CHAPTER-II
REVIEW OF LITERATURE

A brief review of research work “De novo Sequencing of Ancient Seed Spice Celery (Apium graveolens L.) and Development of Microsatellites” has been reviewed and highlighted under the following headings. To fill up the gap, wherever it was necessary, the review on other relevant crop is included.

2.1 NOMENCLATURE
   2.1.1 Scientific Classification
   2.1.2 Introduction and Spread History
   2.1.3 Cultivation
   2.1.4 Economic Importance

2.2 GENOME SEQUENCING
   2.2.1 Flow Cytometry
   2.2.2 DNA Isolation
   2.2.3 Next Generation Sequencing Technologies
   2.2.4 Genome Sequenced of Various Plants
   2.2.5 De Novo Genome Assembly Approaches
   2.2.6 Problems of assembling complex genomes
   2.2.7 Bioinformatical Tools

2.3 MOLECULAR MARKERS
   2.3.1 Simple sequence repeats (SSRs)

2.1 NOMENCLATURE

Scientific name: Apium graveolens L., English name: celery, leaf celery, stalk celery, celeriac, turnip-rooted celery, Sanskrit name: Ajmoda, Apium: from the German Eppich. Graveolens: from the Latin Gravis meaning "grave, heavy" and Olens "smelling” from the verb olere. The Greek writer Homer referred to celery as "selinon". The Latin name was "selinun" and the French name "celeri" is similar to the name used today.
2.1.1 Scientific Classification

Scientific classification of celery is as described.

**Scientific Classification**

<table>
<thead>
<tr>
<th>Domain</th>
<th>Eukaryota</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kingdom</td>
<td>Plantae</td>
</tr>
<tr>
<td>Subkingdom</td>
<td>Tracheobionta</td>
</tr>
<tr>
<td>Division</td>
<td>Spermatophytes</td>
</tr>
<tr>
<td>Sub-division</td>
<td>Angiospermae</td>
</tr>
<tr>
<td>Class</td>
<td>Mangnolisisa</td>
</tr>
<tr>
<td>Sub-class</td>
<td>Rosidae</td>
</tr>
<tr>
<td>Order</td>
<td>Apicedes</td>
</tr>
<tr>
<td>Family</td>
<td>Apiceae</td>
</tr>
<tr>
<td>Genus</td>
<td>Apium</td>
</tr>
<tr>
<td>Species</td>
<td>graveolens</td>
</tr>
<tr>
<td>Binomial name</td>
<td>Apium graveolens L.</td>
</tr>
</tbody>
</table>

2.1.2 Introduction and Spread History

*Apium dulce* Mill. is often listed as a synonym of celery. In Latin language, celery was called *sedano* (which gave rise to celery and its Italian, German and French cognates) or *apium*, which is found in Catalan and Provencal and the regional German word *Eppich*. The ultimate origin of both names is dark.

Celery leaves and inflorescences were part of the garlands found in the tomb of Tutankhamun, pharaoh of ancient Egypt and celery mericarps dated to the seventh century B.C.E. were recovered in the Heraion of Samos (Zahory and Hopf, 2000). Another archeological find of celery, dating to the ninth century B.C.E., was at Kastanas, Greece. However, the literary evidence for ancient Greece is far more abundant. In Homer’s *Iliad*, the horses of Myrmidons graze on wild celery that grows in the marshes of Troy, and in *Odyssey* there is mention of the meadows of violet and wild celery surrounding the cave of Calypso (Fragiska, 2005). Celery is believed to be originally from the Mediterranean basin. Ancient literature documents that celery, or a similar plant form, was cultivated for medicinal purposes before 850 B.C. It’s claimed medicinal purposes were probably attributable to its volatile oils, contained in all portions, but
mostly the seed. During ancient times Ayurvedic physicians used celery seed to treat the following conditions: colds, flu, water retention, poor digestion, various types of arthritis, and liver and spleen ailments. Woven garlands of wild celery are reported to have been found in early Egyptian tombs. Although celery is thought to be from the Mediterranean, indigenous "wild" relatives of celery are found in southern Sweden, the British Isles, Egypt, Algeria, India, China, New Zealand, California and southernmost portions of South America. However it is doubtful that it’s center of origin was that extensive.

The native habitual of celery is the lowland of Italy from where it spreads to Sweden, Egypt, Algeria and Ethiopia in Asia to India. Celery was firstly cultivated as a food plant in France in 1623. In India it is cultivated in north–western Himalayas, Punjab, Haryana and western Uttar Pradesh in an area of about 5000 ha. Punjab produces about 90% of the total Indian production (Fazal and Singla, 2012).

Celery was considered a holy plant in the classical period of Greece and was worn by the winners of the Nemean Games, similar to the use of bay leaves at the Olympic Games. The Nemean Games were conducted every second year, starting in 573, in the small city of Nemea in southern Greece in the Poloponnes peninsula. The Romans valued celery more for cooking than for religion although much superstition was connected with it. The celery plant was thought to bring bad fortune under certain circumstances.

The Italians domesticated celery as a vegetable in the 17th century resulting in selections with solid stems. Early stalk celery had a tendency to produce hollow stalks. After years of domestication, selection eliminated this characteristic as well as bitterness and strong flavors. Early growers found that the naturally strong flavors could be diminished if grown in cooler conditions and also if blanched. Blanching is the practice of pushing dirt up around the base of the stalks to prevent sunlight from turning the stalks green.

There are two types of stalk celery varieties, self-blanching or yellow, and green or Pascal celery. In North America green stalk celery is preferred and mainly eaten raw, although it is also eaten cooked. In Europe and the rest of the world self-blanching varieties are preferred. Celeriac is very popular in Europe where it is eaten cooked or raw. Smallage is grown in Eastern Europe and Asia for its seed as well as to use the aromatic leaves to flavor cooked food and to garnish plates. In some areas celery and celery seed is consumed to treat high blood pressure. Celeriac is becoming popular as a part of trendy American gourmet eating.
2.1.3 Cultivation

There are two main varieties of celery grown today: Pascal celery, which is pale green and the most common variety, and Golden celery, which is grown under cover (a layer of soil or paper) to prevent chlorophyll from developing and imparting the green colour.

In North America, commercial production of celery is dominated by Pascal celery. Gardeners can grow a range of cultivars, many of which differ little from the wild species, mainly in having stouter leaf stems. They are ranged under two classes, white and red; the white cultivars being generally the best flavored, and the most crisp and tender. The wild form of celery is known as smallage. It has a furrowed stalk with wedge-shaped leaves. The whole plant has a coarse, rank taste and a peculiar smell. With cultivation and blanching, the stalks lose their acidic qualities and assume the mild, sweetish, aromatic taste particular to celery as a salad plant.

Celery grows to a height of 60 to 90 cm. It has a shallow tap root system. The stem is branched succulent and ridged. The leaflets are ovate to sub-orbicular three lobes 2-4.5 cm long. The inflorescence is a compound umbel. The flowers are small and white. The carpels are sub-pentagonal. The primary ridges are distinct and filiform. The fruit is a schizocarp, with two mericarps, 1-2 mm in diameter, aromatic and slightly bitter. Celery is naturally cross-pollinated but not self-incompatible. Celery can be successfully cultivated on all soils, except saline, alkaline and waterlogged. Loamy soils are the best. Celery is sensitive to the extremes of soil reaction. The soil pH around 5-7 is suitable. A combination of 12-15°C and 22-25°C day and night temperature respectively, gives 80% seed germination within two weeks. So cold and dry climate is the suitable climate. The seeds are sown in March-April and seeds are transplanted in May and the crop is ready to be harvested in November. In India, during 2014-2015 the area under celery cultivation was 33,000 hectares and production was 33,000 tonnes while in Gujarat the figures were 7,400 hectares and 9,010 tonnes respectively (Anon., 2015).

