Spring 2017

Transcriptome Profiling and Long Non-Coding Rna Identification in Grapevine

Zachary Noel Harris

TRANSCRIPTOME PROFILING AND LONG NON-CODING RNA IDENTIFICATION IN GRAPEVINE

A Master's Thesis

Presented to

The Graduate College of

Missouri State University

In Partial Fulfillment

Of the Requirements for the Degree

Master of Science, Biology

By

Zachary Noel Harris

May 2017

TRANSCRIPTOME PROFILING AND LONG NON-CODING RNA IDENTIFICATION IN GRAPEVINE

Biology

Missouri State University, May 2017

Master of Science

Zachary Noel Harris

ABSTRACT

Next-generation sequencing technologies have provided access to vast quantities of nucleic acid sequence data. The resulting wealth of information enables biologists to address complex biological questions in species for which a high-quality well-annotated reference genome sequence has yet to be generated. The cultivated grapevine, *Vitis vinifera*, has a relatively poorly annotated reference genome. In addition, it is a highly heterozygous species which further hinders the annotation of its genome and the characterization of its transcriptome. Here, I annotated Version 2 of the 12X *V. vinifera* genome using RNA-seq data derived from the variety ‘Riesling’ by employing the most up-to-date computational methods. The results provide the first annotation of ‘Riesling’ and the first profile of its transcriptome in relation to the reference transcriptome of the model grape variety ‘Pinot Noir’. In addition, I develop a computational pipeline for the identification of long non-coding RNAs (lncRNAs) in non-model plant species that lack well-sequenced reference genomes. This pipeline was then applied to ‘Riesling’ RNA-seq data for the first analysis of lncRNAs in that variety.

KEYWORDS: transcriptome, genome annotation, grapevine, lncRNA, computational biology, gene prediction.

This abstract is approved as to form and content

László G. Kovács, PhD
Chairperson, Advisory Committee
Missouri State University

TRANSCRIPTOME PROFILING AND LONG NON-CODING RNA IDENTIFICATION IN GRAPEVINE

By

Zachary Noel Harris

A Master's Thesis
Submitted to the Graduate College
Of Missouri State University
In Partial Fulfillment of the Requirements
For the Degree of Master of Science, Biology

May 2017

Approved:

László G. Kovács, PhD

Sean P. Maher, PhD

Matthew R. Siebert, PhD

Julie Masterson, PhD: Dean, Graduate College

ACKNOWLEDGEMENTS

This project was made possible by Dr. Jason Londo at the USDA-ARS Grape Research Unit who provided RNA-seq data and guidance for my work. Without Dr. Londo's help, this thesis never would have not become a reality. The support and encouragement from my thesis advisor, Dr. Laszlo Kovacs, was another factor that contributed to the accomplishment of this work. From listening to me rant about the minutiae of various aspects of my work, to encouraging me to pursue conferences overseas, he supported every facet of my time here. Further gratitude goes to the rest of my committee, Drs. Sean Maher and Matthew Siebert, for providing invaluable advice and support at every required moment. I would also like to the Kovacs lab for providing an excellent environment for me to work and constant intellectual platform on which ideas could be brainstormed. Of the team, I would especially like to thank Chelsea Campbell for welcoming me into the lab, and over time, becoming one of the most important friends in my life. New to the lab, a few undergrads have been particularly important in furthering projects for which I have had little time. They include Devin Ellebracht, Stephanie Chau, and Thomas Freeman, without whom the project may have gone uncompleted. I would also like to thank the faculty and staff of Missouri State University, particularly those in the Biology Department, as they have cultivated an excellent environment for the development and education of young scientists. Finally, I would like to thank my family for the years of support they have provided as I kicked around ideas of what I wanted to be when I grew up. I would especially like to thank my mother, Kimberly, for encouraging me to never stop asking questions.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

RNA-seq as a Tool for Transcriptome Characterization

The Central Dogma of Molecular Biology states that, in living organisms, biological information stored in the form of DNA is transcribed into RNA, and then translated to protein. The protein is then directly responsible for the expression of a phenotype in the organism. This directionality in the flow of biological information was deduced even before the detailed molecular mechanisms of this process were understood [1]. The Central Dogma of Molecular Biology has been the ground work upon which molecular genetics has been built. The understanding of these mechanisms has shed light on the function of genes and the regulation of their expression. This insight, coupled with recent transcriptomics methods, most notably RNA sequencing (RNA-seq), lead to the identification of vast array of previously unknown types of genetic information and afforded us a global view of the genomic landscape in action. This new comprehensive and functional view of the genome paved the way for the discovery of vast amounts of transcriptional information and led to the recognition that gene regulation is immensely complex even in the simplest eukaryotic genome.

The enabling technology behind RNA-seq was next-generation sequencing, the most common platform of which is the Illumina sequencing-by-synthesis method. The Illumina-based RNA-seq workflow starts with extraction of total RNA from the organism of interest. This is followed by the processing of the resulting sample to eliminate unwanted RNA species, such as rRNA or non-polyA RNA by such methods as ribo-depletion or poly-A capture. The remaining molecules are then fragmented and primed for amplification with random hexamers. Generated complimentary DNA (cDNA) are then tagged with adapters and barcoded. These adapters are then used as the basis for sample identification (indexing). The nucleotide sequence of the resulting adaptor-tagged cDNA fragments is then determined on an Illumina sequence analyzer in a massively parallel manner.

The magnitude of the information produced poses a new problem to biologists as they try to find simultaneously precise and efficient ways of analyzing these data. In terms of RNA-seq analysis, two general methodologies exist for the interpretation of data: genome-guided and genome independent (or *de novo*) approaches. In the genome-guided approach, sequences are aligned to a reference genome, assembled into transcripts, merged across assemblies, and analyzed for the biological question being addressed (Figure 1) [2–6]. These questions could range from the assembly of a novel transcriptome [7, 8] to studying differential gene expression under various treatment conditions [9, 10]. But the advent of new software now enables biologists to assemble transcripts *de novo*, that is, in the absence of a reference genome (demonstrated in Figure 1) [11, 12]. This approach enables us to address questions in organisms for which reference genomes do not exist or are of poor quality, thereby expanding the reach of molecular biology to virtually any organism.

The choice of analysis (whether genome-guided or genome-independent) often depends on the availability of computational resources and the genetic features of the organism under study. Genome-guided approaches are favored when computational resources are limited or when there is little expected nucleotide sequence variation in the organism of question. Genome-guided approaches can also be superior when sample-to-sample comparisons are made as the data are linked to some specific genomic coordinates. However, if the computational resources are available, *de novo* approaches facilitate the assembly of sequences in their own context and not in the preset context of the reference genome. This approach maximizes the ability to identify novel splice variants and pinpoint genetic variation between organisms. Whether the optimal approach is genome-guided or *de novo* is determined by the study organism and the biological question at hand.

