4-30-2015

Assembly and Annotation of the Common Walnut (Juglans regia) Transcriptome

Jeanne F. Whalen
University of Connecticut - Storrs, jeanne.whalen@uconn.edu

Assembly and Annotation of the Common Walnut (*Juglans regia*) Transcriptome

Jeanne Fraher Whalen

B.S., University of Connecticut, 2014

A Thesis

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Science

At the

University of Connecticut

2015

APPROVAL PAGE

Masters of Science Thesis

Assembly and Annotation of the Common Walnut (*Juglans regia*) Transcriptome

Presented by

Jeanne Fraher Whalen, B.S.

Major Advisor_____
Rachel O'Neill

Associate Advisor_____
Jill Wegrzyn

Associate Advisor_____
John Malone

Associate Advisor_____
Yaowu Yuan

University of Connecticut
2015

Table of Contents

I. Introduction

II. Background

1. History
2. Description and Uses
3. Genetic Resources
4. Applications
5. Sequencing Methods
6. Transcriptome Sequencing
7. Quality Control and Trimming
8. *De novo* Transcriptome Assembly
9. Open Reading Frame Prediction
10. Annotation
11. Genome Alignment

III. Methods

1. Sequencing and Quality Control
2. Assembly
3. Annotation
4. Genome Alignment
5. Orthologous Gene Families

IV. Results

1. Sequencing and Quality Control

2. Assembly

3. Annotation

4. Genome Alignment

5. Orthologous Gene Families

V. Discussion

VI. References

Abstract

The Common walnut (or Persian walnut), *Juglans regia*, is native to the region spanning the Balkans, Himalayas and southwest China with the largest native populations in forests through Kyrgyzstan. Following its dispersal to Western Europe and Southern Europe, and later introduction into North America, the walnut tree is now grown commercially in several countries with the largest producers being the United States, China, Iran, and Turkey. Walnuts are of great economical value in the United States and globally. In 2011, walnuts were valued at $1.35 billion in California. Worldwide, production is increasing, currently set at 1.7 million tons of in-shell walnuts per year. The walnut stone fruit is a source of nutrient-rich food as well as a source of rare antioxidants. To understand the genetic diversity and improve existing breeding programs, a study to characterize both the genome and the transcriptome is underway. Both will provide insight on concerns facing the walnut industry in the United States including drought and disease resistance as well as other conditions for optimal seed growth. In this study, the transcriptome is assembled *de novo* from a combination of 19 different tissues. This comprehensive tissue sampling includes reproductive tissues as well as roots, leaves, flower, and fruit. Varied developmental stages are also included in the RNA libraries. Paired-end 85 bp sequencing was performed on individual libraries with the Illumina Genome Analyzer II. The transcriptome was assembled independently of the concurrent genome sequencing of *J. regia*. The Trinity assembled transcriptome which defined the genes and transcripts in each library was annotated with a combination of open-source tools: USearch, BLAST2GO and InterProScan. Individual assemblies ranged from 17,257 to 23,666, with a combined assembly of all reads yielding 29,785 unique genes. Functional annotations were provided through a combination of sequence similarity searches, Gene Ontology term assignment and identification of orthologous genes

families. All sequences queried were assigned at least one gene ontology (100%), with the top 20 ontologies being involved in important regulatory processes. Tribe-MCL analysis identified 10,092 families among the ten species compared. A total of 31 gene families were unique to the common walnut transcriptome.

1. Introduction

The common walnut, *Juglans regia*, is one of the oldest cultivated trees and is of great importance worldwide. This diploid species is wind-pollinated and self-fertile. Common walnut is of economic value, not only for its nutrient rich food source, but also for medicinal and hardwood properties. Since its migration from central and east Asia, it has been grown commercially worldwide in temperate climates. Walnut breeders are interested in increasing the yield of young trees, expanding the range of harvest dates and reducing the need for chemical inputs for resistance to bacteria and viral infections.

As is the case with many economically important crop species, there has been increasing interest in developing genomic resources to support and improve upon existing breeding programs. Illumina short read RNA sequencing was utilized to develop a comprehensive transcriptome resource to enhance the on-going efforts to assemble a full genome sequence. The transcriptome is the full set of expressed genes in an organism necessary to characterize the gene space. Samples from 19 tissues of a single cultivar were used in this analysis. Here, the individual libraries as well as the combined read set was *de novo* assembled. The resulting transcripts were analyzed for completeness, functional assignment, and unique qualities of walnut. This resource will provide the foundation for future expression studies that may shed light on areas of interest, such as disease resistance, cold hardiness, and optimal seed production.

2. Background

2.1 History

Thriving in a temperate climate, *Juglans regia* is a member of the family Juglandaceae. They are native to central Asia and southwest China. In the fourth century BC, Alexander the Great introduced them to southern Europe. During Roman times, they were brought to western

and northern Europe before eventually being brought to Turkey by way of the Silk Road. Later, the common walnut was introduced to the Americas by English colonists and became an important food source in the New World (Sze-Tao, 2000). Walnut is now grown in southeastern Europe, Eastern Asia and North America.

2.2 Description and Uses

Juglans regia is a flowering, nut producing tree which has an equal height and spread of about 40-60ft. It is intolerant of shade, requiring full sun to grow, but tolerant of drought. Female flowers produce nuts encased in a smooth green husk which reach maturity in autumn. The walnut seed is highly desired as a high-energy source of nutrients. The seed contains more than 15% protein, by dry weight, and has profuse polyunsaturated fatty acids (Sze-Tao and Kathe, 2000). Additionally, they are a rich source of iron, potassium, and various vitamins, including B6 and E (Sze-Tao and Kathe, 2000). Studies support that consumption can reduce the risk of high cholesterol and protect against heart disease (Ozkan, 2005). With proper cultivation, the trees will begin to produce small seeds in the fourth year. After 20 years of production, they typically reach optimum harvest, producing the greatest yield of seeds. An orchard will produce an average of 5400 lbs per acre (Edstrom et al. 2012). The United States, particularly California, and China together contribute to over 80% of the walnut production (Halstead, 2014). Globally, walnut production is second in nut production behind almonds (Sze-Tao and Kathe, 2000). The production of walnuts continues to grow rapidly with output more than doubling in the last ten years (Halstead, 2014). Production is predicted to continue to expand and global exports are forecasted to rise by 6% (Halstead, 2014).

Outside of seed production, *Juglans regia* is also grown commercially as a hardwood for furniture and saw-timber. This wood is valued for its rich, durability, and hardness (Burtin et al.

1998). Walnut trees are also a common source of traditional medicine. The flowers, leaves and bark of *J. regia* have a wide-range of applications related to the medicinal properties. The flowers have anti-inflammatory, antioxidant, antidepressant and antihypoxic activities. The methanol extracts from the flowers contain phenol and polyphenolic compounds that are believed to be responsible for these effects (Nabavi et al. 2011). Both the leaves and the bark are known to have antimicrobial properties (Burtin et al. 1998) which have proved useful in human respiratory tract and gastrointestinal infections, particularly *Stapholococcus aureus* (Pereira et al. 2007).

2.3 Genetic Resources

DNA is a double-stranded molecule that is composed of four nucleotides, A, C, G and T. It encodes the genetic instructions, including molecular functions, of all living organisms. The sum of all DNA in an organism is the genome. The genome is made up of both the coding and non-coding sequences of DNA. During the process of transcription, DNA is copied into RNA to begin the process of gene expression. The transcriptome is the set of all RNA molecules in a cell: rRNA, tRNA, mRNA and non-coding RNA. Unlike the genome, the transcriptome is only composed of genes being transcribed in a cell at a particular time. Over 2,177 eukaryotic species have been sequenced to date with a far greater number associated with transcriptomic resources.

As of now, 140 land plant genomes have been sequenced, assembled, annotated and published ("Information by Organism", 2015). Although many of these genomes are considered 'incomplete', they have led to advances in agriculture, particularly crop production, drought tolerance, and disease resistance. Sequencing the genomes of plants allows for the mapping of desirable traits for selective breeding (M. Bolger et al. 2014). In certain cases, including the sugar beet and the potato, the genome sequence has been used to gain information to regulate the

maturity and life cycles of the plants (M. Bolger et al. 2014). A major challenge in plant genome sequencing is the genome complexity and size, which can vary widely, ranging from approximately 135 Mb (*Arabidopsis thaliana*) to over 20Gb (loblolly pine). *A. thaliana* is a model plant species and the first plant genome sequenced (2000) (Arabidopsis Genome Initiative, 2000). As improvements in sequencing technology were realized and costs decreased, the number of complete plant genomes steadily increased. Nearly all major clades within the plant kingdom have some genomic representation today.

Until now, the genetic resources available for walnut were fairly limited. Expressed sequence tags (ESTs) are short, single-pass reads of mRNA derived from cDNA library clones (Parkinson and Blaxter, 2009). Sanger-sequenced ESTs served as the primary and most inexpensive method for gene discovery in many species. In walnut, 5,025 ESTs from leaf tissue were sequenced to identify simple sequence repeat (SSR) markers (Zhang, 2010). The EST-SSR markers have been used to create a genetic map of walnut, detect SNPs and identify microsatellites (Z. Zhang et al. 2013; Liao et al. 2014). A subsequent study generated the first genomic resources via Bacterial artificial chromosome (BAC) end sequencing of 31.2 Mbp (5.1%) of the genome (Wu et al. 2012). This analysis characterized a portion of the gene space as well as the non-coding, primarily retrotransposon components. The draft walnut genome is diploid (2n = 32) with an estimated size of 687 Mbp and an N50 of 464,955 bp.

Six DNA sequences from three plant genomes (plastid, mtDNA and nuclear) of families from the order Fagales were used to determine the phylogenetic relationship within the order. In this analysis, it was determined that *Juglans regia* falls within the Fagales subclade *Myriacaceae*(*Rhoiptelaceace*(*Juglandaceae*)) (R. Li et al. 2004). The Fagales consists of many deciduous hardwood trees. Of the Fagales, only *Betula nana* (dwarf birch of the family

Betulaceae) (Wang et al. 2013) has a published genome. No other species from the family Juglandaceae have been fully sequenced. Currently, whole genome sequencing of the close relative, *Juglans nigra*, black walnut, is underway with BAC and transcriptome resources available. Efforts are also on going to generate a genome for the English oak (*Quercus robur*) ("Quercus robur", 2015) and *Castanea mollissima* (Chinese chestnut of the Fagaceae) ("Chinese Chestnut Genome", 2015).

2.4 Applications

Goals for breeding walnut cultivars through marker-assisted methods or genomic selection include: late leafing, lateral bud fruitfulness, high protein/oil content in seeds, and resistance against blight (Keles et al. 2014; Mohan et al. 2014). A common disease that plagues walnut is Armillaria root disease, of which more than 60% of walnut trees are susceptible (Baumgartner et al. 2013). Identifying the genes related to these traits of interest or involved in causal pathways, may greatly improve overall productivity.

2.5 Sequencing Methods

2.5.1. Sanger Sequencing

Sanger machines dominated the sequencing market for over 25 years and they remain in use for small-scale projects where long reads and high quality are important. In this method, DNA polymerases copy single-stranded DNA templates by adding nucleotides to a growing chain. This method requires a single strand DNA (ssDNA) template, a DNA primer, DNA polymerase, dNTPs and chain terminating ddNTPs, which lack a 3' -OH group which is required for elongation. The sequencing is separated into four reactions with all the components, except each of the individual reactions will only contain one of the ddNTPs. The ddNTPs will be incorporated to the strand at random and this will give a DNA fragment of a certain length.

These fragments are then run on an electrophoresis gel so that the position of the ddNTP is known, and therefore the base at that position is known. The sequence can then be determined by reading the gel from smallest fragment to largest fragment (Sanger et al. 1977). After further automated improvements, including capillary electrophoresis, this method produces average read-lengths of 700 bases, can detect up to 500,000 bases per day and costs approximately $1/read (Liu et al. 2012).