In the past, celery was grown as a vegetable for winter and early spring; because of its antitoxic properties, it was perceived as a cleansing tonic, welcomed after the stagnation of winter.

2.1.4 Economic Importance

Commercially, celery is available as seeds, vegetable, seed oil and oleoresin. The seed contains 2.2% essential oil (2.2-3.0 %). The oil is pale yellow in colour, contains d-
Review of Literature

limonene (60%), ß-selinene (10-12%), sedanoic acid anhydride (0.5%) and sedanolide (2.5-3.0%). The latter two components contribute to its characteristic odour. The leaves are rich source of minerals like Ca, P, Fe, vitamin A and vitamin C for which it is consumed in salad.

The dried, ripe seeds are used as spice to flavour food and liquids, the seed is a stimulant and carminative and is used as a nerve tonic in indigenous medicine systems. It is also used as a remedy for rheumatism. The seed oil is used for flavouring food items and in the perfumery and pharmaceutical industries. The fatty oil obtained from the fruit is used as an antispasmodic and nerve stimulant.

2.2 GENOME SEQUENCING

The term genome sequencing refers to sequencing methods for determining the order of the nucleotide bases - adenine, guanine, cytosine, and thymine in a molecule of DNA. Knowledge of DNA sequences has become indispensable for basic biological research, other research branches utilizing DNA sequencing, and in numerous applied fields such as diagnostics, biotechnology, forensic biology and biological systematics.

2.2.1 Flow Cytometry

The expression ‘genome size’ is often used for the DNA content of the monoploid genome or chromosome set, whereas ‘DNA C-value’ stands for the DNA content of the whole chromosome complement or karyotype irrespective of the degree of generative polyploidy of the organism. However, ‘genome size’ and ‘DNA C-value’ are often also used synonymously.

Nath et al. (2014) tried to standardize a methodology for estimation of genome size by flow cytometry in five mucilaginous species of Hyacinthaceae family. Five reported nuclei isolation buffers were tested in Drimia indica to overcome the problem of presence of mucilage that caused error with nuclei sample preparation. The Galbraith’s buffer, which gave comparatively better results after being modified by increasing pH, detergent concentration and replacement of sodium citrate by citric acid, allowed better sample preparation with inflated yield, depleted debris and improved DNA peak CV i.e. coefficient of variation.

Bennett et al. (2003) performed the flow cytometry of Arabidopsis. Galbraith's buffer used to suspend nuclei for flow cytometry consisted of, per litre, 4.26 g MgCl₂, 8.84 g sodium citrate, 4.2 g of 3-(N-morpholino) propane sulfonic acid, 1 ml Triton
X-100 and 1 mg boiled ribonuclease A, adjusted to pH 7.2. This buffer was successfully used for *Arabidopsis* and gave high stability of plant and animal nuclei for up to 30 h. Seedlings grown on agar in a petri dish, or newly expanded rosette leaves from flowering plants of *Arabidopsis* were manually chopped using a razor blade to release nuclei into Galbraith buffer and filtered through a 53 µm nylon mesh to recover 1 ml, which was kept on ice. For comparison in mixtures with *Arabidopsis*, a small aliquot (approx. 50 µl) was added to a sample of *Arabidopsis* nuclei and propidium iodide (PI) was then added. The mean fluorescence of co-stained nuclei in at least ten replicate samples was quantified using a Coulter Epics Elite (Coulter Electronic, Hialeah, FL, USA) flow cytometer with a laser tuned at 514 nm. Fluorescence at >615 nm was detected by a photomultiplier screened by a long pass filter. This gave a 1C DNA amount of approx. 153 Mb for *Arabidopsis*.

### 2.2.2 DNA Isolation

Healey *et al.* (2014) described simple modifications to the established CTAB-based extraction method that allows for reliable isolation of high molecular weight genomic DNA from difficult to isolate plant species Corymbia (a eucalypt) and Coffea (coffee). The simplified protocol does not require multiple clean up steps or commercial based kits, and the isolated DNA passed stringent quality control standards for whole genome sequencing on Illumina HiSeq and TruSeq sequencing platforms. With this robust protocol, whole genome sequencing is possible from recalcitrant plant species using established DNA sequencing technologies for advanced bioinformatics.

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 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 analyze diversity and relationships among 48 pepper (*Capsicum* spp.) genotypes originating from nine countries.

Sharma *et al.* (2012) standardized a protocol for rapid extraction of high quantity and quality of genomic DNA for *Ferula jaeschkeana* Vatke (Apiaceae). The protocol utilized preserved leaf material for DNA extraction. Liquid nitrogen was used for grinding the leaves. The isolated DNA was essentially free of polyphenolics,
polysaccharides and other major contaminants as revealed by color difference of DNA isolated. The quality and quantity of DNA was also visualized on 0.8% agarose gel.

Chen et al. (1999) developed a rapid DNA mini-preparation method for rice and other plant species. An eppendorf tube and 1ml pipette tip was used to grind plant tissues, and only one transfer for DNA isolation was required. The yields of the DNA samples ranged from 2.3-5.2 mg from 25-50 mg fresh leaf tissue. DNA samples extracted using this method from rice were completely digested with five restriction enzymes (EcoR I, EcoR V, Hind III, Mse I and Pst I) and were successfully used for AFLP and other PCR applications.

2.2.3 Next Generation Sequencing Technologies

Many crop genomes are large, complex and often polyploid, making genome sequencing a major challenge. With the decreasing cost of second generation DNA sequencing technologies, genome size in itself would not prevent the application of whole genome sequencing approaches.

The remarkable improvement of sequencing methods from Sanger to NGS in the past few years has enabled large-scale genome sequencing and evolutionary comparisons. However, the large size of many crop genomes is predominantly due to the amplification of repetitive elements as well as whole or partial genome duplication. One study on the repetitive elements of the rice genus found the percentage of repetitive elements in the samples to range from 25% of reads in Oryza coarctata (HHKK) to 66% of reads in Oryza officinalis (CC) (Zuccolo et al., 2007). A similar study for maize estimated that the repetitive content ranges from 64 to 73% (Meyers et al., 2001).

The ability to sequence the DNA of an organism has become one of the most important tools in modern biological research. Until recently, the sequencing of even small model genomes required substantial funds and international collaboration. The development of ‘second-generation’ sequencing technology has increased the throughput and reduced the cost of sequence generation by several orders of magnitude. These new methods produce vast numbers of relatively short reads, usually at the expense of read accuracy. The first approach to second-generation sequencing was pyrosequencing, and the first successful pyrosequencing system was developed by 454 Life Sciences and commercialized by Roche as the GS20. This was capable of sequencing over 20 million base pairs (Margulies et al., 2005). During 2007, the GS FLX model replaced the GS20 which was capable of producing over 100 million base pairs of sequence in a similar
amount of time. Due to recent advancements in this technology, the production has been increased up to 400 Mbp with the introduction of the Titanium chemistry. Several other companies including Illumina, Applied Biosystems, Helicos Biosciences and Pacific Biosciences have joined the competition (Imelfort and Edwards, 2009).