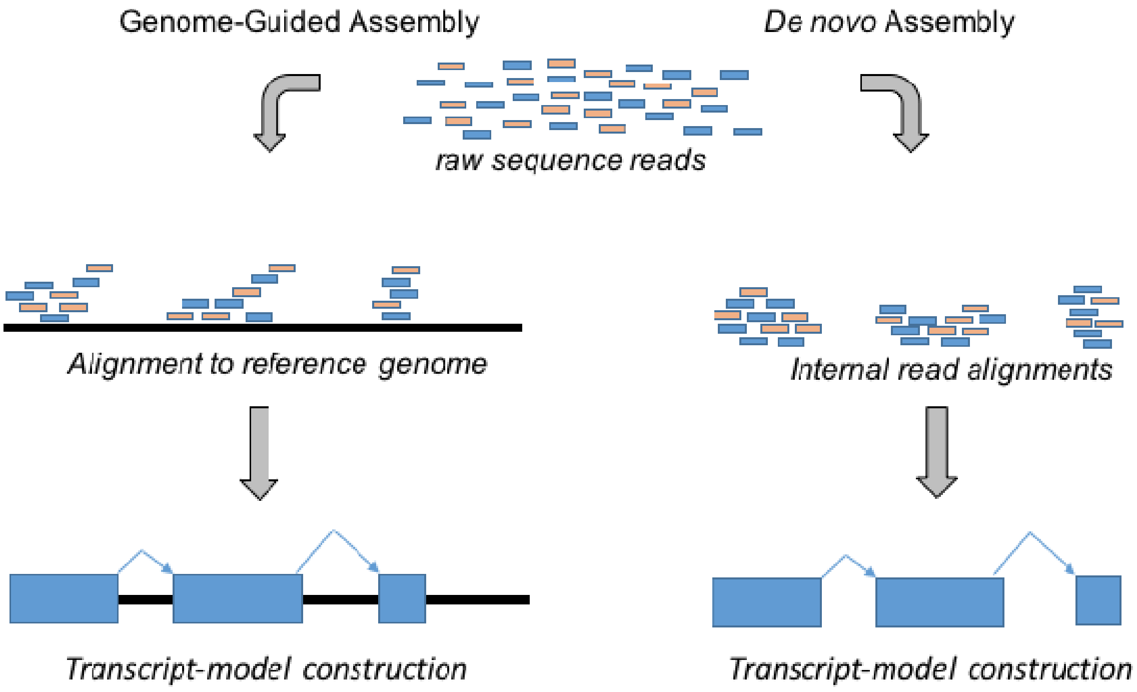
Genome-Guided Assembly
De novo Assembly
raw sequence reads
Alignment to reference genome
Internal read alignments
Transcript-model construction
Transcript-model construction

Figure 1. Algorithmic approaches to genome-guided and genome-independent transcriptome assembly.

At present, various computer programs exist for the *de novo* assembly of transcriptomes. These include Bridger [13], Oases [14], SOAPdenovo-Trans [12], TransAByss [15], and Trinity [11]. Reviews of the various software often conclude that choosing the optimal *de novo* assembler is to be based on what is valued in any given analysis [16, 17]. For example, using a single *k-mer* scheme, SOAPdenovo-Trans was able to outcompete both Trinity and Oases in gene construction at 95% coverage, but it paled in comparison to Trinity at 100% coverage and the ability to make use of strand-specific information [12]. Analyses comparing polyploids to diploids demonstrated that Trinity was able to outperform SOAPdenovo-Trans at full-length gene assembly and performed better than TransAByss at assembling complex tetraploid transcriptomes [16]. In a killifish analysis, various tools were used under various k-mer strategies, and it was found that even using a single k-mer strategy, Trinity was still able to find the most full-length transcripts [17].

Genome Annotation and Gene Model Prediction

While this early step in the analysis of RNA-seq data is highly important, the downstream analysis of these data have the potential to answer highly complex questions that were previously unyielding to scientific inquiry. For example, they can be used to improve upon current genome annotations [18, 19], and train new gene prediction models [20, 21]. Genome annotation is the procedure of defining known genetic elements in genome space. For example, characterizing the locations of repetitive elements, and non-coding and protein-coding gene spaces. Gene prediction can be accomplished by training machine learning algorithms and using higher-order statistical approaches to characterize regions of the genome that appear to have function. Together, these strategies can be used to address such wide-ranging tasks as functional annotation and genome evolution.

Genome annotation is hardly a trivial problem, as genome sizes can range from 5.4 kbp in the case of the ϕ_{X174} bacteriophage [22] to 23.2 Gpb in the case of the loblolly pine [23]. Defining any given genomic characteristic for either of these extremes represents a challenging computational problem. The two most common annotated spaces in genomes are repeated elements and gene spaces. Repeated elements often are identified by probing databases of previously described repeat elements using hidden Markov model approaches [24–26]. Gene spaces are often defined by alignments of expressed sequences to the reference genome. The characterization of both repeated elements and gene spaces tend to be computationally expensive, but can be accomplished successfully. Many curated genome databases [27, 28] employ their own algorithms to annotate genomes using the wealth of data they maintain. Smaller projects can benefit from software, such as Maker, that carry out the same processes relying on data in publicly accessible databases.

While expressed sequence alignment to the genome provides evidence of gene models, they are often incomplete, as processes like alternative splicing alter gene space in a tissue- and development-specific manner. Several programs have been written to study aligned gene models and predict the presence of novel gene spaces. One widely used program, AUGUSTUS, uses Hidden Markov Models to draft potential gene structures. For example, each gene structure should contain several key items in a particular order including a start codon, some number of exons, donor splice sites, acceptor splice sites, introns, and a stop codon [21]. This statistical framework, coupled with aligned expressed sequence information, can inform the prediction of novel genes. Other programs, such as SNAP, have incorporated similar frameworks but tend to focus more on such genome-specific features as AT- or CG-content in addition to splice sites [20].

Non-Coding RNA Species

Arguably, the most significant recent advance of molecular biology is the finding that an unexpectedly large portion of the nuclear genome is transcribed into RNA. In humans, the extent of transcribed information accounts for as much as ~80% of the entire genome [29]. Most of this information does not code for polypeptides. A subset of these non-coding RNA species has been known for several decades and were thought to play a supporting role in facilitating the flow of genetic information to proteins. Among these species were ribosomal RNA (rRNA) and transfer RNA (tRNA), both of which are strongly conserved across diverse branches of life. Subsequently, such short non-coding RNAs species have been discovered as micro-RNA (miRNA) [30, 31], short-interfering RNA (siRNA) [32], small nuclear RNA (snRNA) [33], small cajal body RNA (scaRNA) [34], and piwi-protein-interacting RNA (piRNA)[35], some of which play important gene regulatory roles, others are components of essential ribonucleoprotein complexes [36].