2.5.2. Roche 454 Pyrosequencing

Next generation sequencing introduced parallel, high-throughput sequencing methods for substantially lower costs. Pyrosequencing is responsible for the first next-generation sequencing platform which took the form of Roche 454 and arrived on the market in 2004. The parallelized pyrosequencing method amplifies DNA inside water droplets in oil (emulsion PCR). Each droplet contains a single DNA template attached to a primer-coated bead that allows the formation of clonal fragments. An enzyme-mediated light reaction indicates base incorporation. This will generate a light signal that is recorded as a peak. If multiple nucleotides are incorporated, meaning the adjacent locations have the same base, the peak will be proportional to the number of nucleotides incorporated, due to a more intense light. If no nucleotide is incorporated, there will not be a peak and the sequence repeats itself with the next nucleotide. (Fakruddin et al. 2013). Roche 454 is no longer supported but machines are still in operation at many sequencing centers. This technology costs $850/Gb with read lengths of 700bp and takes one day to run (Fakruddin et al. 2013).

2.5.3. Ion Torrent ion semiconductor sequencing

Ion Torrent is an ion semiconductor sequencing method created to be faster and less expensive for small, targeted regions of the genome, with lower coverage. The method relies on

detecting the release of protons during DNA polymerization. This technique requires a DNA template strand and relies on the release of a proton when a nucleotide is incorporated into the growing complementary strand. The release of the proton will change the pH of the solution which is detected by the ion semiconductor. Each nucleotide is tested separately and then washed away before the next nucleotide is added (Quail et al. 2012). Ion semiconductor is fast and cost efficient, but the read lengths are short compared to other sequencing methods and it does not generate accurate reads for homopolymer repeats. The cost of Ion Torrent semiconductor sequencing is $1000/Gigabase (Gb) (Quail et al. 2012).

2.5.4. PacBio SMRT

Single molecule real time sequencing (SMRT) is a method developed by Pacific Biosciences (PacBio). This was the first long read technology that did not involve an amplification step. The real time nature of the platform therefore increased the read lengths substantially but also increased error rate and reduced the throughput. This technique uses a zero-mode waveguide (ZMW), and requires a single DNA polymerase and a single molecule of the DNA template which are affixed to the bottom of the ZMW. The ZMW is an optical waveguide which guides light energy into a small volume, in comparison to the wavelength of light. The dNTPs are fluorescently labeled with a unique color for each of the four bases. When the dNTP is incorporated to the complementary strand, the fluorescent label is cleaved, emitting a light and a base call is made from the color of this light. Sequencing takes place on a chip that contains many ZMWs. PacBio SMRT sequencing has a high error rate of approximately 13% (McCarthy, 2010). The runtime is two hours, and it produces average read-lengths of 15,000 bases, and read-lengths can be as long as 50,000 bases. These long read-lengths improve both

genome and transcriptome assembly by simplifying the alignment process. However, each run is more expensive than other sequencing technologies and the throughput is comparatively low.

2.5.6. Illumina

The Illumina platform is the most widely adopted, highest throughput and least expensive (per base) technology. Since its inception, several platforms have been made available with variations on throughput, processing time, and read length. These include the Genome Analyzer, HiSeq, MiSeq, and NextSeq. This platform operates via bridge PCR, an amplification method where primers attached to a solid surface and DNA colonies are formed from their extension. The four nucleotides are each labeled with a differently colored fluorescent dye and a blocking group. When a base is added by complementary base-pairing, the other nucleotides are washed out, and a laser is used to excite the dye of the newly added base. Illumina supports paired-end or single-end sequencing. Paired-end sequencing with a known insert size generates higher quality data for *de novo* assemblies (Morozova and Marra, 2008). Paired-end reads are obtained by sequencing both ends of a sequence and the size between the reads, the insert size, is of known length (Figure 1). The maximum read-length for Illumina HiSeq is 150bp (300bp for MiSeq). Although the run time can be up to ten days, the cost of sequencing is only $41/Gb (Quail et al. 2012).

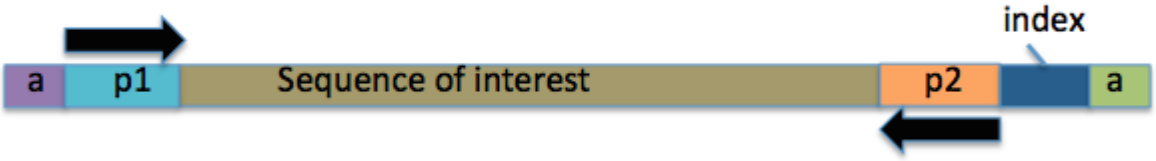

Figure 1. Paired end Illumina sequencing. Standard insert length is 300bp for paired-end reads. The arrows represent the direction sequencing will proceed. Boxes labeled 'a' are the adapter molecules which link together separate DNA molecules, and those labeled 'p' are the primers, short complementary sequences of DNA for targeted amplification of sequences. The index is a unique stretch of DNA for the purpose of multiplexing (and subsequently identifying) multiple samples per lane.

Instrument	Method	Time/run	Read-length (bp)	Reads/run	Error rate
Life Technologies Ion Torrent	Ion semiconductor sequencing	4 hours	400	4×10^6	2%
PacBio RSII SMRT	Single molecule	2 days	10,000-15,000	8×10^5	13%
Roche 454 GS FLX Titanium XL + Pyrosequencing	Sequencing by synthesis	23 hours	700	1×10^6	0.01%
Illumina HiSeq2000	Sequencing by synthesis	12 days	2 x 100	3×10^9	0.1%

Table 1 Overview of widely adopted NGS platforms.

2.6. Transcriptome Sequencing

A transcriptome is the entire set of all RNA molecules, including mRNA, rRNA, tRNA and all other non-coding RNA transcribed in one cell or a population of cells. It varies across tissues and in response to various environmental factors and therefore provides a snapshot in

time of the genes expressed in a given tissue for a specific organism. RNA sequencing (RNA-Seq) to generate a transcriptome is performed by converting the RNA to cDNA by using a reverse transcriptase (RT). The RNA is first extracted from the cell or population of cells of interest. RT is an enzyme that is associated with retroviruses and has the ability to create a complementary DNA (cDNA) strand from template RNA. The RNA can be removed from the now double strand molecule by an enzyme, RNAse. The cDNA can then be sequenced on one of the next generation sequencing platforms described above (Surget-Groba and Montoya-Burgos, 2010). Because cDNA is derived from RNA, it will not have the non-coding or regulatory elements like genomic DNA. The selection of the platform depends heavily on the depth of sequencing required which often relates to the quality of existing genomic resources. While a transcriptome is seldom considered fully complete, the first goal is often to describe the majority of the genes present in a given organism.

Future studies may use RNA-Seq studies in a comparative manner to measure relative changes in gene expression. This can take the form of time course studies that examine development or disease progression. It can also form the basis of comparisons between tissues within one individual or across individuals representing distinct populations. Gene expression can be evaluated by comparing transcriptomes (genes) directly or more accurately with the assistance of a good reference genome sequence (Surget-Groba and Montoya-Burgos 2010). In an assembly with a reference genome the reads align to the exonic regions of the genome. In a *de novo* transcriptome assembly, only exon (coding) regions will be sequenced, so these reads must be assembled by overlapping reads.

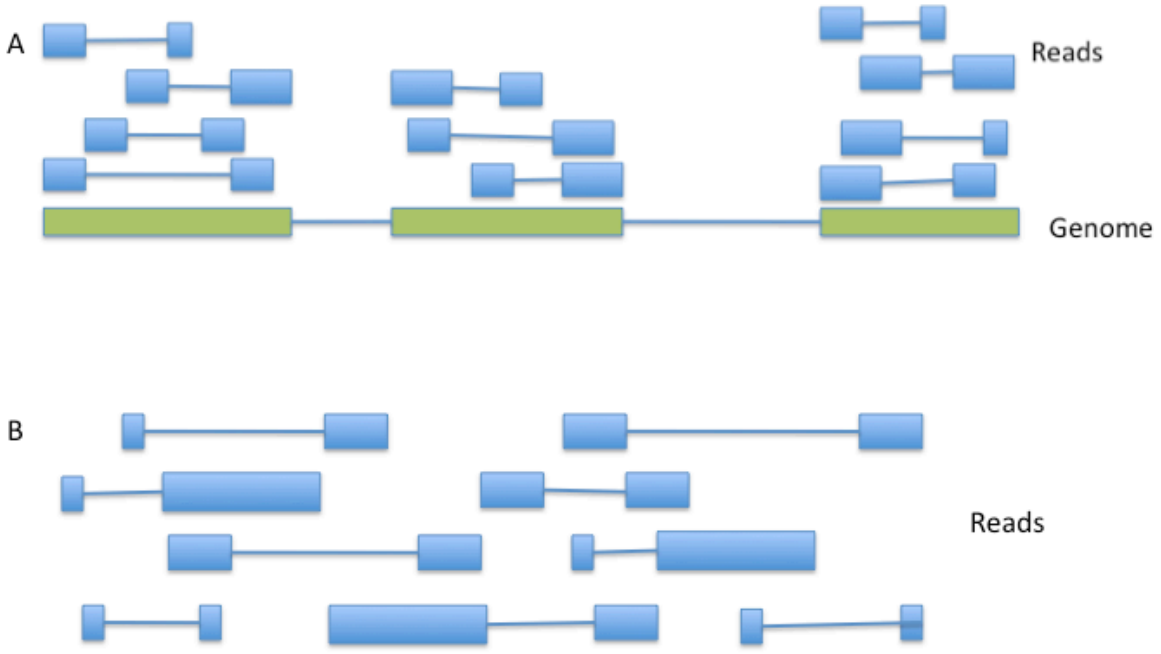

Figure 2. A. Expression analysis through read mapping. Green regions of the reference genome indicate exons. Significant differences in the raw counts between samples indicates differences in expression. B. *De novo* assembly. The reads cannot be mapped to a reference genome, so the assembly is performed through software that identifies overlapping reads and forms contigs. This technique relies on greater depth of sequencing and is greatly improved through the use of paired-end reads. The reads have an insert size of known length, so the relative position of each read can be determined.

2.7 Quality Control and Trimming

After the RNA is sequenced with the Illumina NGS platform, the graphical output depicts colored-coded peaks for each nucleotide and the position of the base call. The base calls are stored in a binary BCL file, which is then converted to FASTQ file (Andrews, 2010). The FASTQ format is a four line, text-based file that includes quality scores for each base. The sequence id and optional description are in line one, the base calls are in line two, a '+' and same optional description (line 1) are on the third line and the fourth line contains the quality scores. Each nucleotide in the sequence is assigned a quality score which corresponds to the probability of a correct base call for that position. The quality scores are encoded by ASCII characters, with

quality values ranging from 1-40. Quality values are determined by the amount of coverage, or number of bases/location on a sequence (Cock et al. 2010). The ends of sequences tend to have lower quality so the sequences must undergo quality control. The 3' ends can have an increased error rate between five- and ten-fold due to degradation (Minoche et al. 2011). The Illumina GAIIx platform has increased coverage in GC-rich regions which increases the potential error rate in AT regions (Minoche et al. 2011). After trimming degraded regions, very short reads are discarded because they do not provide enough information to align or assembly uniquely. Software options for quality control are detailed below.

```
@SOLEXA1_0001:3:1:1:487#0/1
NAAACACCCACATGGGACTCCAACAATAGCAGACAAAAACCAACCCAACCACGTACGAAACAATCGCTAAAGG
AAACTTTCACTA
+
Ba`ab_bbbba\aaa_]abb_\_`_]abbba_\aa`a^_a`a`a`_[^abbaa\aaa\_aba`a_\_a``_^_
_aaaaaa`aa`_
```

Figure 3. FastQ Format. Line 1 begins with '@' and contains sequence id and description. Line 2 is the nucleotide sequence. Line 3 '+'. Line 4 is the quality scores. Symbols explained in Table 2.