Using current Sanger sequencing technology, it is technically possible for up to 384 sequences of between 600 and 1,000 nucleotides in length to be sequenced in parallel. The 454 Genome Sequencer (GS) platform was the first of the new high-throughput sequencing platforms on the market when released in October 2005. It is based on the pyrosequencing approach. The current 454/Roche GS FLX Titanium platform makes it possible to sequence about 1.5 million such beads in a single experiment and to determine sequences of length between 300-500 nt. The reversible terminator technology used by the Illumina Genome Analyzer employs the sequencing by synthesis concept that is most similar to that used in Sanger sequencing i.e. the incorporation reaction is stopped after each base, the label of the base incorporated is read out with fluorescent dyes and the sequencing reaction is then continued with the incorporation of the next base.

SOLiD sequencing platform was the third new high-throughput system. The SOLiD system currently allows sequencing of more than 300 million beads in parallel, with a typical read length of between 25 and 75 nt. Helicos was the first company to sell a sequencer able to sequence individual molecules instead of molecule ensembles created by an amplification process. Single-molecule sequencing has the advantage that it is not affected by biases or errors introduced in a library preparation or amplification step, and may facilitate sequencing of minimal amounts of input DNA. While on the one end of the spectrum instrument throughput increases, some vendors also recently started to offer budget versions of their instruments (e.g. Illumina Genome Analyzer IIe, Illumina MySeq or 454/Roche GS Junior) with lower sequencing capacity. The different sequencing platforms differ in several ways, such as read length and sequencing chemistry (Table 1.1).

**Table 2.1: Technical properties of the next generation platforms**

<table>
<thead>
<tr>
<th>Properties</th>
<th>Roche/454</th>
<th>ABI/SOLiD</th>
<th>Solexa/Illumina</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequencing chemistry</td>
<td>Pyrosequencing</td>
<td>Sequencing by Ligation</td>
<td>Bridge amplification</td>
</tr>
<tr>
<td>Read length (bp)</td>
<td>800</td>
<td>75</td>
<td>2x101</td>
</tr>
<tr>
<td>Number of reads</td>
<td>1 million</td>
<td>5 billion</td>
<td>3 &amp; 6 billion (single and pair end)</td>
</tr>
</tbody>
</table>
The Solexa sequencing platform was commercialised in 2006, with Illumina acquiring Solexa in early 2007. The Solexa Genome Analyzer (currently the GAIIx) system uses reversible terminator chemistry to generate up to 50,000 million bases of usable data per run. On a flow cell surface, the sequencing templates are immobilized and clusters of identical copies of each DNA molecule are created due to solid phase amplification. Sequencing then uses four proprietary fluorescently labeled nucleotides to sequence the millions of clusters on the flow cell surface. These nucleotides possess a reversible termination property, allowing each cycle of the sequencing reaction to occur simultaneously in the presence of the four nucleotides. Illumina sequencing has been developed predominantly for re-sequencing, with more than 10-fold coverage usually ensuring high confidence in the determination of genetic differences. In addition, each raw read base has an assigned quality score, assisting assembly and sequence comparisons. The GAII is used for transcriptome profiling (RNA Seq) and Chromatin Immuno Precipitation sequencing (ChIP Seq) and the relatively low error rate of this system supports de novo sequencing applications. This technology has now been applied for the sequencing of several crop species including cucumber (Huang et al., 2009).

The AB SOLiD System (currently version 3) enables parallel sequencing of lonally amplified DNA fragments linked to beads. The method is based on sequential ligation with dye labelled oligonucleotides and can generate more than 20 Gb of mapable data per run. The system can be used for tag-based applications such as gene expression and Chromatin Immuno Precipitation sequencing, where a large number of reads are required and where the high throughput provides greater sensitivity for the detection of lowly expressed genes. The (ABI) SOLiD system is predominantly used for re-sequencing where comparison to a reference and high sequence redundancy enables the identification and removal of erroneous sequence reads.

In contrast to the short-read sequencers, the Roche 454 FLX system produces read lengths on average 300–500 bp and is capable of producing over 400 Mbp of sequence with a single-read accuracy of >99.5%. Amplification and sequencing is performed in a
Review of Literature

highly parallelized picoliter format. Emulsion PCR enables the amplification of a DNA fragment immobilized on a bead from a single fragment to 10 million identical copies, generating sufficient DNA for the subsequent sequencing reaction. Sequencing involves the sequential flow of both nucleotides and enzymes over the beads, which converts chemicals generated during nucleotide incorporation into a chemiluminescent signal that can be detected by a CCD camera. The light signal is quantified to determine the number of nucleotides incorporated during the extension of the DNA sequence. Newbler software has been developed by 454 life sciences specifically to assemble this type of data and it has been successfully applied for the assembly of a bacterial genome (Poly et al., 2007).

In September 2015, Thermo Fisher Scientific launched their latest Next Generation Sequencers (NGS), the Ion S5 and Ion S5 XL. The systems employ Ion Torrent semiconductor technology on multiple chip formats that enables to run multiple research applications such as sequencing of gene panels, custom assays, small genomes, exomes and transcriptomes – all on a single instrument. The S5 systems are not replacing their Ion siblings, the Ion PGM and Ion Proton. Instead, they span the gap between the two. This is made possible because of the three new chips that are on offer; Ion 520, Ion 530 and Ion 540. The Ion 520 chip performs similarly to the Ion PGM 318 chip with 200 bp or 400 bp reads, while an Ion 530 chip delivers three times the output of a 318 chip, again with 200 bp or 400 bp reads. The top-end Ion 540 chip is most similar to a Proton chip and supports 200 bp reads.

The different chip configurations enables to run both smaller panels and larger panels on the same instrument, generating outputs of 3-80 million reads with run times of just 2.5-4 hours, depending on the read length. The extra computing power of the Ion S5XL is realized during the data analysis stage that can be completed in just 1-5 hours, which is 3-4X faster than the Ion S5.

2.2.4 Genome Sequenced of Various Plants

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. 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).

The availability of complete genome sequence of Arabidopsis thaliana (Arabidopsis Genome Initiative, 2000) and rice (Goff et al., 2002; Yu et al., 2002) has provided a reference platform for plant genomics. In depth analysis of the known and predicted coding sequences has provided an invaluable resource on the gene contents for these model plants and covers the basic sample of gene repertoire needed for dicotyledonous and monocotyledonous plants. However, there exists a tremendous amount of biological diversity among different plant species that require more even sampling of other plant genome sequences to understand the diversity of gene content and basic functional resolution of the plant genomes in general. 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) and 250,000 Mbp (Pislotum nudum) (Obermayer et al., 2002).

The flowering plant Arabidopsis thaliana is an important model system for identifying genes and determining their functions. Its genome sequence was published in 2000 by the Arabidopsis Genome Initiative and it was the first plant to be sequenced. The sequenced regions covered 115.4 Mb of 125 Mb genome and extended into centromeric regions. The evolution of Arabidopsis involved whole genome duplication, followed by subsequent gene loss and extensive local gene duplications. The genome contains 25,498 genes encoding proteins from 11,000 families, similar to the functional diversity of Drosophila and C. elegans.