The function of many of these RNA species has been well characterized. For example, miRNA is processed from transcripts, termed pri-miRNA, cleaved by the RNase Type-III endonuclease DROSHA into precursor miRNA (pre-miRNA), which then is transported from the nucleus through exportin-5, and finally processed by a Dicer complex, resulting in 21 to 24 nucleotide-long miRNAs [37–39]. We also know the processes by which siRNA are processed[40], but the mechanism for the genesis of piRNA remains unknown. Some of these RNA species are not thought to be handled uniformly across transcription, but may have more complex, variable processing patterns. Therefore, it is prudent to identify these species rather than look for large-scale pathways by which they are processed. Recent analyses have elucidated the presence of an entirely new category of non-coding transcripts, the long non-coding RNAs (lncRNAs) [36].

lncRNA Identification Pipelines and Techniques

The key approach to the identification of non-coding RNA (especially long non-coding RNA), is to assess the ability of the transcript to encode proteins [41–45]. The software Coding Potential Calculator (CPC) builds a support vector machine (SVM) based on six features. Three of these features rely on BLAST searches to identify most closely related homologs, and the remaining three features deal with the presence and quality of an open reading frame (ORF) [46, 47]. This method has been considered optimal, because it is relatively computationally inexpensive, requiring only a BLAST search and an ORF analysis.

Other methods rely on the alignment of putative non-coding transcripts to transcripts derived from other biological species. This alignment allows the generation of phylogenetic trees whereby sequence information, branch length, tree topology and codon substitution are used to model coding potential [48]. The main program, named PhyloCSF, by which these methods are implemented is computationally expensive, taking nearly 50 hours to complete when implemented in a 60-way genome alignment on mammals [49]. The time required to complete PhyloCSF coupled with the lack of genomic information for many species makes this method excessively cumbersome for many biological analyses.

In addition to these analytical methods, there have been efforts to build models that rely on classifying sequences based on nucleotide frequencies alone. Two programs have attempted to analyze nucleotide frequencies in similar manners. The first program, named Coding Non-Coding Index (CNCI), uses adjoining nucleotide triplets (hexamers); the second, named PLEK uses k-mer frequencies where *k* can take on the value from 1 to 5 [49, 50]. Both models calculate the frequency of nucleotide usage and use these frequencies to build a support vector machine. Using these frequencies, CNCI builds a model with up to 4,096 parameters, and PLEK builds a model with up to 1,394 parameters. Both models have proved to be very effective in classifying species evolutionarily close to humans, but an analysis of distantly related taxa requires the use of new model-training data. Unfortunately, much of these data are unavailable or only predicted to be present.

Current computational pipelines for the identification of lncRNAs employ some variation of a process starting with raw RNA-seq reads being aligned to a reference genome by a short read aligner, usually TopHat [5]. These alignments are used for the construction of transcripts via programs like Cufflinks [6]. These transcripts can then be used for homology searches using BLAST [47] or HMMER [51] against multi-species, highly collaborative, expansive reference databases [52, 53]. Known homologies are filtered out and further filtering is executed by one or more of the identification tools above.

Efforts in the annotation of these regions in model organisms has been expansive. For example, ~1,600 lncRNA were identified in *Mus musculus* by analyzing chromatin-state maps outside of known protein coding areas of the genome [54]. More than 8,000 lncRNAs were identified in human by use of RNA-seq [55], and over 9,600 genomic loci were identified in humans as giving rise to 15,512 lncRNA transcripts in the project GENCODE [29]. GENCODE highlighted problems with these analyses, notably, the apparent inability to generate results that are comparable across methods for identification. To underscore this point, GENCODE revealed the failure to independently validate the presence of 70% of 5,446 previously described lncRNA loci in humans [41], 61% of the 1,600 long intergenic non-coding RNAs identified [55], and 88% of very long intergenic non-coding RNAs (vlincRNAs) in humans [56].

The lack of reproducibility in the most heavily studied of species calls into question the validity of the methods used thus far. With such a low recovery rate, how can we be sure that these methods are not generating mostly RNA molecules “accidentally” transcribed by RNA polymerase enzymes, or simply finding artifacts of the assembly algorithms? The question prompted a simple analysis for the identification of lncRNA in rainbow trout [45]. Here, lncRNA were considered present if they occurred in multiple, independently sequenced data sets. This comparison works to reduce the false discovery rate under the assumptions that sequences annotated as lncRNAs in multiple RNA-seq data sets, are likely to be true transcripts.

Long Non-Coding RNAs in Plants

Plant lncRNAs have attracted less attention than those in animals, and therefore are poorly characterized in their modes of operation. But even with this paucity of attention, a few specific lncRNAs have been analyzed. For example, it is known that the cold-assisted intronic non-coding RNA (COLDAIR) binds to a protein in the PRC2 complex resulting in the methylation of a lysine residue at position 27 of the third histone [57]. Further, the lncRNA induced by phosphate starvation 1 (IPS1) serves as a binding site for miR399, suppressing its function [58]. Of course, for the discovery of these mechanisms, we must first compile a comprehensive catalog of putative lncRNA sequences. Therefore, a broad analysis for the identification of these species *en masse* is necessary.

Grape Genomics and Transcriptomics

In 2007, the first complete genome sequence of a fruit crop was published. This was the genome of the cultivated grapevine *Vitis vinifera* cv. 'Pinot Noir' [59]. This genome was named the 8X-coverage reference genome sequence. Accompanying this achievement was the first genome annotation of grapevine, using protein alignments from the Uniprot-Swiss-Prot database, *ab initio* gene predictors, and 301 known gene sequences from grapevine. The resultant annotation model found 30,434 putative genes. This model was improved upon using gene predictors trained with 601 known gene sequences and expressed sequence tags generate the v1 Centro di Ricerca Interdipartimentale per le Biotecnologie Innovative (CRIBI) annotation consisting of 29,971 unique genes [60]. This annotation was used as the reference for the transcriptome assembly of *Vitis vinifera* cv. 'Corvina' in 2013, where only 15,161 genes were validated [61]. From 'Corvina', 2,321 new gene models and 9,463 novel isoforms were proposed. These differences were attributed to the genotypic variation and poor annotation of the reference. This idea prompted the reannotation of the grapevine genome by use of more complex data in the form of RNA-seq from the cultivar 'Cabernet sauvignon', and rootstock cultivars ‘1103P’ and ‘M4’ [10]. This approach utilized the genome annotation tool PASA coupled with the gene prediction tool AUGUSTUS to generate a final annotation of 32,922 unique protein-coding genes. This model is now referred to as the Version 2 (v2) annotation.

In 2016, the genome of the *V. vinifera* cultivar, ‘Cabernet sauvignon’ was sequenced [62]. This genome sequence maintained a Benchmarking Universal Single Copy Orthologue (BUSCO) score of 80%, but only 16,981 of the 29,971 v1 annotation genes aligned to the reference. This corroborates previous conclusions that there is extensive genetic variation among grape cultivars within the species *V. vinifera*, and it is readily detectable even at the level of the transcriptome.