Symbol	Quality Value	Symbol	Quality Value	Symbol	Quality Value	Symbol	Quality Value
A	1	K	11	U	21	_	31
B	2	L	12	V	22	`	32
C	3	M	13	W	23	a	33
D	4	N	14	X	24	b	34
E	5	O	15	Y	25	c	35
F	6	P	16	Z	26	d	36
G	7	Q	17	[	27	e	37
H	8	R	18	\	28	f	38
I	9	S	19	]	29	g	39
J	10	T	20	^	30	h	40

Table 2. PHRED quality scores from 0 to 93 are represented using ASCII 33 to 126 in Illumina FASTQ files

Quality control and trimming of the resulting reads in FASTQ format is implemented through independent computational approaches that evaluate the overall success of the library and remove problematic reads/bases. Applications that implement simple trimming methods work with specific user-provided thresholds for quality and read-length. Since longer, high quality reads provide more reliable information, the user can also set a minimum read length, which will remove reads shorter than the specified minimum (A. Bolger et al. 2014). The minimum read length is typically no less than 35bp to ensure the reads can be uniquely assembled or mapped. The sliding window approach examines a small portion of the read at a time by sliding across the sequence. This approach allows portions of the sequence with quality above the threshold to be saved, rather than discarding the entire read. In general, the quality threshold should be set to a minimum of 30 which translates to a chance of 1 in 1000 that the

base is called incorrectly. Both the 5' and 3' ends are trimmed to achieve maximum quality. One implementation of this method is the program, Sickle, which uses a window size of 10% of the sequence length (Joshi and Fass, 2011).

Figure 4. A. Illumina read before and after trimming by Sickle. 25 bases at the end of the sequence were trimmed due to poor quality. B. A sequence that will be discarded entirely since the quality scores are less than 20 across the read.

2.8 *De novo* Transcriptome Assembly

Since many organisms do not have a reference genome, the transcriptome must be assembled *de novo*. The era of next-generation sequencing has produced a several fold increase in the number of reads available for assembly compared to traditional Sanger methods. Early assemblers such as Arachne, Celera, and Phrap implement what is known as overlap-layout-consensus algorithms (Gordon et al. 1998; Myers et al. 2000; Batzoglou et al. 2002). These approaches are highly accurate, require overlap between reads, and can be compared to trying every piece of a puzzle in the missing space until you find the one that fits. They provide an all versus all comparison of the available reads in order to extend the contig (Zhao et al. 2011). The nature of this approach is very computationally intensive and does not scale well as the number

of reads grows. In general, this method is ideal for few reads with high degree of overlap (Zhao et al. 2011).

To reduce the computational resources and time needed for assembling large sets of short read data, various implementations of de Bruijn graphs are used in both genome and transcriptome assembly software packages (Compeau et al. 2011). De Bruijn graphs align *k*-mers, sequences of length *k* that are shorter than the read length, by a *k*-1 (edge) overlap to create contigs. In this method, all the *k*-mers in a sequence are determined and a graph is created by the connecting pairs of *k*-mers (nodes) with a *k*-1 overlap. In a given sequence, it is common for a *k*-mer to appear multiple times. A de Bruijn graph is composed of unique *k*-mers, so when this multiplicity happens, the amount of nodes in the graph will decrease (the identical nodes merge into one). This creates a multiedge, in which the node branches off to all possible edges for that specific *k*-mer. A node's indegree refers to the number of edges leading into that node, while the outdegree is the amount of edges leading out of the node (Langmead, 2014). Some nodes will have multiple edges which complicates the assembly, by creating multiple assembly possibilities, but this can be improved by implementing multiple *k*-mers (Compeau et al. 2011). Cost-effective from a computational perspective in that they are much faster than the overlap-layout method, de Bruijn graphs do not handle sequencing errors or highly polymorphic reads well and do not preserve positional information.

There are several implementations of de Bruijn graph methods in transcriptome assembly for short read Illumina data. The use of multiple *k*-mers in programs such as Oases (Schulz et al. 2012), SOAP-denovo-Trans (Luo et al. 2012) and Trans-ABySS (Robertson et al. 2010) improves assemblies by decreasing the complexity of nodes with multiple edges. Trinity-RNAseq uses a single *k*-mer of 25 and then creates a *k*-mer library of all overlaps with length 25

(Grabherr et al. 2013). A single *k*-mer decreases the memory requirements. Although the runtime for Trinity is longer than other assemblers, this three module program often outperforms the others in short-read assembly (Zhao et al. 2011). Trinity consistently produced the most full-length transcripts, the least fused-transcripts and performed best for both small and large data sets (Zhao et al. 2011). It should be noted that longer read data originating from technologies such as 454 may rely on assemblers such as Newbler (Miller et al. 2008) or MIRA (Loman et al. 2012) which implement a hybrid approach of de Bruijn and overlap graphs to improve accuracy with less read depth.

2.9 Open Reading Frame Prediction

To understand the function of a nucleotide sequence, the open reading frame (ORF) must be determined. An ORF is the portion of DNA that has the ability to code for a protein. A reading frame is the sequence of DNA or RNA that is composed of non-overlapping triplets, codons. In eukaryotes, an ORF must begin with a start codon (typically AUG in RNA), and end with a stop codon (UAA, UAG, UGA). The 5' untranslated region (UTR) is sometimes translated into a protein. This portion can then regulate translation, transcription or protein exporting, among other functions. The 5' UTR can also complicate the identification of the start site. Without a reference sequence, the ORF is typically predicted through sequence evidence from other organisms. A number of tools are available to determine the ORF in a single sequence or set of sequences.

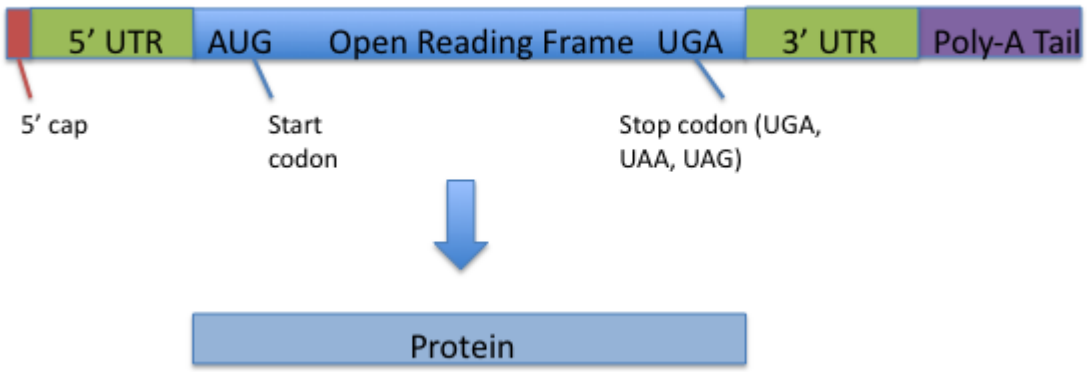

Figure 5. Open Reading Frame. The mRNA (top) has a 5' cap, 5' and 3' UTRs, a Poly-A Tail and ORF. For most eukaryotic sequence, the ORF begins with the start codon (AUG) and ends with a stop codon (UGA, UAA, UAG). ORF codes for protein which is produced following translation.

The most basic approach to ORF identification is to simply identify the longest frame by analyzing the six possible reading frames (3 forward beginning with the nucleotide at position x, x+1 and x+2; 3 with the reverse following the same rules). One such tool, OrfPredictor implements this with a user-defined minimum frame length. It also searches for ORFs by sequence homology with BLAST (Min et al. 2005). An improved implementation involves a combination of reading frame translation, sequence similarity, protein domains, and machine learning from experimentally validated proteins. TransDecoder, a program integrated with Trinity, implements these considerations in a multi-phase approach. This method also has the ability to identify multiple ORFs that represent alternatively spliced candidates.

2.10 Annotation

Annotation is the process of identifying key features of the genome or transcriptome, particularly genes and their protein products (Stein, 2001). In addition to the structural annotation of the ORF, the functional annotation provides insight to the gene products role and importance to the organism. This may also include the pathways this gene is involved in.

Typically a multi-step process, the annotation of *de novo* assembled transcripts generally begins with BLASTing the assembled sequences against databases of known proteins. Non-model organisms may have elements unique to themselves, which complicates the annotation. Unique sequences often cannot be annotated from sequence similarity approaches alone (Carpentier et al. 2008).

2.10.1. BLAST

Sequence homology, the hypothesis that similar sequences have the same or similar functions from a shared ancestor, is commonly used in annotation. Basic Local Alignment Search Tool (BLAST) is an algorithm used to compare unknown sequences against a database of curated sequences. The query sequence is aligned to the database and if the alignment is above a certain threshold, the alignment is considered significant and possibly informative. BLAST can be used with amino acid sequences for proteins or nucleotides for DNA sequences. The algorithm works by finding similar sequences through short matches (local alignment), rather than whole sequence alignment (global alignment), through a process called seeding. Seeding results in a list of all the three-letter 'words' possible in a sequence, and then finds sequences in the database by searching for these 'words'. The database sequences that have the most common 'words' compared to the query sequence are then aligned with the query sequence. BLAST alignments are measured for significance through an E-value, or Expected value. This measure is the number of hits that are expected to be found between the query and the database sequences of the same length at random, relative to database size. A lower E-value (less than 1) indicates a more significant match (Altschul et al. 1990). Improvements have been made to the BLAST algorithms by other applications in terms of both accuracy and speed. These programs are refined for particular searches, such as high- or low-identity. An algorithm refined for low-

identity searches, UBLAST is appropriate for searches between sequences that diverged from an ancient common ancestor (Edgar, 2010).

Query label	Target label	Percent identity	Alignment length	Number of mismatches	Number of gap opens	Start position in query	End position in query	Start position in target	End position in target	E-value (calculated)	Bit score

Table 3. Example output for UBLAST. Each alignment is summarized in a tab-delimited text format with information on the scores (bit score and E-values), start and end positions of each sequence, and the alignment length.

2.10.2. Gene Ontology

The Gene Ontology project began in 1998 and enables biologists to assign consistent nomenclature to curated gene products. The GO nomenclature is organized in a hierarchy implemented as directed acyclic graphs with the top nodes representing each of the main three categories (cellular component, molecular function and biological process). Subsequent terms in the hierarchy are organized by defined relationships (*is-a*, *part-of*, *regulates*, etc). Each GO term has been curated by one or more biologists and these are generally well established for model organisms. Each term has a unique identifier, name, definition, and how this term was originally assigned (computational or experimental methods) (Conesa and Götz, 2008). Terms can be defined in non-model organisms by sequence homology to model organisms. Several packages exist to assign terms to sequences, including GO::TermFinder (Boyle et al. 2004), TermGenie (Dietze et al. 2014) and Blast2GO (Conesa and Götz, 2008). Blast2GO is a robust solution that considers sequence similarity, source database, GO hierarchy, and the quality of the annotation through a weighted algorithm (Conesa and Götz, 2008).

2.10.3. Protein Domain Identification

A protein domain is the part of the protein that is independent of the rest of the protein, can evolve and has a function (Veretnik et al. 2009). In a full-length protein sequence, there can be one or many domains. Their functions, order, and spacing are of primary interest. By searching against multiple models, a protein can be defined by its patterns, profiles, and fingerprints. Protein patterns refer to motifs in the secondary structure, such as hairpin loops, helices and zinc fingers. Patterns alone are not enough to determine the function. Profiles and

fingerprints include the patterns but also their order and relative distance. Multiple programs are available to predict protein families and functional domains within the protein. One such program is InterProScan which cross-references the protein family and domain results with Gene Ontology before reporting a definition for a more complete and high-quality report (Jones et al. 2014). InterProScan can run against a variety of protein signature databases including, but not limited to: Panther (Mi et al. 2012), Pfam (Finn et al. 2014) and SMART (Letuncik et al. 2015).

Figure 6. Protein motifs are determined from sequence alignments, producing a profile for each motif. The fingerprint profile represents the correct order and spacing of the profiles which can be subsequently identified and searched against in databases.