Lee et al. (2016) carried out the complete chloroplast genome sequences of Epimedium koreanum, which is a perennial medicinal plant distributed in Eastern Asia, by de novo assembly using whole genome next generation sequences. The chloroplast genome of E. koreanum was 157,218 bp in length and separated into four distinct regions such as large single copy region (89,600 bp), small single copy region (17,222 bp) and a pair of inverted repeat regions (25,198 bp). The genome contained a total of 112 genes including 78 protein-coding genes, 30 tRNA genes and 4 rRNA genes. Phylogenetic analysis with the reported chloroplast genomes revealed that E. koreanum is most closely related to Berberis bealei, a traditional medicinal plant in the Berberidaceae family.
Prabhudas et al. (2016) performed shallow whole genome sequencing for the assembly of complete chloroplast genome sequence of Arachis hypogaea L. The size of the complete CP genome of A. hypogaea L. was found to be 156,391 bp. The genome coverage was calculated to be 2122x with 3,863,475 quality filtered reads mapped to the assembled CP genome. The CP genome exhibited a quadripartite structure consisting of LSC and SSC regions of 85,946 bp and 18,797 bp respectively, with a pair of inverted repeats (IRa and IRb) of 25,824 bp each separating them. The overall GC content of the complete chloroplast genome was 36.4% and the individual GC content for LSC, SSC and IRs was 33.8%, 30.2%, and 42.8% respectively. A total of 110 genes were annotated including 76 protein coding genes, 30tRNA genes and 4rRNA.

Lee et al. (2015) carried out the complete chloroplast genome of Glehnia littoralis F. Schmidt ex Miq. by de novo assembly using whole genome sequencing data. It is an oriental medicinal herb belonging to Apiaceae family. The chloroplast genome of G. littoralis was 1,47,467 bp in length and divided into four distinct regions: large single copy region (93,493 bp), small single copy region (17,546 bp) and a pair of inverted repeat regions (18,214 bp). A total of 114 genes including 80 protein-coding genes, 30 tRNA genes and 4 rRNA genes were predicted and accounted for 57.1% of the chloroplast genome.

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.

Chen et al. (2014) estimated the size of the haploid Aquilaria agallocha genome to be approximately 736 Mb by flow cytometry. Five DNA libraries were constructed for a total of 144.3 Gb, which represented approximately 196X sequencing coverage. The DNA libraries contained one pair-end library with a fragment length of 300 bp and four mate-pair libraries of various fragment lengths. For the A. agallocha transcriptome, two RNA libraries were constructed to represent MJ treatment conditions (0 mM and 0.5 mM MJ) where polyA + RNA from in vitro shoot was sequenced for a total of 12.5 Gb. The DNA libraries were utilized in a de novo assembly procedure where the resulting draft

18
Review of Literature

The genome (NCBI BioProject: PRJNA240626) contained 28,482 scaffolds with an N50 of 126.4 kb, a 1.3 Mb longest sequence and a total size of 728.5 Mb, approximately 98% genome coverage.

Hirakawa et al. (2014) dissected the eggplant genome and built a draft genome dataset with 33,873 scaffolds termed SME_r2.5.1 that covered 833.1 Mb, ca. 74% of the eggplant genome. Approximately 90% of the gene space was estimated to be covered by SME_r2.5.1 and 85,446 genes were predicted in the genome. Clustering analysis of the predicted genes of eggplant along with the genes of three other solanaceous plants as well as Arabidopsis thaliana revealed that, of the 35,000 clusters generated, 4,018 were exclusively composed of eggplant genes that would perhaps confer eggplant-specific traits. Between eggplant and tomato, 16,573 pairs of genes were deduced to be orthologous and 9,489 eggplant scaffolds could be mapped onto the tomato genome.

Kang et al. (2014) constructed a draft genome sequence of mungbean to facilitate genome research into the subgenus Ceratotropis. Based on the de novo assembly of additional wild mungbean species, the divergence of what was eventually domesticated and the sampled wild mungbean species appears to have predated domestication. We prepared five libraries for sequencing by Illumina HiSeq2000 including 180 and 500 bp paired-end libraries and 5, 10 and 40 kb mate-pair libraries. These libraries provided a 320-fold base pair coverage of the estimated genome size. In addition, long reads providing approximately five fold genome coverage were produced by sequencing using GS FLXþ. The short reads were assembled using ALLPATHS-LG software, producing 2,800 scaffolds with an N50 length of 1,507 kb. The total length of the scaffolds was 843 Mb. The long reads generated by GS FLXþ were assembled into 180,372 contigs using Newbler 2.5.3 software. In total, 144,213 of the GS FLX contigs were consistent with the scaffolds from ALLPATHS-LG. The non-matched GS FLXþ contigs were divided into 5 kb pseudo-mate-pair reads and assembled using ALLPATHS-LG software to improve the quality of the assembly, resulting in 2,748 scaffolds with N50 length of 1.52 Mb. The total length of the produced scaffolds was 431 Mb, representing 80% of the genome size of 543 Mb estimated from 25-base kmer frequency distribution.

Kim et al. (2014) reported genome sequence of hot pepper providing insights into the evolution of pungency in Capsicum species. Whole-genome sequencing and assembly of the hot pepper (Mexican landrace of Capsicum annuum cv. CM334) was done with 186.6x coverage. They suggested integrative genomic and transcriptomic analysis that change gene expression and neofunctionalization of capsaicin synthase, which shaped
capsaicinoid biosynthesis. They found differential molecular patterns of ripening regulators and ethylene synthesis in hot pepper and tomato.

Moghe et al. (2014) sequenced the genome of wild radish (Raphanus raphanistrum). The first step in generating a draft assembly for the R. raphanistrum genome was the estimation of genome size using flow cytometry. The estimated size of 515 Mb was comparable to genome size estimates of related species, including Brassica (529 Mb), Brassica oleraceae (696 Mb) and R. sativus (573 Mb). They sequenced the genome of a 5th generation inbred plant and the reads were assembled with a hybrid approach. The final assembly size was 254 Mb, representing 49.3% of the estimated genome size, with N50 of 10.1 kb.

Polashock et al. (2014) estimated the genome size of Vaccinium macrocarpon to be about 470 Mb. Genomic sequences were assembled into 229,745 scaffolds representing 420 Mbp (N50 = 4,237 bp) with 20X average coverage. The number of predicted genes was 36,364 and represented 17.7% of the assembled genome. Of the predicted genes, 30,090 were assigned to candidate genes based on homology. Genes supported by transcriptome data totaled 13,170 (36%).

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 Chiltepin (C. annuum var. glabriusculum). Results show that the pepper genome expanded ~0.3 million years ago 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.

Wang et al. (2014) 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 2.1Mb, containing an estimated 27,148 genes. Candidate genes and oil biosynthetic pathways contributing to high oil content were discovered by comparative genomic and transcriptomic analysis. It 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.
The first whole-genome sequences of 'DC-27' carrot were de novo assembled and analyzed by Xu et al. (2014). Transcriptomic sequences of 14 carrot genotypes were downloaded from the Sequence Read Archive (SRA) database of National Center for Biotechnology Information (NCBI) and mapped to the whole-genome sequence before assembly. The first web-based genomic and transcriptomic database for *D. carota* (CarrotDB) was developed on the basis of these data sets by which users can search certain target genes and simple sequence repeats along with designed primers of 'DC-27'. Assembled transcriptomic sequences along with fragments per kilobase of transcript sequence per millions base pairs sequenced information (FPKM) of 14 carrot genotypes were also provided. CarrotDB offers the tools of Genome Map and Basic Local Alignment Search Tool. In the entire genome sequences, a total of 2,826 transcription factor (TF) genes, classified into 57 families, were identified.