Objective and Justification

Here, I present a reannotation of the grapevine genome based on prior predictions of gene models. I expanded this annotation by training more complex prediction models and re-annotated the genome using RNA-seq data specific to a different cultivar of grape, ‘Riesling’. Because the annotation software is unable to detect putative non-coding regions, I set out to construct a computational a pipeline that can *de novo* identify lncRNAs from RNA-seq data. The resulting pipeline has the capacity to solve several of the problems that currently plague lncRNA identification. First, it provides a standard order of operations for analysis; second, it eliminates the requirement for a high-quality reference genome; and third, it allows for the ability to answer questions that are more complicated than the number of lncRNAs encoded in a genome. These annotations, improved gene prediction models, and annotated lncRNA allow us to understand the grapevine at a fundamental level which may enable us to answer long-standing biological questions and sustain the cultivation of the crop in the face of global climate change.

The cultivated grapevine, *Vitis vinifera*, is one of the most economically important crops in the United States. The USDA-NASS Census of Agriculture in 2007 and 2012 indicated that, in the US fruit production sector, grapes are the highest-ranking crop in terms of both number of farms and the acreage of cultivated land [63]. This acreage led to the 2015 estimate that the United States produced 2.9 billion liters of wine, nearly 10.5% of the world's overall wine production [64]. In addition to being a beverage of increasing popularity in the New World and East Asia, wine also carries cultural significance in many European countries. Beyond wine, high value-added products such as juice, grape seed oil, dietary supplements, jelly, raisin, and fresh table grapes are also made from grapevine and further enhance the economic importance of the species. Consequently, a thorough understanding of the genomic content of this crop is well warranted.

METHODS

Data Acquisition and Characterization

RNA-seq data were obtained from the USDA-ARS Grape Genetics Research Unit. These data were derived from RNA of two ‘Riesling’ accessions, 588673 and Ventosa. The former accession is from the USDA-ARS cold hardy germplasm, and the latter from Ventosa Vineyards, both sites located in Geneva, New York. Accession 588673 is comprised of dormant bud, leaf, tendril, flower, rachis, unripe berry, and post-véraison berry tissues. Ventosa is comprised of tendril, flower, root, and oversampled leaf tissue composed of field-collected leaves, and leaves exposed to chilled and freezing temperatures. Collectively, these samples represent an atlas of gene expression for ‘Riesling’. RNA was extracted using the commercially available Sigma Spectrum RNA kit and were sequenced as 150-bp paired-end reads using an Illumia HiSeq2000 sequence analyzer at Cornell University.

Genome Annotation and Transcriptome Profiling

Raw RNA-seq reads from each sequenced library were assorted into bins corresponding to a directional, trial-specific barcode using the FASTX tool fastx_barcode_splitter. Reads then were processed via FastQC, and adapters were removed using the Trimmomatic tool. Several independent sequencing runs were conducted. Reads generated in different lanes then were concatenated into FASTA files representing all libraries sequenced from the left terminus and all libraries sequenced from the right terminus. These then were paired using the tool pairfq_lite [65]. The resulting transcripts were *de novo* assembled using the program Trinity v2.0.6 [11] with the following parameters: --seqType fq --max_memory 25G --SS_lib_type FR --CPU 1. Flags for --left and --right were given both paired and unpaired reads from pairfq_lite delimited by a comma.

Because some tissues were not available in both accessions, and because some differences may exist in these two clones, transcriptomes for each accession were assembled independently. The final transcriptome was assembled by first uniquely naming all Trinity accessions from one ‘Riesling’ accession followed by concatenation with the second accession to form a master Riesling.fasta file. The unmasked *V. vinifera* cv. ‘Pinot noir’ 12Xv2 reference genome, all Trinity-assembled transcripts, and the Uniprot-Swiss-Prot database as FASTA files then were passed to Maker v2.32 [19] for annotation. All settings applied can be found in the combined_maker_opts.ctl file hosted at the GitHub link above, with the most notable change from default being the declaration of the “hidden setting” est_forward=1. FASTA. Finally, gff3 files were merged across the entire genome using the *fasta_merge* and *gff3_merge* tools in the Maker suite to generate a preliminary transcriptome.

Generating Gene Prediction Models:

Using the Maker framework, gene prediction models were created for future analysis of ‘Riesling’ data. Using all Maker-generated transcripts, I trained SNAP v1.0beta.17 [20] to generate a gene prediction model following the guidelines in the Maker Wiki [66]. The SNAP model was trained twice, iteratively. The first model was trained on the preliminary transcriptome, and the second was trained on that output. The resultant model, named Riesling.hmm, can be found in the GitHub repository described above.

Protein Domain Characterization and Functional Annotation of the Transcriptome

A boilerplate SQLite database (also available on GitHub) was constructed by tying each Gene Ontology [67] term to its respective class and function, and tying each Pfam domain to its respective Gene Ontology term. Protein domains then were characterized against Pfam30.0 [52] as a reference protein database using the hmmscan tool of the HMMER v3.1b2 software suite [68]. The program was executed with default parameters and then filtered for an E-value (a similarity metric) of less than or equal to 1×10^{-5}. Transcript accessions then were tied to GO terms, classes, functions using 'inner join' statements in SQlite. A script for this analysis has been provided in the GitHub repository. In addition to Pfam comparisons, both the annotated set of proteins and the 'non-overlapping' set of proteins were searched against the reference Uniprot-Swiss-Prot protein database and the annotated set was searched against Uniprot-Uniref90. Searches were completed using the blastp algorithm of the BLAST suite v2.29 [47] with default parameters and were filtered *post hoc* for E-value thresholds. BLAST results were filtered for the top hit by each protein sequence using the sortBlast function found in the master_lncRNA_pipeline.source file on GitHub.

Anchoring the Annotated Transcriptome to the Reference Transcriptome

To anchor the newly annotated transcriptome to the reference 'Pinot Noir' transcriptome, I developed a novel approach to reciprocal best hit (RBH) analyses. In short, this iterative method takes the result of a reciprocal BLAST search and identifies RBHs, which are then removed from the master file and saved. The reciprocal BLAST filter is then repeated on the transcripts remaining in the file, and the newly identified RBHs are then removed and saved. By default, this process is repeated 25 times with no inherent E-value threshold. An E-value threshold of $1\times10^{-5}$was applied *post hoc* after concatenating all reciprocal best hit files (.rbh) generated by the algorithm. A script for this analysis can be found on GitHub. Further, I employed MCScanX [69] to find homologs by a combination of collinear order and reciprocal best hits. MCScanX was executed following the guidelines of the manual. SQLite was used to determine the number of anchors that were identical in both the novel RBH and MCScanX algorithms.