2.10.4. Gene family identification

Orthologous groups are those genes shared by different species based on a common ancestor. Orthologous gene family analysis identifies these genes to assist in studies related to evolution and comparative genomics. The Markov Cluster (MCL) algorithm is one method used for clustering proteins into families based on their sequence similarity (L. Li et al. 2003). TRIBE-MCL is one implementation that is able to cluster multi-domain proteins, fragmented

proteins and promiscuous proteins (proteins that perform different functions). While traditional methods rely solely on sequence similarity, TRIBE-MCL uses sequence similarity in addition to flow simulation, the stochastic fluctuation of graphs, by MCL. Orthologous group analysis provides insight to conserved gene families across closely related species as well as unique gene families (Enright et al. 2002).

2.11 Genome Alignment

Once a complete *de novo* transcriptome has been acquired, if a reference genome or genome of a closely related species is available, the transcriptome can be mapped back to it. A variety of splice aware aligners are available for this task: TopHat2 (Kim et al. 2013), STAR (Dobins et al. 2012) and GMAP (Wu and Watanabe, 2005). GMAP is able to align the sequences despite polymorphisms, and sequencing errors, and without probability models for alternative splicing. The transcriptome should align to the reference genome with high coverage and identity, but when using related species, both the coverage and identity should be decreased (Wu and Watanabe, 2005). This step is a useful verification of the quality of the transcriptome assembly. The transcriptome may also be used to improve the genome assembly by scaffolding coding regions. Mapping programs can facilitate this process as well.

3. Methods

3.1. Sequencing and Quality Control

The 22 samples representing 19 unique tissues were collected from a single population, Chandler, and frozen in liquid nitrogen and then transferred to a -80°C freezer. Poly-A selected RNA was isolated from each sample using the hot borate method (Wilkins and Smart, 1996) followed by purification and DNAse treatment using an RNA/DNA Mini Kit (Qiagen, Valencia, CA) per the manufacturer's protocol. High quality RNA was confirmed by running an aliquot of

each sample on an Experion Automated Electrophoresis System (Bio-Rad Laboratories, Hercules, CA). The cDNA libraries were constructed following the Illumina mRNA-sequencing sample preparation protocol, TruSeq (Illumina Inc., San Diego, CA). Final elution was performed with 16 µL RNase-free water. The quality of each library was determined using a BioRad Experion (BioRad, Hercules, CA). Each library was run as an independent lane on a Genome Analyzer II (Illumina, San Diego, CA) to generate paired-end sequences of 85 bp in length. Quality control of the resulting reads was performed via Sickle (version 1.210) with a minimum Phred-scaled quality score of 35 and a minimum length of 45bp. Following quality control, the 19 tissues were assembled individually as well as a combination of all reads for a complete transcriptome.

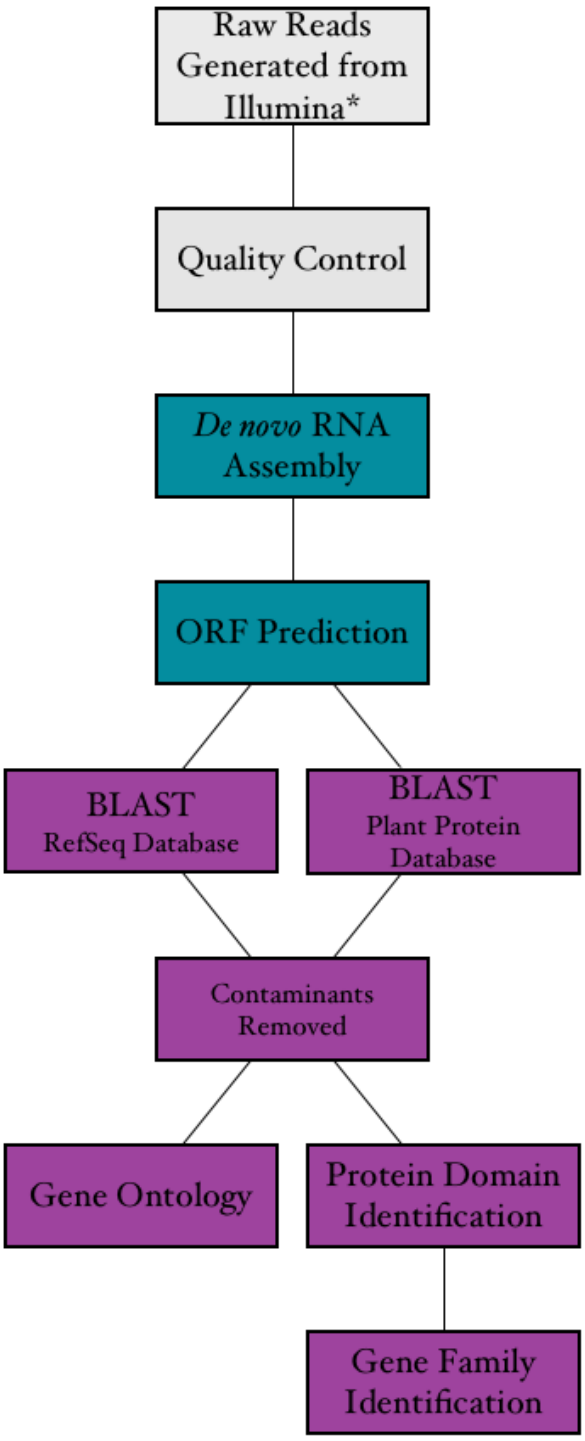

Figure 7. Analysis pipeline for the transcriptome assembly and annotation of *Juglans regia*. *Reads were generated from UC Davis genome sequencing center. Gray represents the sequencing and quality control portion of the pipeline; Teal represents assembly steps; Purple represents annotation.

3.2. Assembly

Trinity RNA-Seq (r20140413p1) *de novo* assembler was run on the trimmed sequence set with standard parameters and a minimum contig length set to 300bp. Before determining the ORF, the Trinity 'genes' were extracted using a custom in-house Python script. When calculating the N50, the contigs were ordered by increasing length, therefore giving greater weight to longer contigs. Specifically, the N50 statistics tells us that the 50% of the sequences in our set have a length of this value or greater. The Trinity genes were processed for ORF detection by TransDecoder (version 2.01), a program that is integrated with Trinity. TransDecoder was run with the 'train' option. This option employs machine learning approaches so the algorithm can learn from full-length protein sequences from the same or closely related species. The top five related species, as determined through concurrent functional annotation efforts using the RefSeq database, were *Vitis vinifera* (5399 shared genes), *Ricinus communis* (3379), *Cucumis sativus* (2650), *Fragaria vesca* (2337) and *Glycine max* (1179) were provided as a training set. The "search-pfam" option allowed for scans against the PFAM domain database (PFAM-A, version 27). The output included all predicted ORFs and the subset of full-length sequences, which have unambiguous protein-coding regions. Since replicates were not available to accurately evaluate expression differences between the individual *de novo* assemblies, the sequences were clustered with 0.5 identity using the cluster_fast option from USEARCH to provide a general view of unique contributions.

3.3. Annotation

The transcripts and their translated proteins derived from individual assemblies and the combined assembly underwent functional annotation. First, the complete or longest (if complete was unavailable) sequences from the TransDecoder output were extracted using a custom in-

house Python script. To find local alignments below an E-value threshold, these sequences were run through the USEARCH (version 7.0.1090_i86linux64) UBLAST algorithm which implements a modified version of NCBI's BLASTx. The databases queried were NCBI *RefSeq*, a curated collection of non-redundant protein sequences from all organisms and NCBI *RefSeq Plant Protein Full*, a custom curated database of full-length plant proteins. The E-value threshold was set to 1e-9, while the weak E-value was 0.001. Custom scripts were used to filter contaminants based on three databases (fungal, insect, and bacteria). Analysis of the translated sequences was run with InterProScan (v5.0) against the Pfam and Panther protein domain databases. An in-house python script was employed to further refine the full annotation results. This script considers the UBLAST results from multiple searches (in this case RefSeq and Plant Protein), XML results from InterProScan, and the GO terms in all three categories assigned by Blast2GO (and generates the search input for this program). It selects the best (based on score) and most informative (based on description details) functional annotation from these results. Additionally, the script reports summary statistics of the transcriptome assembly, including the N50, species information on the BLAST hits, and provides a final comprehensive annotation report.

3.4. Genome Alignment

The transcriptome was mapped to the draft *Juglans regia* genome ("Genome *Reju*", 2015) (which was assembled concurrently) using the splice-aware aligner GMAP (v2014-12-28). The current draft is 687 Mb, with an N50 length of 46,148bp, and a total of 221,640 contigs. The transcripts were mapped with 98% coverage and identity. Less stringent parameters of 95% coverage and identity were also used in this analysis.

3.5. Orthologous Gene Families

The orthologous groups analysis was implemented with TRIBE-MCL to cluster 352,562 protein sequences from 9 angiosperm species (two monocots and seven dicots) and one bryophyte: *Arabidopsis thaliana* (27,416 proteins), *Glycine max* (54,257), *Oryza sativa* (40,738), *Physcomitrella patens* (32,400), *Populus trichocarpa* (41,434), *Ricinus communis* (31,221), *Theobroma cacao* (29,484), *Vitis vinifera* (26,504), *Zea mays* (39,323) and *Juglans regia* (29,785). These proteins were chosen from the PLAZA 3.0 set and only full-length sequences and those greater than 21 amino acids were included in the analysis. PLAZA 3.0 is a tool for comparative plant genomics that contains curated protein sets for all sequenced plant genomes. Proteins whose genomic coordinates did not match the reported protein coding sequence or translate into the reported protein were eliminated. The process begins with pairwise NCBI *blastp* v2.2.27+ (*E*-value cutoff of 1*e*-05) against the full set of protein sequences. The negative $\log_{10}$ of the *blastp E*-values produces a network graph which is the input necessary to define orthologous groups. A moderate inflation value of 4.0 was used. The inflation value is user-selected and used to simulate random walks in the network graph. Next, Pfam domains were assigned from the PLAZA annotations to the individual sequences. InterProScan 5.0 was applied to the *de novo* assembled walnut transcripts as described earlier. Pfam domains and related GO term assignments with *E*-values < 1*e*-05 were retained and the GO terms were normalized to level four of the classification tree. Families in which all predicted elements were classified as retroelements were removed.

4. Results

4.1. Sequencing and Quality Control

Sequencing by Illumina Genome Analyzer II produced a combined 1,062,838,572 reads. Following quality control by Sickle, read counts (total) were reduced to 978,128,921. The read counts after QC for each library ranged between 24,361,590 (leaf-young) and 61,797,713 (packing tissue). The 22 trimmed libraries were then concatenated into one combined library for a single assembly of a full transcriptome (Table 4).

Tissue Source	Developmental Stage	Reads (Before QC)	Reads (After QC)
Callus Exterior	N/A	30,577,642	28,100,032
Callus Interior	N/A	60,902,168	58,222,286
Catkins	Immature	39,254,844	37,113,012
Embryo	Mature	37,247,600	34,387,782
Pistillate Flower	Vegetative	43,330,174	40,571,295
Pistillate Flower	Vegetative	33,983,772	29,328,010
Hull Cortex	Mature	64,446,528	61,424,051
Hull Immature	Immature	62,364,320	59,066,242
Hull Immature	Immature	57,673,738	54,040,229
Hull Peel	Mature	44,029,546	42,804,402
Hull-dehiscing	Senescent	61,885,858	54,850,389
Fruit Immature	Immature	58,683,826	55,371,217
Leaves	Vegetative	61,822,738	58,571,355
Leaf-Mature	Vegetative	53,333,298	42,408,263
Leaf-Young	Vegetative	43,732,392	40,809,417
Leaf-Young	Vegetative	32,326,356	24,361,590
Packing Tissue	Mature	59,283,694	49,505,694
Packing Tissue	Immature	64,903,726	61,797,713
Pellicle	Mature	43,392,812	41,426,900
Root	Vegetative	39,839,020	37,852,140
Somatic Embryo	Immature	28,357,550	27,180,861
Vegetative Bud	Vegetative	41,466,970	38,936,041

Table 4. Read counts for each tissue before and after quality control by Sickle. Highlighted libraries are duplicates and were assembled together.