Chen et al. (2013) reported whole-genome sequencing of *Oryza brachyantha*, revealing mechanisms underlying *Oryza* genome evolution. The 261 Mb de novo assembled genome was sequenced 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 predicted 32,038 protein-coding genes in the *Oryza brachyantha* genome, of which only 70% were located in collinear positions in comparison with the rice genome.

Fu et al. (2013) developed markers, de novo assembly and gene annotation in celery (*Apium graveolens* L.) by using Illumina paired-end transcriptome sequences. 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 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.

Guo et al. (2013) performed de novo sequencing of watermelon by utilizing the Illumina platform, resulting in 46.18 Gbp reads, corresponding to 108.6x coverage of an estimated 425 Mbp genome size of this species. Subsequently, a total of 23,440 protein-coding genes were identified using ab initio predictions, cDNA/EST- and homology-
mapping methods. Furthermore, 20 watermelon accessions were resequenced following the paired-end Illumina strategy. Among them 6,784,860 candidate SNP and 965,006 small indels were identified, representing a germplasm biodiversity that can contribute to the species plant breeding. Additionally, the comparative analyses of the transcriptome data should contribute to the understanding of the genetic diversity and molecular mechanisms underlying some biological processes in watermelon populations. Thus, the evolutionary scenario proposed in this study should shed light on the genetic backgrounds of the modern cultivars.

Jain et al. (2013) generated the draft sequence of a desi-type chickpea genome using next generation sequencing platforms, bacterial artificial chromosome end sequences and a genetic map. They generated 13.354 Gb of high-quality sequence data for chickpea ICC4958 by sequencing whole genome shotgun (WGS) libraries and mate-pair (MP) libraries of 3-20 kb insert sizes using a 454/Roche GS FLX Titanium platform. The 520 Mb assembly covered 70% of the predicted 740 Mb genome length and more than 80% of the gene space. Genome analysis predicted the presence of 27,571 genes and 210 Mb as repeat elements. The gene expression analysis performed using 274 million RNA-Seq reads identified several tissue-specific and stress-responsive genes.

Varshney et al. (2013) obtained 153.01 Gb of sequence data for chickpea, representing 207.32x genome coverage, by Illumina sequencing of 11 genomic libraries with insert sizes ranging from 180 bp to 20 kb. After filtering, 87.65 Gb of high-quality sequence data were assembled into 544.73 Mb of genomic sequence scaffolds with 50% of all bases in scaffolds larger than 645.3 kb (N50) and a maximum of 6.17 Mb. Based on k-mer statistics, the chickpea genome was estimated to be 738.09 Mb in size, which indicates that 73.8% of the genome was captured in scaffolds. They predicted a non-redundant set of 28,269 gene models, with average transcript and coding sequence sizes of 3,055 bp and 1,166 bp, respectively.

Yang et al. (2013) presented the reference genome sequence (241 Mb) of Eutrema salsugineum at 8 coverage sequenced using the traditional Sanger sequencing-based approach with comparison to its close relative Arabidopsis thaliana. The E. salsugineum genome contained 26,531 protein-coding genes and 51.4 % of its genome was composed of repetitive sequences that mostly reside in pericentromeric regions. Comparative analysis of the genome structures, protein-coding genes, microRNAs, stress-related pathways and estimated translation efficiency of proteins between E. salsugineum and A. thaliana suggested that halophyte adaptation to environmental stresses may occur via
global network adjustment of multiple regulatory mechanisms. The *E. salsugineum* genome provides a resource to identify naturally occurring genetic alterations contributing to the adaptation of halophytic plants to salinity and that might be bioengineered in related crop species.

Brenchley et al. (2012) reported the analysis of the bread wheat genome using whole-genome shotgun sequencing. 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 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.

Iorizzo et al. (2012) used the sequencing data from a carrot 454 whole genome library to develop a *de novo* assembly of the mitochondrial genome. Genome annotation allowed identification of 44 protein coding genes. PCR amplification and sequence analysis across different Apiaceae species revealed consistent conservation of the fragment in the mitochondrial genomes and an insertion in *Daucus* plastid genomes, giving evidence of a mitochondrial to plastid transfer of DNA. This study confirmed that whole genome sequencing is a practical approach for *de novo* assembly of higher plant mitochondrial genomes.

Krishnan et al. (2012) sequenced genome and transcriptomes of *Azadirachta indica* using multiple sequencing platforms and libraries. The *A. indica* genome is AT-rich, bears few repetitive DNA elements and comprises about 20,000 genes. The molecular phylogenetic analysis grouped *A. indica* together with *Citrus sinensis* from the Rutaceae family validating its conventional taxonomic classification. Comparative transcript expression analysis showed either exclusive or enhanced expression of known genes involved in neem terpenoid biosynthesis pathways compared to other sequenced angiosperms. Various combinations of sequencing libraries were used to obtain the best scaffold N50s and N90s using SOAPdenovo. The *de novo* assembly of *A. indica* genome produced a scaffold N50 length of 452,028 bp with corresponding scaffold N90 length of 56,222 bp, contig N50 length of 740 bp and contig N90 length of 172 bp.

Kumbo et al. (2012) sequenced and assembled a draft genome of *Gossypium raimondii*. Over 73% of the assembled sequences were anchored on 13 *G. raimondii*
chromosomes. The genome contained 40,976 protein coding genes, with 92.2% of these further confirmed by transcriptome data. A whole-genome shotgun strategy was used to sequence and assemble the G. raimondii genome. A total of 78.7 Gb of next-generation Illumina paired-end 50 bp, 100 bp and 150 bp reads was generated by sequencing genome shotgun libraries of different fragment lengths (170 bp, 250 bp, 500 bp, 800 bp, 2 kb, 5 kb, 10 kb, 20 kb and 40 kb) that covered 103.6 fold of the 775.2 Mb assembled G. raimondii genome. The assembly performed by SOAPdenovo consisted of 41,307 contigs and 4,715 scaffolds and accounted for approximately 88.1% of the estimated G. raimondii genome. Over 73% of the assembly was in 281 chromosome-anchored scaffolds, with 228 of them both anchored and oriented.

Mas et al. (2012) reported the genome sequence of melon. They assembled 375 Mb of the double-haploid line DHL92, representing 83.3% of the estimated melon genome. A whole-genome shotgun strategy based on 454 pyrosequencing was used, producing 14.8 million single shotgun and 7.7 million paired-end reads. Both 454 and Sanger reads were assembled with Newbler 2.5 into 1,594 scaffolds and 29,865 contigs, totaling 375 Mb of assembled genome. The N50 scaffold size was 4.68 Mb and 90% of the assembly was contained in 78 scaffolds. They predicted 27,427 protein-coding genes which was analyzed by reconstructing 22,218 phylogenetic trees, allowing mapping of the orthology and paralogy relationships of sequenced plant genomes.

Prochnik et al. (2012) obtained the draft genome sequence of cassava by a whole genome shotgun (WGS) strategy. The cassava genome spans an estimated 770 Mb in N=18 chromosomes. A total of 22.4 billion bp of raw sequence data was generated, enough to cover the genome ~29 times. These reads were assembled using Roche’s GS de novo assembler (Newbler) v.2.5 into 12,977 scaffolds that span a total of 532.5 Mb. Although the genome assembly is in nearly 13,000 pieces, half of it is captured in only 487 scaffolds, each longer than 258 kbp and containing 49 or more genes.