Gene Duplication and Tandem Arrayed Genes:

Duplicate genes were identified using an all-by-all self-blastp using the total transcriptome for both the query and target. Using MCScanX, collinear duplicates were identified by providing the software both the output of the blastp and the Maker annotated .gff. MCScanX was executed using the following parameters: $path/to/MCScanX ./self_blast -e $i, where *i* represented various threshold E-values. The MCScanX tool duplicate_gene_classifier was used to classify genes into one category of several categories. By default, MCScanX returns both collinear duplications and tandem duplications in two separate files. Tandem duplications were further identified from the .tandem output file

Clustering and Expression-Level Filtering Transcripts for lncRNA

The output from Trinity was prepared for lncRNA analysis by filtering redundant and low-level expressed transcripts. To filter out redundant transcripts, the cd-hit-est algorithm of CD-HIT v4.6 [70] was used with the following parameters: -i Trinity.fasta -n 5 –o clust_Trinity.fasta -c 0.90 -m 8000 -T 6. Filtering by expression was executed with RSEM v1.2.28 [71] implemented by the Trinity-provided script align_and_estimate_abundance.pl with the following flags: --seqType fq –transcripts clust_Trinity.fasta –SS_lib_type FR –est_method RSEM –ali_method bowtie –trinity_mode –prep_reference. Transcripts with expression levels below FPKM=1.50 were filtered from the data set and removed from further analysis on the assumption they were spuriously present in the data set.

Identification of Putative Protein-Coding Genes

The transcripts were searched for open reading frames (ORFs) by Transdecoder v2.0.1 using the methods described by Haas [72]. Transcripts and ORFs identified by Transdecoder then were subjected to a series of analyses including blastx and blastp of the BLAST+ suite. Trinity-assembled sequences were searched against the Uniprot-Swiss-Prot and the Uniprot-Uniref90 databases to identify putatively encoded proteins at a threshold E-value of 1×10^{-20} using the blastx algorithm. Transdecoder-derived protein sequences were searched against the same databases using the blastp algorithm with the same settings. HMMER v2.3.2 searches were executed against the Pfam26.0 database using the Transdecoder-derived protein sequences. The best BLAST hit for each transcript (as determined by bit score, E-value, and percent identity, related yet separate similarity metrics) and all HMMER hits were loaded into Trinotate v2.0.2 to generate an annotation report. The output report was then mined in SQLite to sort transcripts into three bins according to BLAST hits (HMMER hits were ignored). The three bins contained transcripts with (1) significant homology to Viridiplantae proteins, (2) significant homology to non-Viridiplantae proteins, and (3) no homology to any protein sequence. RNA sequences of the latter category were considered putatively non-coding transcripts and were further analyzed in the pipeline. The script for this sorting process can be found on GitHub.

The Final Set of Non-Coding Transcripts and Their Validation

To minimize false discovery of non-coding transcripts, putative non-coding RNA sequences identified in the two independently generated RNA-seq data sets (of 588673 and Ventosa) were compared to one-another. Only sequences that were present as putative non-coding RNA sequences in both data sets with homology over at least 200 nucleotides (as determined by blastn) were carried forward. This was accomplished by filtering the 'Alignment Length' column in output format 6 using the programming language AWK. Of the sequences that were found homologous to one-anther in the two data sets, the longest one (as determined by Bioawk) was chosen to represent that transcript. These transcripts then were filtered against the RFAM v12.0 [73] database by the cmscan algorithm implemented by Infernal v1.1 [74]. The Infernal output was executed to write results to a tab-delimited file. Any hit that Infernal considered significant by default parameters was filtered out. Putatively non-coding transcripts then were validated by CPC [46] to label transcripts as either coding or non-coding based on blastx homologies to the Uniprot-Swiss-Prot database.

Free Energy Levels of Non-Coding Transcripts

The minimum free energy of each transcript was calculated using the rnafold algorithm implemented by the ViennaRNA-2.2.5 [75] software package using the following options to define the partition calculation algorithm, to ensure that 'dangling ends' are treated with the same energy requirements as paired bases, and to disallow lonely pairs: -p –d2 --noLP. The provided script, free_energy_calculations_v2.sh, took a list of sequences, retrieved the sequence using Samtools v 0.1.19 [76], and calculated the free energy using this method. Results were written to a tab delimited file whereby the name of the sequence, the minimum free energy (MFE), centroid free energy, and ensemble diversity were reported. The minimum free energies of the transcripts then were compared to the minimum free energy of similarly-sized randomly selected set of putative protein coding genes as annotated by Maker.

RESULTS

Genome Annotation and Transcriptome Profiling

The focus of this thesis research program was to construct a new and improved transcriptome of the cultivated grapevine *V. vinifera.* The transcriptome was assembled from RNA-seq reads derived from accessions 588673 and Ventosa of the cultivar 'Riesling'. Quality filtering and trimming of the raw RNA-seq data resulted in 14,190,809 paired-end reads for accession 588673. Following quality control, reads were paired. Pairing reads in accession 588673 generated 6,679,255 reads with paired sequences on both DNA strands, 514,591 unpaired sequences on the forward, and 317,708 unpaired sequences on the reverse DNA strand. Of the paired reads, 91.45% aligned to the 12Xv2 *V. vinifera* reference genome sequence derived from the variety ‘Pinot noir’ and hosted at CRIBI [77]. Of the unpaired reads, 79.45% aligned to the reference genome which resulted in a total alignment rate of 79.74%. Both paired and unpaired reads were *de novo* assembled into transcripts, which generated 62,745 contigs with an average contig length of 859 nucleotides and a median contig length of 551 nucleotides. The contig N50 (a weighted median) for the assembly was 1,325 nucleotides. The 62,745 contigs assembled were represented by 49,330 clusters (a term used by Trinity to designate separate genes). In accession Ventosa, quality control and trimming resulted in 103,677,027 reads. Pairing these sequences resulted in 48,639,916 reads paired on both DNA strands, 4,393,048 reads unpaired on the forward, and 2,004,147 reads unpaired on the reverse DNA strand. Of the paired reads, 91.14% aligned to the reference genome. Of the unpaired reads, 77.06% aligned to the reference genome leading to a total alignment rate of 80.64%. Assembly of these reads generated 157,779 contigs with an average length of 840 nucleotides, median length of 373 nucleotides, and N50 of 1,434 nucleotides. These 157,779 contigs were grouped into 109,215 clusters. Additional statistics of the Trinity assemblies are presented in Table 1.

Table 1. Various metrics assessing the quality of the transcript assembly by Trinity. Note: ExN50 represents a recalculated weighted median at the xth expression quantile.