4.2 Assembly

The full assembly produced 114,944 transcripts, including 78,645 unique genes. The unique genes are those transcripts that clustered separately by shared sequence content. These unique genes are either full-length or shorter isoforms; a custom Python script was used to extract the full-length, or if unavailable, the longest isoform. The mean length of the transcripts was 1180 bp while the N50 was 1833. In this assembly, a total of 66,020,861 bases were combined into 1833 transcripts. TransDecoder (v1.0) identified 29,785 total sequences, of which 16,594 were full-length (46%). After ORF selection, the average sequence length was 1,047 bp and the N50 was 1226. The average N50 for the individual library assemblies of full and partial ORFs was 1405 (Table 5). The total number of full and partial ORFs ranged from 18,618 (embryo) to 25,071 (Hull immature). The number of full-length genes ranged from 5,294 (embryo) to 12,734 (Leaf) (Table 6). The sequences of the individual libraries were clustered to roughly determine the unique genes within each tissue. The immature hull tissue showed the greatest number with 2.78% of its transcripts not forming clusters with others (Table 7). The young leaf showed the least diversity with only 0.61% of sequences not clustering.

Total number of reads	1,062,838,572
Total number of quality control reads	978,128,921
Total number of assembled bases	132,041,772
Total number of transcripts	111,944
Mean length of transcripts	1180
Minimum contig length	297
Maximum contig length	15,345
N50 of transcripts	1833
Trinity Genes	78,645
Partial & full-length sequences (ORFs)	29,785
Full-length sequences	16,594

Table 5. Statistics for the full *de novo* transcriptome assemblies

Library	Total Number of Sequences (Total ORFs)	N50	Total Number of Sequences (Selected Frame)	N50	Total Number of Sequences (Full Length)	N50
Callus Exterior	22328	1218	20437	1242	7402	1404
Callus Interior	23690	1296	21177	1323	9463	1428
Catkins	23212	1257	21258	1272	8507	1431
Embryo	18618	1137	17261	1152	5294	1314
Hull Cortex	22271	1179	20268	1200	7255	1353
Hull dehiscing	20629	1257	18433	1290	7282	1416
Hull Immature	25071	1293	22566	1314	10164	1467
Hull Peel	22187	1272	19839	1302	8587	1434
Fruit Immature	24247	1299	21537	1335	9615	1473
Leaf	26837	1347	22775	1389	12734	1494
Packing Tissue	22601	1206	20536	1230	7073	1365
Packing Tissue	24376	1338	21261	1377	10482	1497
Pellicle	21863	1278	19697	1305	8575	1473
Pistillate Flower	24439	1353	21394	1389	10954	1479
Root	25023	978	23717	990	5357	1197
Somatic Embryo	22736	1293	20465	1317	9011	1434
Vegatative Bud	23715	1269	21540	1293	8885	1416

Table 6. Partial and Full-length sequences with the associated N50 for individual tissue libraries.

Sequenced Tissue Libraries	Unique Sequences from Assemblies	Number of Sequences Clustered	Unique Sequences from Assemblies (%)
Callus Exterior	166	20437	0.81
Callus Interior	210	21177	0.99
Catkins	285	21258	1.34
Embryo	214	17261	1.24
Hull Cortex	191	20268	0.94
Hull-dehiscing	207	18433	1.12
Hull Immature	590	21221	2.78
Hull Peel	208	19839	1.05
Fruit Immature	223	21537	1.04
Leaf	488	21394	2.28
Leaf Immature	159	21788	0.73
Leaf Mature	177	18283	0.97
Leaf young	140	22775	0.61
Packing Tissue Immature	362	21261	1.70
Packing Tissue Mature	207	20536	1.01
Pellicle	249	19697	1.26
Pistillate Flower	195	22566	0.86
Root	556	23717	2.34
Somatic Embryo	357	20465	1.74
Vegetative Bud	180	21540	0.84

Table 7. Tissue comparison between libraries. All sequences from each tissue were clustered. The counts above are the sequences that did not cluster (are significantly different from rest).

4.3. Annotation

Of the 29,785 sequences queried from the full assembly, 25,357 were annotated (85%). Of the 16,594 that were determined to be full-length, 14,528 (86%) were annotated. Of the annotated sequences, 61% were informative (known function) and no contaminants (bacteria, fungal, or insect) were identified. The individual library informative annotations ranged from 82% (Hull Immature) to 90% (Embryo) (Figure 8). All sequences with an informative BLAST result had at least one Gene Ontology term assigned. 73% of sequences had at least one

Biological Process term, 63% of sequences had at least one Molecular Function term, and 70% had at least one Cellular Component term. The top five molecular functions are ATP binding (1399 transcripts with function assigned), DNA binding (731), zinc ion binding (675), protein serine/threonine kinase activity (471) and metal ion binding (444) (Figure 9). The top five biological processes are oxidation reduction process (994), metabolic process (893), protein phosphorylation (618), regulation of transcription, DNA templated (605) and transcription, DNA-templated (331) (Figure 10). In total, 26.2% of sequences did not have protein domain information from either the Pfam or Panther databases used with InterProScan. Based on annotation, the top five species with which *J. regia* shares the highest number of genes with are: *Vitis vinifera* (5760), *Populus trichocarpa* (3693), *Ricinus communis* (2902), *Fragaria vesca* (2418) and *Citrus sinensis* (2173) (Figure 11).

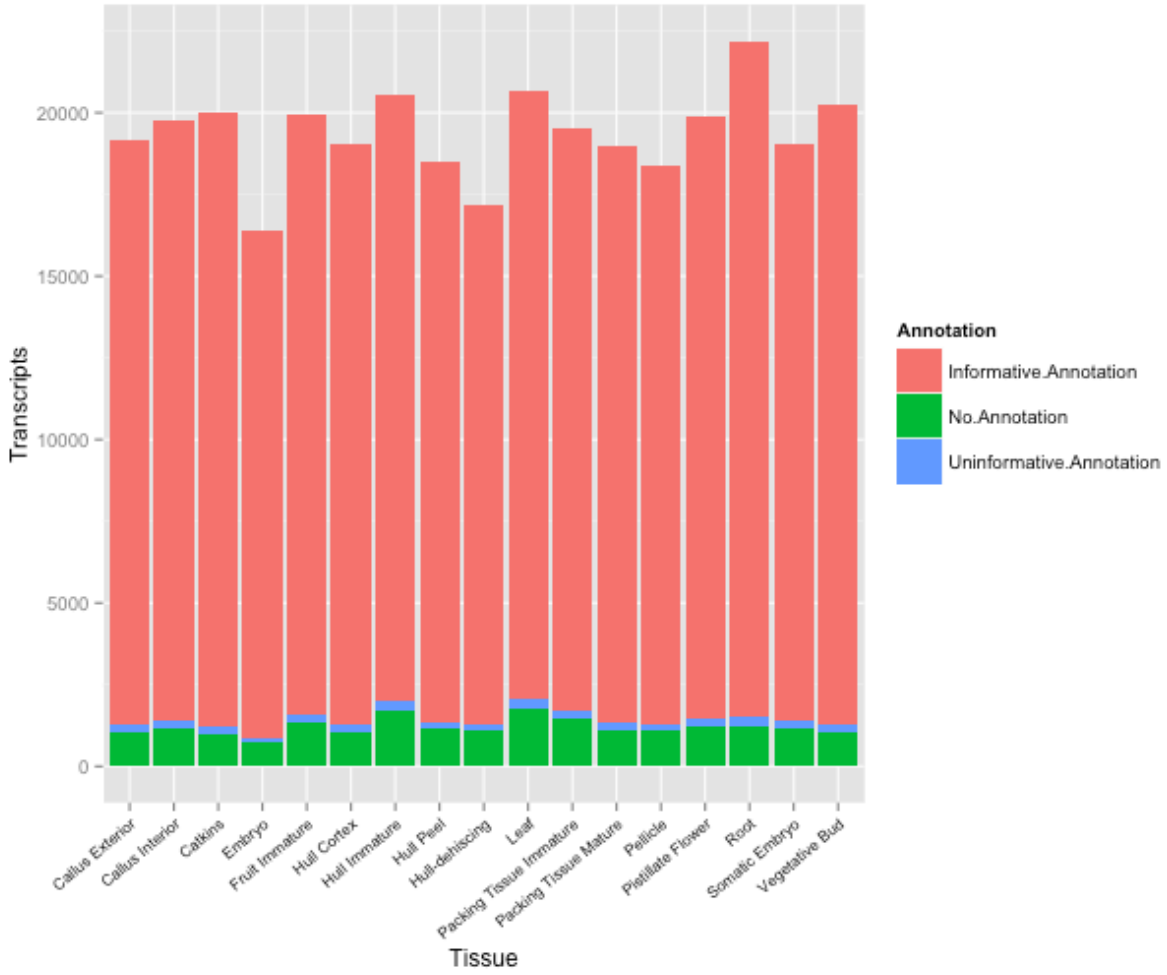

Figure 8. Distribution of annotation (informative, uninformative, no annotation) on all transcripts from the individual *de novo* transcriptome assemblies by tissue.

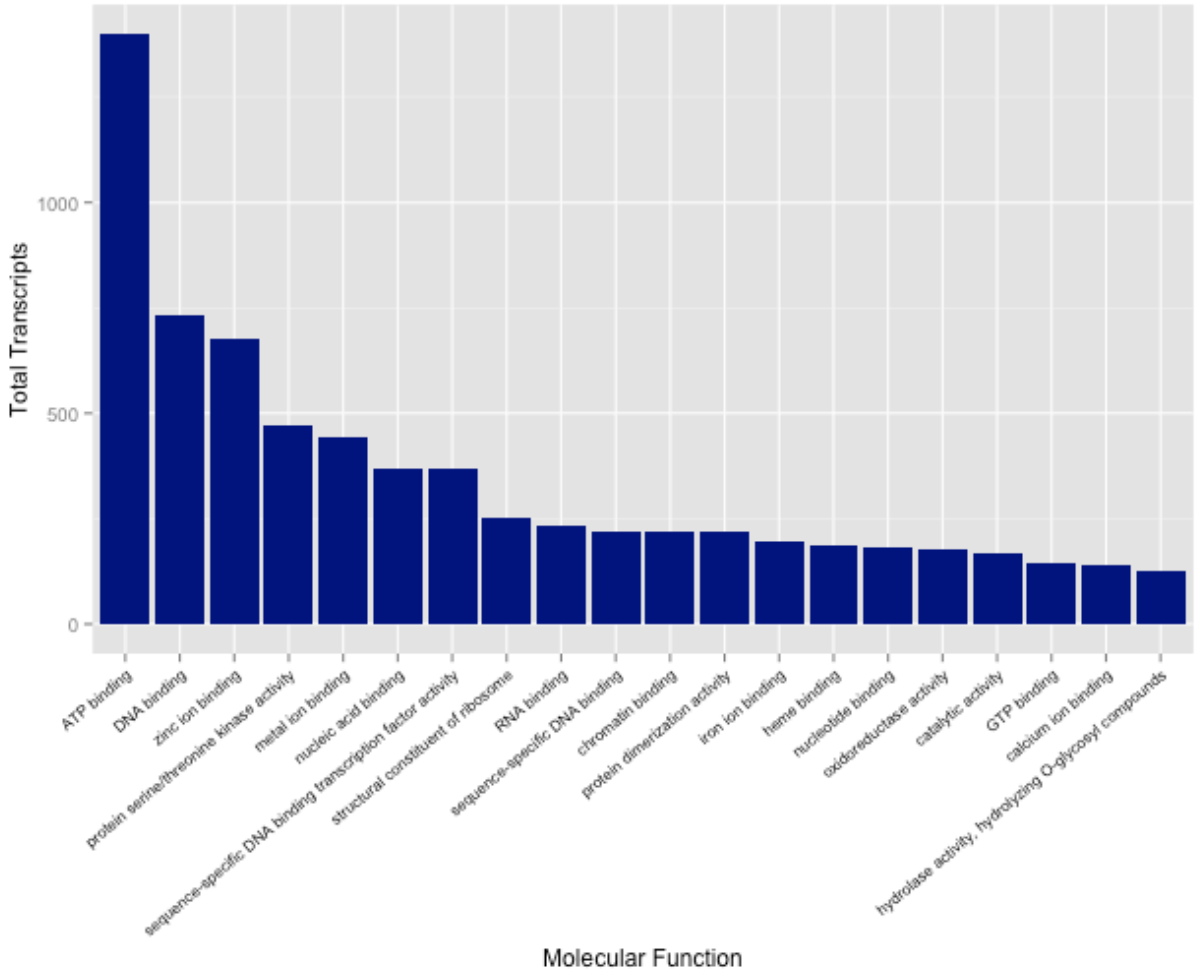

Figure 9. Distribution of the top 20 Molecular Function Gene Ontology terms for the full *de novo* assembly