Wua et al. (2012) presented the draft sequence of the Thellungiella salsuginea genome, assembled based on ~134-fold coverage to seven chromosomes with a coding capacity of at least 28,457 genes. They sequenced the genome of T. salsuginea (Shandong ecotype) using the paired-end Solexa sequencing method (Illumina GA II system). Based on flow cytometry of isolated nuclei stained with propidium iodide, the expected genome size was ~260 Mb. Thus, with a total of 34.8 Gb of high-quality sequences, the genome was covered ~134-fold. The final length of the assembled sequences amounted to ~233.7 Mb, covering about 90% of the estimated genome size. The assembly consisted of 2,682
scaffolds, the 10 longest of which range from 1.9-6.8 Mb and represent 17% of the assembled genome.

Zhang et al. (2012) used whole genome shotgun combined with next-generation sequencing to assemble a draft genome of the foxtail millet strain ‘Zhang gu’. The final genome assembly was 423 Mb, ~86% of the estimated genome size, with repeats comprising ~46% of the genome. They annotated 38,801 protein-coding genes, of which ~81% were expressed. Key chromosome reshuffling events were detected through collinearity identification between foxtail millet, rice and sorghum including two reshuffling events fusing rice chromosomes 7 and 9, 3 and 10 to foxtail millet chromosomes 2 and 9, respectively, that occurred after the divergence of foxtail millet and rice, and a single reshuffling event fusing rice chromosome 5 and 12 to foxtail millet chromosome 3 that occurred after the divergence of millet and sorghum.

Argout et al. (2011) sequenced and assembled the draft genome of *Theobroma cacao*, an economically important tropical-fruit tree crop that is the source of chocolate. They used the Roche/454 and Sanger raw data to produce the assembly. This represented ×16.7 coverage of the 430 Mb genome of B97-61/B2, whose size was estimated by flow cytometry. This assembly, performed with Newbler software, consisted of 25,912 contigs and 4,792 scaffolds. Eighty percent of the assembly was in 542 scaffolds, and the largest scaffold measured 3.4 Mb. They determined the N50 (the scaffold size above which 50% of the total length of the sequence assembly can be found) to be 473.8 kb. The total length of the assembly was 326.9 Mb, which represents 76% of the estimated genome size of the *T. cacao* genotype B97-61/B2 (430 Mb). In addition, they used a high coverage of Illumina data (×44 coverage of the genome), which had a different error profile than 454 data, to improve accuracy of the *T. cacao* genome sequence. This assembly corresponds to 76% of the estimated genome size and contains almost all previously described genes, with 82% of these genes anchored on the 10 *T. cacao* chromosomes.

Dassanayake et al. (2011) presented the draft genome for *Thellungiella parvula*. By the next generation sequencing, they obtained the de novo assembled genome in 1,496 gap-free contigs, closely approximating the estimated genome size of 140 Mb. They anchored these contigs to seven pseudo chromosomes without the use of maps and showed that short reads can be assembled to a near-complete chromosome level for a eukaryotic species lacking prior genetic information. The sequence identifies a number of tandem duplications that, by the nature of the duplicated genes, suggest a possible basis for *T. parvula’s* extremophile lifestyle. Our results provide essential background for
developing genomically influenced testable hypotheses for the evolution of environmental stress tolerance.

Sato et al. (2011) sequenced the whole genome of *Jatropha curcas* L. using a combination of the conventional Sanger method and new-generation multiplex sequencing methods. Total length of the non-redundant sequences thus obtained was 285,858,490 bp consisting of 120,586 contigs and 29,831 singlets. They accounted for 95% of the gene-containing regions with the average G+C content was 34.3%. A total of 40,929 complete and partial structures of protein encoding genes have been deduced. Comparison with genes of other plant species indicated that 1529 (4%) of the putative protein-encoding genes are specific to the Euphorbiaceae family. A high degree of microsynteny was observed with the genome of castor bean and, to a lesser extent, with those of soybean and *Arabidopsis thaliana*. In parallel with genome sequencing, cDNAs derived from leaf and callus tissues were subjected to pyrosequencing, and a total of 21,225 unigene data have been generated. Polymorphism analysis using microsatellite markers developed from the genomic sequence data obtained was performed with 12 *J. curcas* lines collected from various parts of the world to estimate their genetic diversity.

Varshney et al. (2011) used the Illumina next generation sequencing platform to generate 237.2 Gb of sequence, which along with Sanger-based bacterial artificial chromosome end sequences and a genetic map, they assembled into scaffolds representing 72.7% (605.78 Mb) of the 833.07 Mb pigeonpea genome. Filtering and correction of the sequence data for very small and/or bad-quality sequences yielded 130.7 Gb of high-quality sequence, ~163.4× coverage of the pigeonpea genome. Genome analysis predicted 48,680 genes for pigeonpea and also showed the potential role that certain gene families, for example, drought tolerance–related genes, have played throughout the domestication of pigeon pea and the evolution of its ancestors. This reference genome sequence would facilitate the identification of the genetic basis of agronomically important traits, and accelerate the development of improved pigeon pea varieties that could improve food security in many developing countries.

Parchman et al. (2010) obtained 586,732 sequencing reads from a 454 GS XLR70 Titanium pyrosequencer (mean length: 306 base pairs) for *Pinus contorta*. A combination of reference-based and de novo assemblies yielded 63,657 contigs, with 239,793 reads remaining as singletons. Based on sequence similarity with known proteins, these sequences represent approximately 17,000 unique genes, many of which are well covered by contig sequences. This sequence collection also included a surprisingly large number
of retrotransposon sequences, suggesting that they are highly transcriptionally active in the tissues which are sampled. To assess the coverage and quality of the assembly, BLASTx was used to align both contigs and singletons to the UniRef50 15.4 and the TAIR9 Arabidopsis thaliana annotated protein databases using an E value threshold of $10^{-6}$. Of 63,687 contigs, 20,301 (32%) had BLAST hits to known proteins in UniRef50 and matched 8,316 unique protein accessions. High quality PCR primers were designed for a substantial number of the SSR loci, and a large number of these were amplified successfully in initial screening.

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) estimated 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.

Sato et al. (2008) reported structural features of the Lotus japonicus genome. The 315.1 Mb sequences determined in this and previous studies correspond to 67% of the genome (472 Mb), and are likely to cover 91.3% of the gene space. Linkage mapping anchored 130-Mb sequences onto the six linkage groups. A total of 10,951 complete and 19,848 partial structures of protein-encoding genes were assigned to the genome. Comparative analysis of these genes revealed the expansion of several functional domains and gene families that are characteristic of L. japonicus. Synteny analysis detected traces of whole-genome duplication and the presence of syntenic blocks with other plant genomes to various degrees. Functional annotation for deduced L. japonicus proteins was performed by a similarity search against the genes of known function and a domain analysis. The similarity search was performed using the gapped BLASTP function of BLAST against the UniProtKB database and protein-encoding genes deduced in A. thaliana, rice, Populus trichocarpa and grapevine. For the analysis of gene families and functional domains, the predicted proteome was searched against InterPro.