Metric	'Riesling' Accession	
	Ventosa	588673
N10	3337	2953
N20	2599	2301
N30	2130	1912
N40	1762	1603
N50	1434	1325
E10N50	1413	1008
E20N50	1349	1119
E30N50	1290	1149
E40N50	1413	1259
E50N50	1520	1345
E60N50	1671	1480
E70N50	1815	1612
E80N50	1947	1620
E90N50	1922	1499
E100N50	1437	1328

All transcripts from both ‘Riesling’ accessions were used for a complete genome annotation of the *V. vinifera* cv. ‘Pinot Noir’ v2 reference genome. Using these transcripts and the entire Uniprot-Swiss-Prot reference protein database, 65,342 putative transcripts were identified. All transcripts had some level of support from either RNA-seq or protein alignments as indicated by the calculated Average Edit Distance (AED, a metric defining how similar the predicted gene is to RNA-seq evidence) value of less than 1.00. Annotated transcripts identified were functionally tied to proteins in the Uniprot-Swiss-Prot reference database [78] using the blastp algorithm. This procedure identified 1,680 transcripts with homology to proteins in the database, 1,004 of which were carried forward from the database itself in the annotation.

***Ab initio* Prediction, Protein Domain Characterization, and Functional Annotation**

Using the resultant combined accession output, annotations were used to train *ab initio* SNAP gene prediction models [20]. The raw output from the Maker alignment was used to train the first pass of the SNAP model, and this output was used to train the second pass. The resultant model identified 25,995 predicted genes as compared to 65,342 predicted genes from model-independent annotations. Of these, nearly all were represented by some level of RNA-seq or amino acid sequence alignments as noted by an AED of less than 1.00. A diagrammatic representation of this analysis pipeline, from transcript assembly to SNAP gene prediction models, can be found in Figure 2.

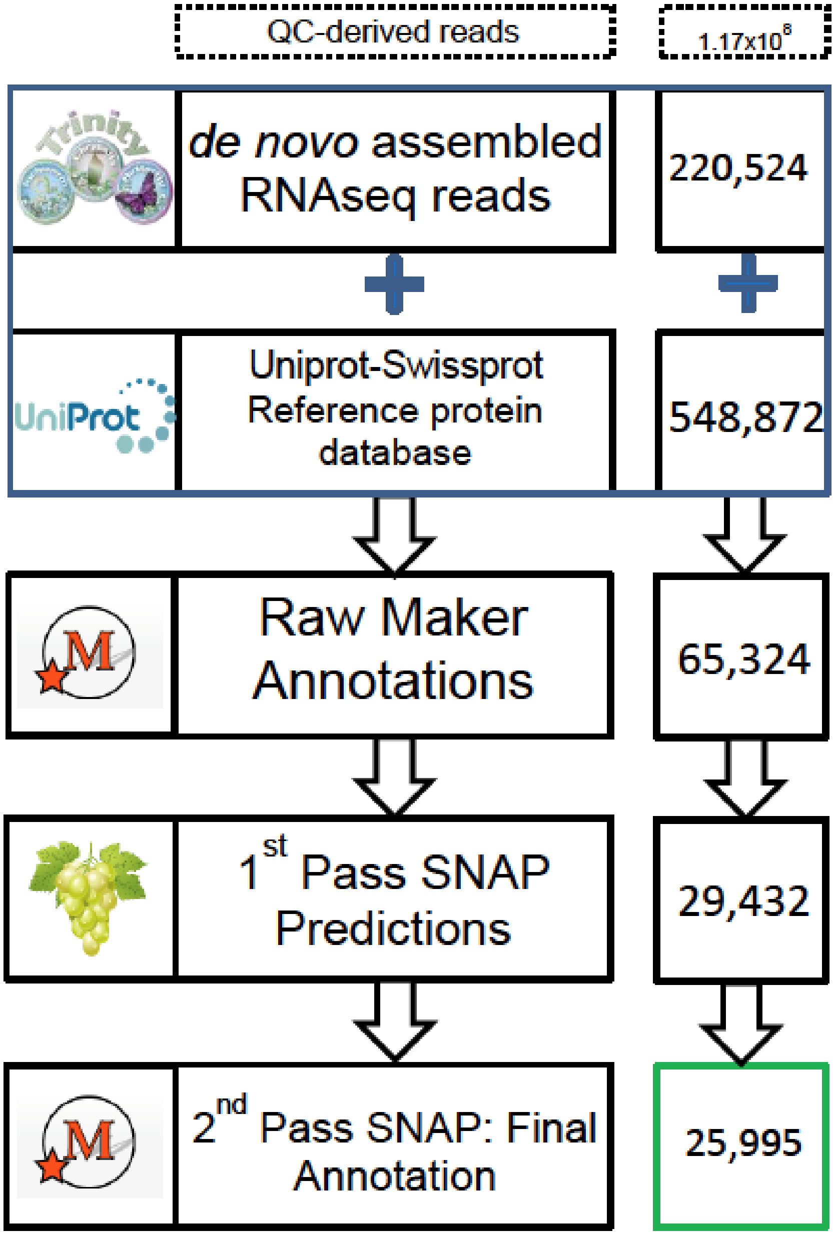
QC-derived reads
$1.17x10^{8}$
de novo assembled RNAseq reads
220,524
Uniprot-Swissprot Reference protein database
548,872
Raw Maker Annotations
65,324
1st Pass SNAP Predictions
29,432
2nd Pass SNAP: Final Annotation
25,995

Figure 2. Computational pipeline for transcriptome profiling as implemented by software Maker and SNAP. Values in the flowchart on the right show the number of transcripts carried forward at each step of the pipeline.

Protein domains were identified against the reference protein database Pfam30.0 [52]. Domains that were found to significantly match an entry in the database (E ≤

1×10^{-5}) were mapped to Gene Ontology (GO) [67] terms, classes, and functions. In total, 23,042 protein domains were found across all 25,995 annotated transcripts representing 14,763 uniquely annotated protein sequences. Collectively, 30,113 domains, representative of 10,794 unique proteins, were mapped to GO terms (accounting for domains that map to multiple GO terms). In total, 32.6% mapped to biological processes, 58.3% mapped to molecular function, and 9.1% mapped to cellular component. Summary data are shown in Figure 3. Proteins were further functionally annotated using blastp with the Uniprot-Swiss-Prot database as a reference. Using this approach, I anchored 20,382 and 16,863 proteins at $E \leq 1\times10^{-5}$, and at $E \leq 1\times10^{-20}$, respectively. Using the larger Uniprot-Uniref90 reference protein database [79], I anchored 21,577 and 19,835 proteins at $E \leq 1\times10^{-5}$and at $E \leq 1\times10^{-20}$, respectively. These outputs are summarized in Figure 4.

In addition to the transcripts that had either RNA-seq evidence or Uniprot-Swiss-Prot evidence, 66,820 transcripts lacking any form of alignment evidence also were identified. These transcripts were also searched for protein domains using the approaches described above. This analysis identified 5,790 domains in 3,709 unique transcripts. Of these, 4,151 mapped to GO terms, classes, and functions, accounting for domains that map to multiple GO terms. In total, 24.8% domains mapped to biological processes, 66.8% mapped to molecular function, and 8.3% mapped to cellular component. Further attempts to characterize these transcripts based on the UniProt-Swiss-Prot database resulted in the putative annotation of 4,383 and 1,686 proteins at E ≤ 1e-05 and at E ≤ 1e-20, respectively.