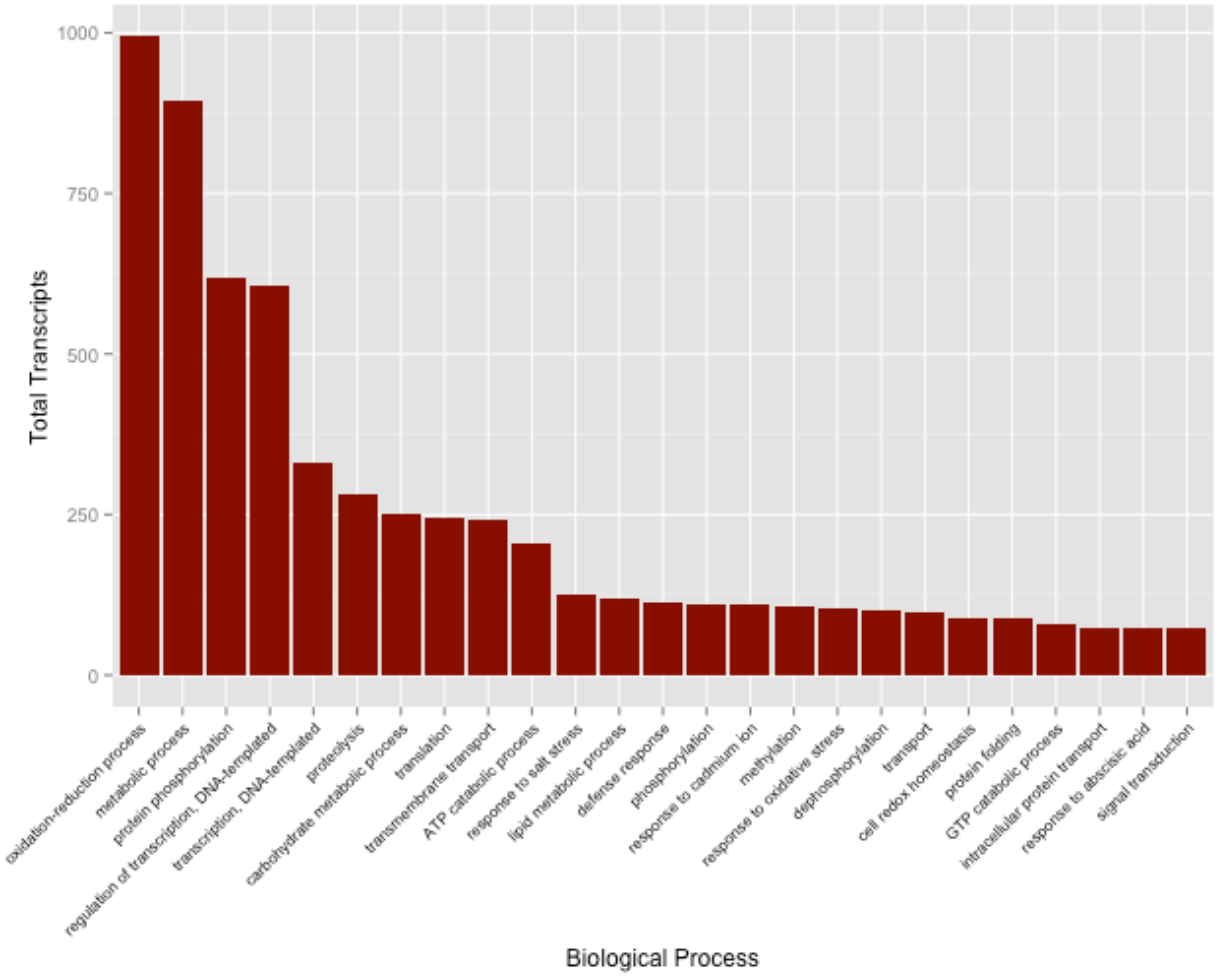

Figure 10. Distribution of the top 20 Biological Process Gene Ontology terms for full the *de novo* assembly.

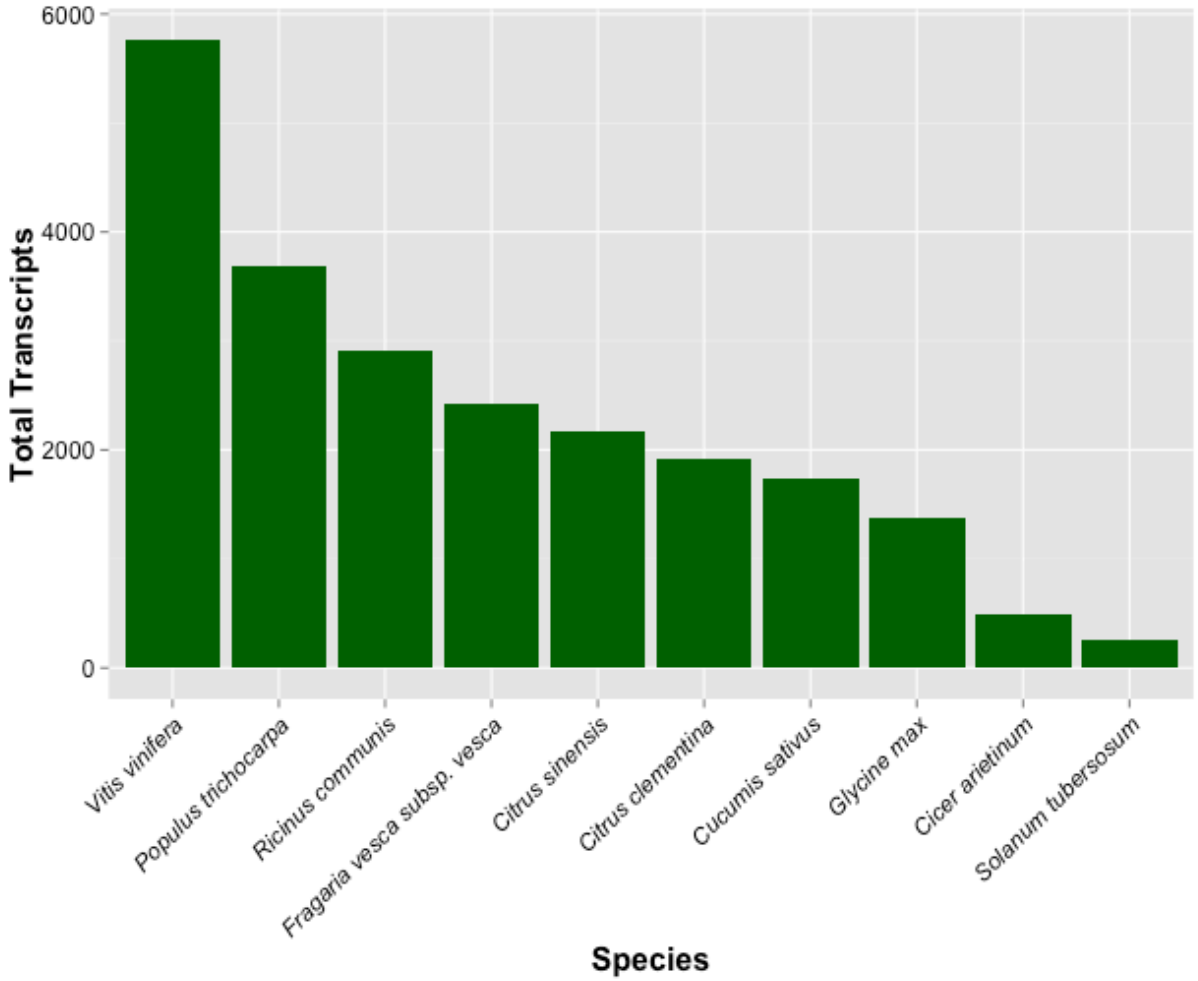

Figure 11. Top ten closely related species by sequence similarity. Total transcripts are the number of unique genes shared with *J. regia* for the full *de novo* assembly.

4.4. Genome Alignment

All ORFs determined by TransDecoder were mapped back to the walnut draft genome which was sequenced separately and concurrently. The draft genome has a total of 221,640 contigs which make up 687 Mbp. Using the parameters 98% coverage and 98% identity, 30,622 (85.45%) transcripts mapped back to the genome. With the less stringent parameters, 95% coverage and identity, 32,036 transcripts (89%) mapped back to the genome (Table 8).

Coverage (%)	Identity (%)	Transcripts Mapped (%)
98	98	85.45
95	95	89

Table 8. Genome alignment summary. Percentage of transcripts that aligned to the draft genome of *J. regia* with two different coverage and identity scores.

4.5 Orthologous Gene Families

The Tribe-MCL gene family analysis of 353,562 protein sequences formed a total of 10,092 gene families with a size of at least 2 sequences or more. Of these, 4,897 genes families were fully conserved among the ten plant genomes with at least one gene per species represented. There were 31 gene families unique to walnut, 19 of which had no protein domain annotation. These gene families do not correspond to a Pfam domain, so the function is difficult to determine. Four families with a high copy number and domains associated with repeats were labeled as transposable elements and eliminated from further analysis. The remaining 31 families ranged in size from 5 to 22 genes. Of the 10,092 gene families, 362 had members from the other nine plant species but had no representation from walnut.

Pfam Family	Pfam ID	Members in Walnut	GO ID
DUF4283 zf-CCHC_4	PF14111 PF14392	22	N/A
PapC_C PapC_N	PF13954 PF13953	15	GO:0005215 GO:0005515
Myb_DNA-bind_3	PF12776	12	N/A
NAM_associated	PF14303	6	N/A
LRR_4 protein binding	PF12799	5	GO:0005515
DUF4283	PF14111	5	N/A
Mannitol-dh oxioreducatase activity	PF01232	5	GO:0005515 GO:004353
BPD_transp_2 transporter activity	PF02653	5	N/A

Table 9. Annotation summary for the gene families unique to walnut as defined by TRIBE-MCL analysis.

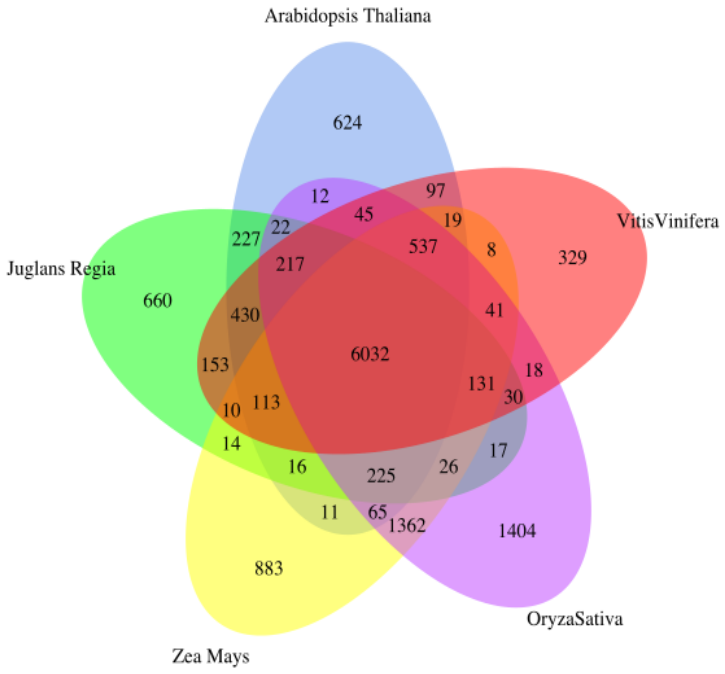

Figure 12. Venn diagram of shared genes for *J. regia* and four other plant species. The full analysis included a total of nine other plant species which reduced the number of gene families unique to walnut

5. Discussion

Trinity RNA-seq *de novo* assembler performed the assembly of Illumina short reads. While other software packages for transcriptome assembly are available, Trinity has been shown to provide the highest quality assembly. In a study by Grabherr et al. (2011), Trinity recovered more full-length transcripts than the other software packages, transcripts over a broad range of transcript expression levels, and alternatively spliced isoforms with a high level of accuracy (Grabherr et al. 2011). Zhao et al. (2011), completed a comparison study on assemblers on the transcriptomes of *Drosophila melanogaster, Camilla sinensis* and *Schizosaccharomyces pombe* in which Trinity out performed the other packages in all categories except speed (Zhao et al. 2011). In a similar study, Zhang et al. (2013) compared five assemblers and two NGS technologies on two Geraniaceae species and found that the Trinity assembly from Illumina sequencing data provided the highest quality assembly (Zhang et al. 2013). In our individual as well as full assembly, we were able to assemble a reasonable number of partial and full-length unique genes at 29,785 which is probably still lower than the total number of genes expressed in common walnut due to incomplete sampling, as well as transcripts not recovered or misassembled. Average gene numbers in other angiosperm species range from 27,029 in *Arabidopsis* (Arabidopsis Genome Initiative, 2000) to 45,555 in *Populus trichocarpa* (Tuskan et al. 2006).