Wicker et al. (2008) generated 574 Mbp of Illumina/Solexa sequences from barley total genomic DNA, representing about 10% of a genome equivalent. From these
sequences they generated a Mathematically Defined Repeat (MDR) index which was then used to identify and mark repetitive regions in the barley genome. Comparison of the MDR plots with expert repeat annotation drawing on the information already available for known repetitive elements revealed a significant correspondence between the two methods. MDR-based annotation allowed for the identification of dozens of novel repeat sequences, though, which were not recognized by hand-annotation. The MDR data was also used to identify gene-containing regions by masking of repetitive sequences in eight de novo sequenced bacterial artificial chromosome (BAC) clones.

Velasco et al. (2007) estimated the genome size of *Vitis vinifera* to be 504.6 Mb. Genomic sequences corresponding to 477.1 Mb were assembled in 2,093 metacontigs and 435.1 Mb were anchored to the 19 linkage groups (LGs). The number of predicted genes was 29,585, of which 96.1% were assigned to LGs. This assembly of the grape genome provides candidate genes implicated in traits relevant to grapevine cultivation, such as those influencing wine quality, via secondary metabolites, and those connected with the extreme susceptibility of grape to pathogens. Single nucleotide polymorphism (SNP) distribution was consistent with a diffuse haplotype structure across the genome. Of around 2,000,000 SNPs, 1,751,176 were mapped to chromosomes and one or more of them were identified in 86.7% of anchored genes. In total 29,585 genes were predicted. BLAST searches against Uniprot and plant protein databases, annotated with GO terms, of various domain libraries were the base for gene annotations GO terms were extracted from BLAST searches against KEGG databases, KOBAS of metabolic pathways and InterproScan.

### 2.2.5 De Novo Genome Assembly Approaches

The first sequence fragment assembly algorithms were developed beginning in the late 1970s. Early sequencing efforts focused on creating multiple alignments of the reads to produce a layout assembly of the data. Finally, a consensus sequence could be read from the alignment and the DNA sequence was inferred from this consensus. This approach is referred to as the overlap-layout-consensus approach and culminated in a variety of software applications such as CAP3 (Huang and Madan, 1999). From the early to mid 1990’s, research began to focus on formalizing, benchmarking and classifying fragment assembly algorithm approaches.
2.2.6 Problems of Assembling Complex Genomes

One challenge of genome sequencing lies in the fact that only a small portion of the genome encodes genes, and that these genes are surrounded by repetitive DNA that are often difficult to characterize. Repeats can cause ambiguity with fragment assembly and thus pose the greatest challenge when assembling genomic data. The most robust method to overcome the problems of repeats when assembling shotgun reads is to increase the read length to a point that every repeat is spanned by at least one read. However, increasing sequence read length has proven technically problematic and most second-generation sequencing technologies compromise read length to deliver an increased number of reads.

Modifications of second-generation sequencing methods attempt to overcome the problem of short reads by producing read pairs, where two reads are produced with a known orientation and approximate distance between them, increasing the ‘effective’ read length. If the distance between the paired ends is large enough, repeats will be spanned by pairs of reads removing ambiguity from the assembly (Figure 2.1). It is expected that the ability to produce increased quantities of paired read sequence data, combined with custom bioinformatics tools, will provide the foundation to sequence large and complex plant genomes.

2.2.7 Bioinformatical Tools

Bioinformatics is a vast and complex multidisciplinary research area where numerous tools have been developed over the years to analyze constantly growing amounts of data. The results of an analysis are made available using various representations (e.g. HTML tables, XML files, images, etc.). In order to produce these representations, each result is converted into a generic category-specific model that is used by a renderer that generates the requested output. The renderers are specific to the model and not to the tool, and thus are available across all the tools in a category. The availability of multiple views of the same data helps the user to interpret and compare results from different tools within a category. Sequence search algorithms produce limited hits annotation. With the new framework it is possible to navigate hits and access related information.

While developments in sequencing technology make it possible to obtain large-scale sequence data in a short time, the assembly and analysis of sequences remains a challenging task. Thus, much of the effort in the recent years has been dedicated to develop and improve bioinformatics tools. Currently, a number of bioinformatics
software packages are available which are essentially used for different purposes, including alignment, assembly, annotation and sequence-variation detection (e.g., identification of SNP) (Imelfort and Edwards, 2009; Scheibye et al., 2009; Lerat, 2010; Paszkiewicz and Studholme, 2010; Bao et al., 2011).

The genome assembly process can be divided into two steps: draft assembly and assembly improvement (finishing). In the majority of the cases, 98% of the genome is covered by draft assembly, while this ratio is 5-fold lower in finished assemblies (Li and Durbin, 2009). Usually, before assembly, repetitive elements are identified and filtered out from the dataset. Repetitive elements are one of the challenging issues for assembly procedures as the majority of the gaps in an assembly are caused by repeated sequences (Cahill et al., 2010). Sequencing with longer reads emerges as a good way out. Paired-end sequencing is also commonly used for this purpose. Depending on availability, repetitive elements are computationally detected by homology searches to known repeat sequences. REPuter (Kurtz et al., 2001), Tandem Repeat Finder (Benson, 1999) and RepeatMasker <http://www.repeatmasker.org> are among the most common programs for detecting such repetitive elements. When there is a lack of a reference genome, repetitive elements are identified de novo. The basic workflow pipeline is composed of masking the known repeats, de novo repeat finding on the masked genome, and classification of the newly identified repeats. RECON (Bao and Eddy, 2002), Repeat Modeler <http://www.repeatmasker.org/RepeatModeler.html>, Repeat Scout (Price et al., 2005) and REPET (Flutre et al., 2011) are examples of the most acknowledged software packages for this purpose. Presently, a number of assembly approaches are applied on short-read assemblies. The first assemblers are based on a simple strategy known as the greedy algorithm, which is an implementation of finding the shortest common super sequence (Narzisi and Mishra, 2011).

The algorithm proceeds as follows: i) pairwise comparison of all sequences to identify overlapping sequences and merging the best overlapped sequences; and ii) these steps are repeated until no more sequences are left to be merged. The greedy algorithm has been used mainly for assembling small genomes. On the other hand, since the algorithm needs local information at each step, the presence of complex repeats may lead to mis-assemblies. There has been a number of sequence assembly tools published in recent years. Successive tools have increasingly improved in terms of both the quality of the output assemblies as well as robustness in the assembly of more complex genomes.
There has been a general trend of improving elements of the algorithm underlying the Eulerian approach.

The gene ontology (GO) is probably the most extensive scheme today for the description of gene product functions but also other systems such as enzyme codes, KEGG pathways, FunCat or COG are widely used within molecular databases. Many bioinformatics tools and methods have been developed to assist in the assignment of functional terms to gene products. Fewer resources, however, are available when it comes to the large-scale functional annotation of novel sequence data of non-model species, as would be specifically required in many plant functional genomics projects. Blast2GO is not a mere generator of functional annotations. The application includes a wide range of statistical and graphical functions for the evaluation of the annotation procedure and the final results. Especially, (relative) abundance of functional terms can be easily assessed and visualized. The Blast2GO annotation procedure consists of three main steps: blast to find homologous sequences, mapping to collect GO terms associated to blast hits, and annotation to assign trustworthy information to query sequences. Once GO terms have been gathered, additional functionalities enable processing and modification of annotation results (Conesa and Gotz, 2008).

