content of S7 and S8 could be related to the higher turpentine note. Sweet and rose-like odour notes of S1 could be due to occurrence of higher levels of geranyl acetate and lemonol. Principal component
analysis showed that samples S7 and S8 loaded with \( \alpha \) pinene, myrcene and undecanal.

Farag et al. (2014) analyzed metabolites of six Nigella species seeds via UPLC-qTOF-MS and GC–MS coupled to chemometrics. Nigella sativa, commonly known as black cumin seed, that contains numerous phytochemicals including terpenoids, saponins, flavonoids, alkaloids. Large scale metabolic profiling including UPLC-PDA-MS and GC–MS with further multivariate analysis was utilized to classify 6 Nigella species. Under optimized conditions, they annotated 52 metabolites including 8 saponins, 10 flavonoids, 6 phenolics, 10 alkaloids, and 18 fatty acids.

Metabolomics and molecular marker analysis to explore pepper (Capsicum sp.) biodiversity Wahyuni et al. (2013). Metabolic diversity in ripe fruits of a collection of 32 diverse pepper (Capsicum sp.) accessions was obtained by measuring the composition of both semi-polar and volatile metabolites in fruit 21 pericarp, using untargeted LC–MS and headspace GC–MS platforms. Species-specific clustering was also observed when accessions were grouped based on their semi-polar metabolite profiles. In total 88 semi-polar metabolites could be putatively identified. A large proportion of these metabolites represented conjugates of the main pepper flavonoids (quercetin, apigenin and luteolin) decorated with different sugar groups at different positions along the aglycone.

Physio-biochemical composition and untargeted metabolomics of cumin (Cuminum cyminum L.) make it promising functional food and help in mitigating salinity stress Pandey et al. (2015). The study is intended to comprehensively analyse physiological parameters, biochemical composition and metabolites under salinity stress. In contrast, total antioxidant activity, \( \text{H}_2\text{O}_2 \), proline and MDA contents increased concurrently with stress treatments. A total of 45 differentially expressed metabolites were identified, including luteolin, salvianolic acid, kaempferol and quercetin, which are phenolic, flavonoid or alkaloids in nature and contain antioxidant activities. Additionally, metabolites with bioactivity such as anticancerous (docetaxel) and antimicrobial (megalogenicin) properties were also identified. Metabolome analysis of 20 taxonomically related benzylisoquinoline alkaloid-producing plants Hagel et al. (2015). They applied a multi-platform approach incorporating four different analytical methods to examine 20 non-model, BIA-accumulating plant species. One dimensional 1H NMR-based profiling quantified 91 metabolites and revealed significant species-and tissue-specific variation in sugar, amino acid and organic acid content. Direct
flow infusion tandem mass spectrometry provided a broad survey of 110 lipid derivatives including phosphatidylcholines and acylcarnitines, and high-performance liquid chromatography coupled with UV detection quantified 15 phenolic compounds including flavonoids, benzoic acid derivatives and hydroxycinnamic acids.

GC-MS Determination of Bioactive compounds of Enicostemma littorale (Blume) (Ambikapathy et al. 2011). In their study, the bioactive compounds of Enicostemma littorale have been evaluated using GCMS. The chemical compositions of the whole plant methanol extract of E. littorale were investigated using Perkin-Elmer Gas chromatography-mass spectrometry. The mass spectra of the compounds, found in the extract was matched with characterization and measurement of the central Electrochemical Research institute. GC-MS analysis of E. littorale 22 whole plant methanol extract revealed the existence of the ether compound-Laminaribiitol (79.93%), 12-hydroxy-9-octadecenoic acid (9.546%). Myricetin (4.7519%), 3,3-Methylenebis (4- hydroxycoumarin) (2.811%), catechin (2.002%).

GC–MS metabolic profiling of Cabernet Sauvignon and Merlot cultivars during grapevine berry development and network analysis reveals a stage- and cultivar-dependent connectivity of primary metabolites Inostroza et al. (2016). They developed metabolic profiling analysis across six stages from flowering to fully mature berries of two cultivars, Cabernet Sauvignon and Merlot is reported at metabolite level. Based on a gas chromatography–mass spectrometry untargeted approach, 115 metabolites were identified and relative quantified in both cultivars. Sugars and amino acids levels show an opposite behavior in both cultivars undergoing a highly coordinated shift of metabolite associated to primary metabolism during the stages involved in growth, development and ripening of berries.