Open reading frame prediction is a difficult task for organisms without a reference genome or close relative. This prediction is further complicated by errors in sequencing and incomplete transcripts that are inherent to the *de novo* assembly process. The TransDecoder package is integrated into Trinity and identifies the coding-region from the assembled output. Because of this integration, as well as its ability to report multiple ORFs from a single transcript,

it is an ideal method for prediction and performed well as shown by the high annotation rate (85%) and relatively high number of full-length proteins identified (16,594). Most importantly, this package relies on multiple forms of evidence to validate the ORFs generated. We were able to provide training sequences based on test annotation runs of the transcripts from the most closely related (fully sequenced) plant relatives. Where sequence examples were not available, the inclusion of the Pfam database helped verify true translation products. These constraints were responsible for the decrease from 78,645 unique genes to 29,785 full and partial length ORFs.

Gene ontology is a dynamic tool for standardizing the plethora of BLAST descriptions assigned to thousands of sequences in a given set. Biologists often find the task of organizing these descriptions into broad classifications to describe their gene sets daunting. In this analysis, the custom BLAST searches were provided to Blast2GO and this application was able to work with a broad set of descriptions and provided at least one GO term to each of the annotated sequences queried. The category molecular function refer to the elemental activities of a gene product at the molecular level. This category is more specific than biological process and therefore defines individual functions at greater resolution. The top five molecular functions in walnut are ATP binding (1399), DNA binding (731), zinc ion binding (675), protein serine/threonine kinase activity (471), and metal ion binding (444). These are expected due to their regulatory nature in the cell and are similarly enriched in other comprehensive plant transcriptome studies (Vandepoele and Van de Peer, 2005; Nillson et al. 2010).

In the orthologous gene family analysis implemented by TRIBE-MCL in which 10,092 gene families were discovered among the 10 species evaluated (including *J. regia*), 4,897 gene families were conserved in all species while 31 gene families were unique to walnut. Of the

remaining eight that were not characterized as retroelements or without functional information, we identified a few conserved domains. Leucine Rich Regions (LRR) typically involved in protein-protein interactions are very prevalent in many species (Finn et al. 2014). Myb_DNA-bind_3 is a type of transcription factor. The NAM-associated domain (no apical meristem) proteins are involved in developmental processes, including formation of the shoot apical meristem, floral organs and lateral shoots, as well as hormonal control (Finn et al. 2014). PapC_N and PapC_C have transporter activity and regulatory roles in protein binding (Finn et al. 2014). In examining the BLAST descriptions of the sequences that make up these families in order to obtain more information, the majority of these sequences were labeled uninformative (hypothetical, predicted, etc) so further information could not be readily derived. A total of 362 gene families were conserved between the nine angiosperm species, walnut excluded. Among these gene families were the protein domains Oleosin and PLAT. Oleosins are structural proteins found in vascular plant oil bodies and plant cells. These proteins are believed to be involved in water-uptake and are found in oil bodies of seeds, but not fruits. PLAT is a protein domain found in lipid and membrane associated proteins. It is possible that these transcripts were yet present but not assembled correctly or that these genes take on different forms in *J. regia*.

The *Juglans regia* genome was assembled concurrently but separately from the transcriptome. At the conclusion of the genome assembly, the transcriptome was mapped back to the genome as a form of validation. The alignment was performed with the parameters 98% coverage and 98% identity. In an alignment between the same species, the identity should be highly conserved so the identity should not be lower than 98%. With these parameters, 85.45% of transcripts mapped back to the genome. As a further validation GMAP was run with 95% coverage and 95% identity and the alignment improved to 89.4% of transcripts mapping back to

the genome. Alternative splicing can potentially account for transcripts that did not align to the genome. Also contributing to unaligned transcripts could be the fragmented nature of an early draft assembly which currently represents 221,640 contigs. It is worth noting that from those transcripts that did not align at the lower identity, 100% did not have a functional annotation which could further indicate that these are artifacts of the *de novo* transcriptome assembly process. Since a reference genome was available, the alignment of another closely related species was not examined. In addition, the closer relatives of *Juglans regia* that have been sequenced to date (*B. nana* and *C. mollisima*) are also still in early draft states.

Walnut and all of the top twenty related species, as determined through annotation, are members of the Rosids. The well-studied Rosids are a major group of eudicots, and make up ~25% of angiosperms. This group is comprised of 17 orders and 176 families (Fay, 2013). The order of walnut is Fagale, which is most closely related to the order Curcurbitales. Only two other Fagales have been sequenced thus far, *Betula nana* and *Castanea mollisima*. In this analysis, *Cucumis sativus* (of the order Curcurbitales*)* annotated 1,736 assembled genes in *J. regia*. The position of the *Vitales* order is unknown within this group, but *Vitis vinifera* had the most annotated genes in common with *J. regia* (5760). The discrepancy between closely related organisms through sequencing and phylogenetic relationships has more to do with the genomes assembled and annotated thus far as well as their representation in the databases we were searching against. *V. vinifera*, the common grapevine, has been extensively studied, particularly in the International Grape Genome Project. Due to the high quality nature of this woody crop genome assembly and characterization, the numerous alignments between the common grapevine and walnut genes is unsurprising.

Transcriptomes are inherently limited. Expression profiles change based on a multitude of factors including, but not limited to developmental stage, time (day versus night), tissue source, and environmental conditions. It is also very difficult to replicate a transcriptome because of these changes in gene expression. In this sense, a transcriptome for an organism is never truly complete. We are further challenged by working with a non-model organism and a full *de novo* assembly compiled from older Illumina Genome Analyzer short read technology. The coverage and read lengths are less than what is available through the Illumina and other platforms today. This study pooled RNA from 19 different tissues to create the first comprehensive transcriptome for *J. regia*. Future experiments could focus on single tissues and the effect of a condition or developmental stage on expression to better profile the gene space and its interactions. In addition, future comparisons between tissue types would benefit from replicated libraries in order to fully evaluate expression differences.

VII. References

1. Quercus robur (2015). Unpublished manuscript.
2. Chinese Chestnut Genome and QTL Sequences v1.0 (2015). Unpublished manuscript.
3. Genome Reju (2015). Unpublished manuscript.
4. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology, 215*(3), 403-410.
5. Andrews, S. (2010). FastQC: A quality control tool for high throughput sequence data. *Reference Source,*
6. Ansorge, W. J. (2009). Next-generation DNA sequencing techniques. *N Biotechnol, 25*(4), 195-203.
7. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: Tool for the unification of biology. the gene ontology consortium. *Nat Genet, 25*(1), 25-29.
8. Association, T. C. W. G. *The california walnut*. Los Angeles, California:
9. Batzoglou, S., Jaffe, D. B., Stanley, K., Butler, J., Gnerre, S., Mauceli, E., et al. (2002). ARACHNE: A whole-genome shotgun assembler. *Genome Res, 12*(1), 177-189.
10. Baumgartner, K., Fujiyoshi, P., Browne, G. T., Leslie, C., & Kluepfel, D. A. (2013). Evaluating paradox walnut rootstocks for resistance to armillaria root disease. *Hortscience, 48*(1), 68-72.
11. Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for illumina sequence data. *Bioinformatics*.

12. Bolger, M. E., Weisshaar, B., Scholz, U., Stein, N., Usadel, B., & Mayer, K. F. X. (2014). Plant genome sequencing - applications for crop improvement. *Curr Opin Biotechnol, 26*, 31-37.
13. Boyle, E. I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J. M., et al. (2004). GO::TermFinder--open source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes. *Bioinformatics, 20*(18), 3710-3715.
14. Brendel, V., Bucher, P., Nourbakhsh, I. R., Blaisdell, B. E., & Karlin, S. (1992). Methods and algorithms for statistical analysis of protein sequences. *Proceedings of the National Academy of Sciences, 89*(6), 2002-2006.
15. Burtin, P., Jay-Allemand, C., Charpentier, J. -. P., & Janin, G. (1998). Natural wood colouring process in juglans sp.(J. nigra, J. regia and hybrid J. nigra 23$\times$ J. regia) depends on native phenolic compounds accumulated in the transition zone between sapwood and heartwood. *Trees, 12*(5), 258-264.
16. Carpentier, S. C., Panis, B., Vertommen, A., Swennen, R., Sergeant, K., Renaut, J., et al. (2008). Proteome analysis of non-model plants: A challenging but powerful approach. *Mass Spectrom Rev, 27*(4), 354-377.
17. Chang, Z., Wang, Z., & Li, G. (2014). The impacts of read length and transcriptome complexity for de novo assembly: A simulation study. *PLoS One, 9*(4), e94825.
18. Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L., & Rice, P. M. (2010). The sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res, 38*(6), 1767-1771.

19. Compeau, P. E. C., Pevzner, P. A., & Tesler, G. (2011). How to apply de bruijn graphs to genome assembly. *Nat Biotechnol, 29*(11), 987-991.

20. Conesa, A., & G\otz, S. (2008). Blast2GO: A comprehensive suite for functional analysis in plant genomics. *Int J Plant Genomics, 2008*, 619832.

21. Dandekar, A. (2005). Juglandaceae. *Biotechnology of Fruit and Nut Crops,* , 297-324.

22. Dietze, H., Berardini, T. Z., Foulger, R. E., Hill, D. P., Lomax, J., Osumi-Sutherland, D., et al. (2014). TermGenie--a web-application for pattern-based ontology class generation. *Journal of Biomedical Semantics, 5*(1), 48.

23. Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics, 29*(1), 15-21.

24. Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics, 26*(19), 2460-2461.

25. Edstrom, J. P., Krueger, W., & Reil, W. (2004). *Presentation* English Walnut Production on Marginal Soils

26. Enright, A. J., Van Dongen, S., & Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research, 30*(7), 1575-1584.

27. Fakruddin, M. R., Chowdhury, A., Hossain, N., Mahajan, S., & Islam, S. (2013). Pyrosequencinga next generation sequencing technology. *World Applied Sciences Journal, 24*(12), 1558-1571.

28. Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., et al. (2014). Pfam: The protein families database. *Nucleic Acids Res, 42*(Database issue), D222-30.

29. Fjellstrom, R. G., Parfitt, D. E., & McGranahan, G. H. (1994). Genetic relationship and characterization of persian walnut (*Juglans regia* L.) cultivars using restriction fragment length polymorphisms (RFLPs). *American Society for Horticultural Science (USA),*

30. Gotz, S., Arnold, R., Sebastian Leon, P., Martin-Rodriguez, S., Tischler, P., Jehl, M. A., et al. (2011). B2G-FAR, a species-centered GO annotation repository. *Bioinformatics, 27*(7), 919-924.