![Figure 2.1: Schematic representation of Blast2GO application.](image)

GO annotations are generated through a 3-step process: blast, mapping, annotation. InterPro terms are obtained from InterProScan at EBI, converted and merged to GOs. GO annotation can be modulated from Annex, GOSlim web services and manual editing. EC and KEGG annotations are generated from GO. Visual tools include sequence color code, KEGG pathways, and GO graphs with node highlighting and filtering options.
Additional annotation data-mining tools include statistical charts and gene set enrichment analysis functions.

2.3 MOLECULAR MARKERS

Molecular markers have acted as versatile tools and have been effectively employed in diverse fields like taxonomy, physiology, embryology and genetic engineering. Molecular markers like Simple Sequence Repeats (SSR) can be used effectively utilized in plant breeding, MAS (marker assisted selection), mapping, fingerprinting, population genetics and phylogenetic studies. A challenge for genotypes research is that, despite the fact that the use of molecular markers for phylogenetic studies is well-established; very few studies have described the development of new markers for genotypes. Some of the findings related to molecular markers are reviewed as under.

2.3.1 Simple Sequence Repeats

Microsatellite or short tandem repeats or simple sequences repeats are monotonous repetitions of very short (one to five) nucleotide motifs, which occur as interspersed repetitive elements in all eukaryotic genomes (Tautz and Renz, 1984). Variation in the number of tandemly repeated units is mainly due to strand slippage during DNA replication where the repeats allow matching via excision or addition of repeats (Schlotterer and Tautz, 1992). As slippage in replication is more likely than point mutations, microsatellite loci tend to be hyper-variable. Microsatellite assays show extensive inter-individual length polymorphisms during PCR analysis of unique loci using discriminatory primers sets.

The PCR amplification protocols used for microsatellites employ loci-specific either unlabeled primer pairs or primer pairs with one radiolabelled or fluoro-labeled primer. Analysis of unlabeled PCR products is carried out using polyacrylamide or agarose gels. The employment of fluorescent labeled microsatellite primers and laser detection (e.g., automated sequencer) in genotyping procedures has significantly improved the throughput and automatisation. However, due to the high price of the fluorescent label, which must be carried by one of the primers in the primer pair, the assay becomes costly. Schuelke (2000) introduced a novel procedure in which three primers are used for the amplification of a defined microsatellite locus: a sequence-specific forward primer with M13(−21) tail at its 50 end, a sequence-specific reverse primer and the universal fluorescent-labeled M13(−21) primer which proved simple and less expensive. Microsatellites are highly popular genetic markers because of their
codominant inheritance, high abundance, enormous extent of allelic diversity, and the ease of assessing SSR size variation by PCR with pairs of flanking primers. The reproducibility of microsatellites is such that, they can be used efficiently by different research laboratories to produce consistent data.

Cheng et al. (2016) performed a comprehensive characterization of simple sequence repeats in pepper genomes providing valuable resources for marker development in Capsicum. They identified an average of 8,68,047.50, 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 1,13,500 in silico unique SSR primer pairs were 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 ranging from 2 to 6 and 0.05 to 0.64, respectively.

Chen et al. (2015) studied development and validation of EST-SSR markers in adzuki bean (Vigna angularis L.) performed by utilizing 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 mono-nucleotide 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%).

Hu et al. (2015) performed de novo assembly and characterization of fruit transcriptome in black pepper (Piper nigrum). 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.

Sahu et al. (2015) aimed to develop eSSRs and their further experimental validation and cross-transferability of these markers in different genera of the Apiaceae family to which Centella belongs. An in-house pipeline was developed for the entire analysis by combining bioinformatics tools and perl scripts. Out of 1776.5 kb of examined sequences, 417 (15.9%) ESTs containing 686 SSRs were detected with a density of one SSR per 2.59 kb. The gene ontology study revealed 282 functional domains involved in various processes, components, and functions, out of which 64 ESTs were found to have both SSRs and functional domains. Out of 603 designed EST-SSR primers, 18 pairs of primers were selected for validation based on the optimum parameter value.

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 which were assembled into 39,584 and 41,740 unigenes with mean lengths of 683 bp and 690 bp respectively. A total of 1,939 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.

Roomi et al. (2014) studied the molecular diversity study of seventy accessions of Arachis hypogaea. The DNA was extracted and primed with thirty SSR primers and amplified through PCR. Fifteen out of thirty primers generated polymorphic bands among the selected accessions. In total, forty different polymorphic loci were determined across the selected accessions. The range of number of polymorphic loci detected was ranged from 2 to 4 for each primer, with an average of 2.6 loci per primer. Polymorphic Index Content (PIC) value was calculated for each marker. The dendogram was constructed on the basis of allelic data from fifteen SSR markers across the selected accessions. All the accessions were divided into six clusters at 0.67 coefficients of similarity.

Wang et al. (2014) studied 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. 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 (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.

Wei *et al.* (2014) studied Simple Sequence Repeat (SSR) markers of Sesame (*Sesamum indicum* L.) by utilizing genome survey. 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 synthesized primer pairs, 218 polymorphic SSRs were developed and used to screen 31 sesame accessions from 12 countries.

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 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 by 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.

Zhang *et al.* (2012) studied *de novo* assembly and characterization of the transcriptome during seed development and generation of genic-SSR markers in peanut (*Arachis hypogaea* L.). 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 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.

Cavagnaro et al. (2011) studied microsatellite isolation and marker development in carrot 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.

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. About 90% and 88% of tested SSR and SNP primers amplified a product, of which 70% and 46%, respectively, were of the expected size.

Gautami et al. (2009) identified a novel set of SSR markers for germplasm analysis and interspecific transferability in Arachis hypogaea L. Primer pairs were designed for 23 SSR loci, of which 14 (61%) primer pairs yielded scorable amplicons. Eight (57%) primer pairs showed polymorphism among 23 groundnut genotypes. The PIC for polymorphic SSR ranged from 0.13 to 0.36 with an average 0.25.

Sui et al. (2009) developed genomic SSR and potential EST-SSR markers in Bupleurum chinense. Nineteen genomic SSR markers were developed using inter-simple sequence repeat (ISSR)-suppression PCR technique in Bupleurum chinense DC. which is a widely used Chinese medicinal plant. 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.
Chakravarthi and Naravaneni (2006) investigated the genetic diversity and DNA fingerprinting of 15 elite rice genotypes using 30 SSR primers on chromosome numbers 7-12. The results revealed that all the primers showed distinct polymorphism among the cultivars studied indicating the robust nature of microsatellites in revealing polymorphism. Based on this study, the larger range of similarity values for related cultivars using microsatellites provides greater confidence for the assessment of genetic diversity and relationships.

Aranzana et al. (2002) studied a set of 109 microsatellite primer pairs recently developed for peach and cherry in the almond × peach F2 progeny previously used to construct a saturated *Prunus* map containing mainly restriction fragment length polymorphism markers. All but one gave amplification products, and 87 (80%) segregated in the progeny and detected 96 loci. The resulting *Prunus* map contains a total of 342 markers covering a total distance of 522 cm. The approximate position of nine additional simple sequence repeats (SSRs) was established by comparison with other almond and peach maps. SSRs were placed in all the eight linkage groups of this map, and their distribution was relatively even, providing a genome-wide coverage with an average density of 5.4 cm/SSR. Twenty-four single-locus SSRs, highly polymorphic in peach, and each falling within 24 evenly spaced approximately 25 cm regions covering the whole *Prunus* genome, are proposed as a ‘genotyping set’ useful as a reference for fingerprinting, pedigree and genetic analysis of these species.