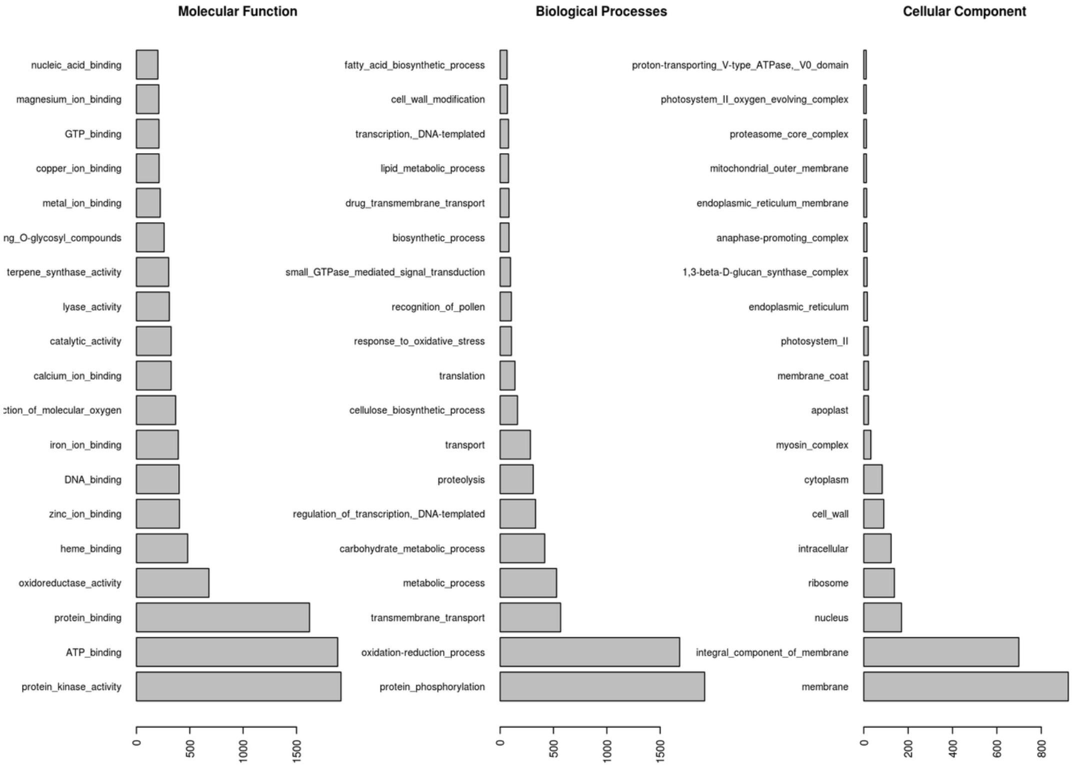

Figure 3. Gene Ontology terms most enriched in the final annotation separated into the three Gene Ontology categories: Molecular functions, biological processes, and cellular components.

Anchoring Annotated Transcripts to Reference Transcriptomes

To anchor newly annotated transcripts to the *V. vinifera* cv. 'Pinot Noir' reference transcriptome, a novel approach to reciprocal best hit (RBH) analysis was designed. This new RBH method is based on the following two assumptions: (1) each gene was present in the same number of copies in both 'Riesling' and 'Pinot Noir' and (2) putative anchors were sufficient. This analysis sorted the outcome of forward and backward blastp alignments in such a way that each transcript only matched each unique target one time. The following processes then occurred iteratively: (1) BLAST results were sorted such that only the highest scoring hit for each query was kept, (2) RBHs were identified, and

3) anything labeled as a RBH (either target or query) was removed from the master file and saved in a separate file at every incidence in the BLAST output. This process was executed 25 times, whereby 14,584 annotated transcripts were putatively anchored to reference transcripts with a threshold expected value of 1×10^{-5}. Transcripts were also tied to the v1 reference transcriptome, whereby 14,212 transcripts were anchored.

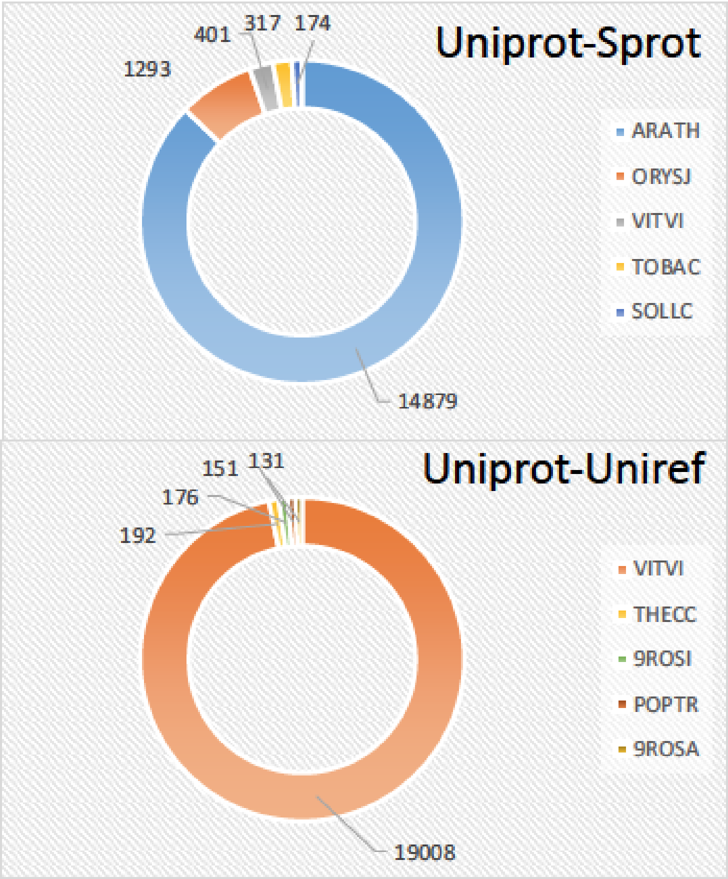

Figure 4. Functional annotation of the transcriptome against both the Uniprot-Swiss-Prot and the Uniprot-Uniref90 reference protein databases. ARATH: *Arabidopsis thaliana*, ORYSJ: *Oryza sativa* sp. *japanicum*, VITVI: *Vitis vinifera,* TOBAC: *Nicotiana tobacum,* SOLLC: *Solanum lycopersicum,* THECC: *Theobroma cacao,* 9ROSI: Rosids, POPTR: *Poplulus trichocarpa,* 9ROSA: Rosales.

Furthermore, a collinearity analysis was performed to anchor annotated and reference transcripts based on collinear order and reciprocal best hit results. This lead to the identification of 12,755 putative homologs. Of these, 6,018 also were present among those identified by the iterative RBH analysis. In addition to anchoring transcripts, collinear segments also were mapped across the genomes of the reference and our newly developed annotation. The differential placement of the chromosomal segements in our new annotation relative to the reference annotation is shown in Figure 5.