31. Gotz, S., Garcia Gomez, J. M., Terol, J., Williams, T. D., Nagaraj, S. H., Nueda, M. J., et al. (2008). High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res, 36*(10), 3420-3435.

32. Gordon, D., Abajian, C., & Green, P. (1998). Consed: A graphical tool for sequence finishing. *Genome Res, 8*(3), 195-202.

33. Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nature Biotechnology, 29*(7), 644-652.

34. Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., et al. (2013). De novo transcript sequence reconstruction from RNA-seq using the trinity platform for reference generation and analysis. *Nat Protoc, 8*(8), 1494-1512.

35. Halstead, T. (2014). *Tree nuts: World markets and trade*United States Department of Agriculture Foreign Agricultural Service.

36. Harismendy, O., Ng, P. C., Strausberg, R. L., Wang, X., Stockwell, T. B., Beeson, K. Y., et al. (2009). Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol, 10*(3), R32.

37. Initiative, A. G. (2000). Analysis of the genome sequence of the flowering plant arabidopsis thaliana. *Nature, 408*(6814), 796.

38. Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., et al. (2014). InterProScan 5: Genome-scale protein function classification. *Bioinformatics, 30*(9), 1236-1240.

39. Joshi, N. A., & Fass, J. N. (2011). *Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files* Software.

40. Keles, H., Akca, Y., & Ercisli, S. (2014). Selection of Promising Walnut Genotypes (*Juglans regia* L.) FROM INNER ANATOLIA. *Acta Scientiarum Polonorum-Hortorum Cultus, 13*(3), 167-175.

41. Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., & Salzberg, S. L. (2013). TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol, 14*(4), R36.

42. Langmead, B. *presentation* De Bruijn graph assembly

43. Letunic, I., Doerks, T., & Bork, P. (2015). SMART: Recent updates, new developments and status in 2015. *Nucleic Acids Res, 43*(Database issue), D257-60.

44. Li, L., Stoeckert, C. J., & Roos, D. S. (2003). OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res, 13*(9), 2178-2189.

45. Li, R. -. Q., Chen, Z. -. D., Lu, A. -. M., Soltis, D. E., Soltis, P. S., & Manos, P. S. (2004). Phylogenetic relationships in fagales based on DNA sequences from three genomes. *International Journal of Plant Sciences, 165*(2), 311-324.

46. Li, Z., Chen, Y., Mu, D., Yuan, J., Shi, Y., Zhang, H., et al. (2012). Comparison of the two major classes of assembly algorithms: Overlap-layout-consensus and de-bruijn-graph. *Brief Funct Genomics, 11*(1), 25-37.

47. Liao, Z., Feng, K., Chen, Y., Dai, X., Li, S., & Yin, T. Genome-wide discovery and analysis of single nucleotide polymorphisms and insertions/deletions in *Juglans regia* L. by high-throughput pyrosequencing.

48. Lin, Y., Li, J., Shen, H., Zhang, L., Papasian, C. J., & Deng, H. W. (2011). Comparative studies of de novo assembly tools for next-generation sequencing technologies. *Bioinformatics, 27*(15), 2031-2037.

49. Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., et al. (2012). Comparison of next-generation sequencing systems. *J Biomed Biotechnol, 2012*, 251364.

50. Loman, N. J., Misra, R. V., Dallman, T. J., Constantinidou, C., Gharbia, S. E., Wain, J., et al. (2012). Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol, 30*(5), 434-439.

51. Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., et al. (2012). SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler. *Gigascience, 1*(1), 18.

52. Mardis, E. R. (2008). The impact of next-generation sequencing technology on genetics. *Trends Genet, 24*(3), 133-141.

53. Martin, J. A., & Wang, Z. (2011). Next-generation transcriptome assembly. *Nature Reviews Genetics, 12*(10), 671-682.

54. Mashayekhi, F., & Ronaghi, M. (2007). Analysis of read length limiting factors in pyrosequencing chemistry. *Analytical Biochemistry, 363*(2), 275-287.

55. Mauricio, R. (2005). Can ecology help genomics: The genome as ecosystem? *Genetica, 123*(1-2), 205-209.

56. McCarthy, A. (2010). Third generation DNA sequencing: Pacific biosciences' single molecule real time technology. *Chem Biol, 17*(7), 675-676.

57. Mi, H., Muruganujan, A., & Thomas, P. D. (2013). PANTHER in 2013: Modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res, 41*(Database issue), D377-86.

58. Miller, J. R., Delcher, A. L., Koren, S., Venter, E., Walenz, B. P., Brownley, A., et al. (2008). Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics, 24*(24), 2818-2824.

59. Min, X. J., Butler, G., Storms, R., & Tsang, A. (2005). OrfPredictor: Predicting protein-coding regions in EST-derived sequences. *Nucleic Acids Res, 33*(Web Server issue), W677-80.

60. Minoche, A. E., Dohm, J. C., & Himmelbauer, H. (2011). Evaluation of genomic high-throughput sequencing data generated on illumina HiSeq and genome analyzer systems. *Genome Biol, 12*(11), R112.

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

62. Myers, E. W., Sutton, G. G., Delcher, A. L., Dew, I. M., Fasulo, D. P., Flanigan, M. J., et al. (2000). A whole-genome assembly of drosophila. *Science, 287*(5461), 2196-2204.

63. Nabavi, S. F., Ebrahimzadeh, M. A., Nabavi, S. M., Mahmoudi, M., & Rad, S. K. (2011). Biological activities of *Juglans regia* flowers. *Revista Brasileira De Farmacognosia, 21*(3), 465-470.

64. Nicese, F. P., Hormaza, J. I., & McGranahan, G. H. (1998). Molecular characterization and genetic relatedness among walnut (*Juglans regia* L.) genotypes based on RAPD markers. *Euphytica, 101*(2), 199-206.

65. Nicese, F. P., Hormaza, J. I., & McGranahan, G. H. (1998). Molecular characterization and genetic relatedness among walnut (*Juglans regia* L.) genotypes based on RAPD markers. *Euphytica, 101*(2), 199-206.

66. Nilsson, L., M\uller, R., & Nielsen, T. H. (2010). Dissecting the plant transcriptome and the regulatory responses to phosphate deprivation. *Physiol Plant, 139*(2), 129-143.

67. Ozkan, G. Physical and chemical composition of some walnut (*Juglans regia* L.) genotypes grown in turkey. *Grasas y Aceites, 56*(2), 141-146.

68. Parkinson, J., & Blaxter, M. (2009). Expressed sequence tags: An overview. *Methods Mol Biol, 533*, 1-12.

69. Pereira, J. A., Oliveira, I., Sousa, A., Valent\ao, P., Andrade, P. B., Ferreira, I. C. F. R., et al. (2007). Walnut (juglans regia L.) leaves: Phenolic compounds, antibacterial activity and antioxidant potential of different cultivars. *Food Chem Toxicol, 45*(11), 2287-2295.

70. Pinney, K., & Polito, V. S. (1983). English walnut fruit growth and development. *Scientia Horticulturae, 21*(1), 19-28.

71. Pop, M., & Salzberg, S. L. (2008). Bioinformatics challenges of new sequencing technology. *Trends Genet, 24*(3), 142-149.

72. Preece, J. E., Vahdati, K., Ibanez, A. M., Compton, P. J., Tran, Q., Gunawan, D., et al. (2012). Regeneration systems for pyramiding disease resistance into walnut rootstocks. *Walnut Research Conference,* pp. 57-64.

73. Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., et al. (2012). A tale of three next generation sequencing platforms: Comparison of ion torrent, pacific biosciences and illumina MiSeq sequencers. *BMC Genomics, 13*, 341.

74. Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S. D., et al. (2010). De novo assembly and analysis of RNA-seq data. *Nat Methods, 7*(11), 909-912.

75. Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences, 74*(12), 5463-5467.

76. Schonenberger, J., & von Balthazar, M. (2006). Reproductive structures and phylogenetic framework of the rosids-progress and prospects. *Plant Systematics and Evolution, 260*(2-4), 87-106.

77. Schulz, M. H., Zerbino, D. R., Vingron, M., & Birney, E. (2012). Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics, 28*(8), 1086-1092.

78. Schuster, S. C. (2008). Next-generation sequencing transforms today's biology. *Nat Methods, 5*(1), 16-18.

79. Sharma, R. M., Kour, K., Singh, B., Yadav, S., Kotwal, N., Rana, J. C., et al. (2014). Selection and characterization of elite walnut ('*Juglans regia*'L.) clone from seedling origin trees in north western himalayan region of india.

80. Stein, L. (2001). Genome annotation: From sequence to biology. *Nat Rev Genet, 2*(7), 493-503.

81. Surget-Groba, Y., & Montoya-Burgos, J. I. (2010). Optimization of de novo transcriptome assembly from next-generation sequencing data. *Genome Res, 20*(10), 1432-1440.

82. Sze-Tao, K. W. C., & Sathe, S. K. (2000). Walnuts (*Juglans regia* L): Proximate composition, protein solubility, protein amino acid composition and protein in vitro digestibility. *Journal of the Science of Food and Agriculture, 80*(9), 1393-1401.

83. Tuskan, G. A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., et al. (2006). The genome of black cottonwood, populus trichocarpa (torr. \& gray). *Science, 313*(5793), 1596-1604.

84. USDA. (2013). *Agricultural statistics*. Washington, DC: United States Department of Agriculture.

85. Vandepoele, K., & Van de Peer, Y. (2005). Exploring the plant transcriptome through phylogenetic profiling. *Plant Physiology, 137*(1), 31-42.

86. Veretnik, S., Gu, J., & Wodak, S. (2009). Identifying structural domains in proteins. *In Genny Gu and Philip Bourne Structural Bioinformatics Second Edition.Wiley-Blackwell,* 485-513.

87. Wang, N., Thomson, M., Bodles, W. J. A., Crawford, R. M. M., Hunt, H. V., Featherstone, A. W. (2013). Genome sequence of dwarf birch (betula nana) and cross-species RAD markers. *Mol Ecol, 22*(11), 3098-3111.

88. Wegrzyn, J. L., Liechty, J. D., Stevens, K. A., Wu, L. -. S., Loopstra, C. A., Vasquez-Gross, H. A. (2014). Unique features of the loblolly pine (pinus taeda L.) megagenome revealed through sequence annotation. *Genetics, 196*(3), 891-909.

89. Wu, J., Gu, Y. Q., Hu, Y., You, F. M., Dandekar, A. M., Leslie, C. A., et al. (2012). Characterizing the walnut genome through analyses of BAC end sequences. *Plant Mol Biol, 78*(1-2), 95-107.

90. Wu, T. D., & Watanabe, C. K. (2005). GMAP: A genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics, 21*(9), 1859-1875.

91. Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., et al. (2014). SOAPdenovo-trans: De novo transcriptome assembly with short RNA-seq reads. *Bioinformatics, 30*(12), 1660-1666.

92. Zdobnov, E. M., & Apweiler, R. (2001). InterProScan--an integration platform for the signature-recognition methods in InterPro. *Bioinformatics, 17*(9), 847-848.

93. Zhang, J., Ruhlman, T. A., Mower, J. P., & Jansen, R. K. (2013). Comparative analyses of two geraniaceae transcriptomes using next-generation sequencing. *BMC Plant Biol, 13*, 228.

94. Zhang, R., Zhu, A., Wang, X., Yu, J., Zhang, H., Gao, J., et al. (2010). Development of juglans regia SSR markers by data mining of the EST database. *Plant Molecular Biology Reporter, 28*(4), 646-653.

95. Zhang, Z. Y., Han, J. W., Jin, Q., Wang, Y., Pang, X. M., & Li, Y. Y. (2013). Development and characterization of new microsatellites for walnut (*Juglans regia*). *Genet Mol Res, 12*(4), 4723-4734.

96. Zhao, Q. Y., Wang, Y., Kong, Y. M., Luo, D., Li, X., & Hao, P. (2011). Optimizing de novo transcriptome assembly from short-read RNA-seq data: A comparative study. *BMC Bioinformatics*.