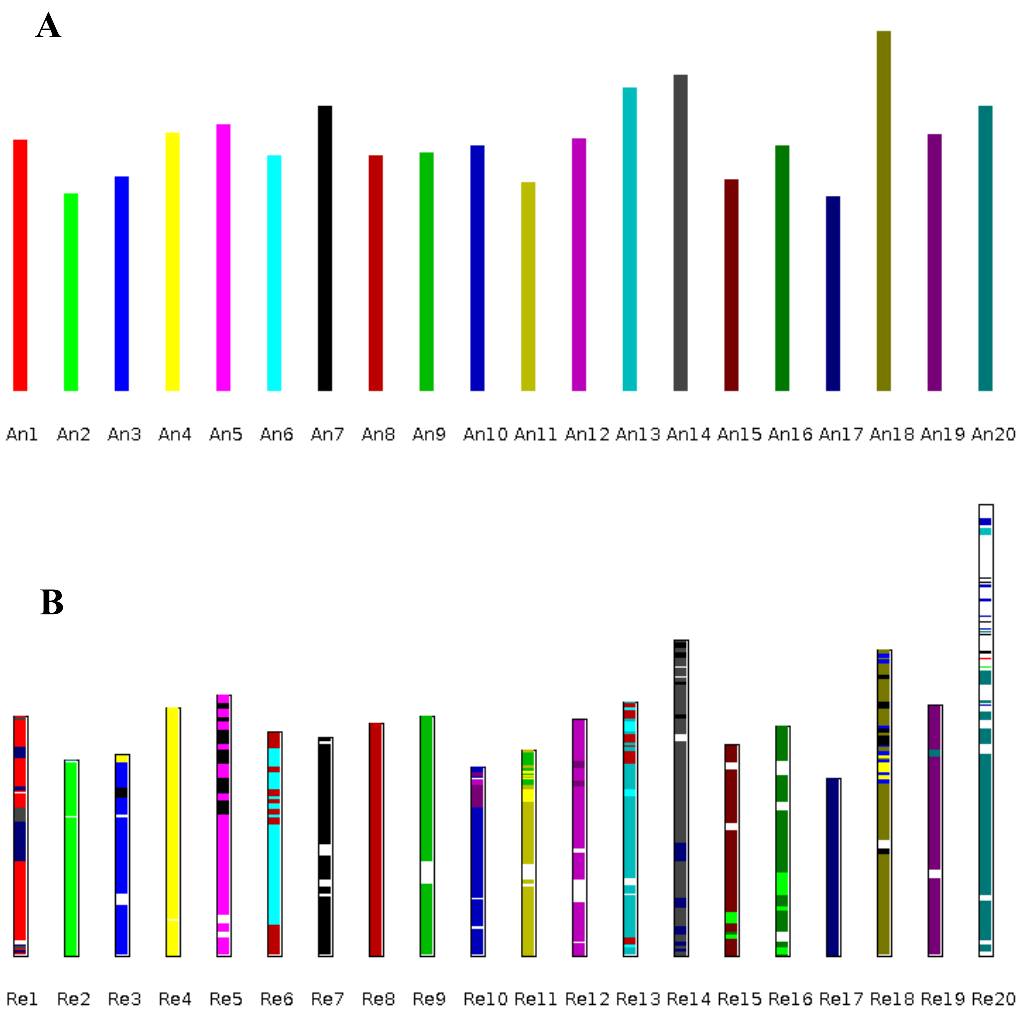

Figure 5. Schematic representation of the grapevine chromosome maps according to the (A) 'Riesling' annotation and (B) the reference annotation. Colored segments indicate the placement of the Riesling annotation-based chromosomes in the map of the reference genome. White segments indicate segments that were missing from the annotated chromosome maps based on the reference genome sequence.

Gene Duplication and Tandem Arrayed Genes

Gene duplication events were also detected using the self-BLAST-based function of the collinearity software (MCScanX, see Methods). The data from this analysis demonstrated that 14,448 genes were likely the results of whole genome duplication, 8,351 were considered dispersed duplications, 1,646 were considered proximal duplications, 1,482 genes represented tandem duplications, and 28 genes were considered singletons. Genes and their classification can be found in the file self_blast.gene_type provided in the Data directory on GitHub: https://github.com/znh1992/Thesis. At the more stringent threshold E-value of 1×10^{-20}, 10,648 duplication events were found across 7,633 genes. Pair counts and unique gene counts can be found in Table 2 for other E-values. These data are summarized in part in the inner track of the Circos plot in Figure 9. Because genes could fall into more than one of the above categories, a separate analysis was done to detect all putative tandem arrayed genes (TAGs). Regardless of threshold expected value, 2,104 tandem duplications were identified across 3,471 (13.3%) unique gene IDs. Of these, 737 (35%) had three tandemly arrayed genes leaving the rest in arrays of two genes.

Table 2. Gene duplication counts and unique gene counts in the duplications series across various E-values.

E-value	Duplications	Unique Genes
1×10^{-10}	30324	14219
1×10^{-15}	18954	10858
1×10^{-20}	10648	7633
1×10^{-50}	1542	1997
1×10^{-100}	296	549

Quality Assessment of the New Annotations

To assess the quality of the new transcriptome annotation, the transcriptome was searched for the presence of Benchmarking Universal Single Copy Orthologs (BUSCOs), a core set of genes occurring in a single copy in all genomes characterized in a particular clade. The reference set of single-copy orthologs for Emryophyta (land plants) currently consists of 1,440 genes. The transcriptome was probed for the number of found BUSCOs represented as both a correct (single copy) match, and a partially correct (duplicated) match. The final transcriptome annotation found 16.8% of expected BUSCOs. This call rate prompted an analysis across all steps of the pipeline presented above. Results for this analysis can be found in Figure 6.

Design of a *de novo* Pipeline for lncRNA Identification

A computational pipeline was constructed to glean lncRNAs from plant transcriptomes. To make the pipeline broadly applicable, it was designed to identify lncRNAs from raw RNA-seq reads in a reference genome-independent manner. The essential function of the pipeline was to remove protein-coding transcripts and short non-coding RNA sequences. First, transcripts were assembled from raw RNA-seq reads, and the resultant transcripts then were purged of redundancy. Clustered transcripts then were filtered by expression level, and the remaining set was further filtered to remove known protein coding genes identified by BLAST analyses. The remaining transcripts were compared across independent RNA-seq data sets from the same genome to ensure a low false positive identification rate [45]. These final transcripts were searched against the reference RNA database Rfam [73] to remove known ncRNAs. The remaining set of putative lncRNAs transcripts then were validated by using a semi-independent lncRNA identification method [46]. A diagram of the pipeline is shown in Figure 7.

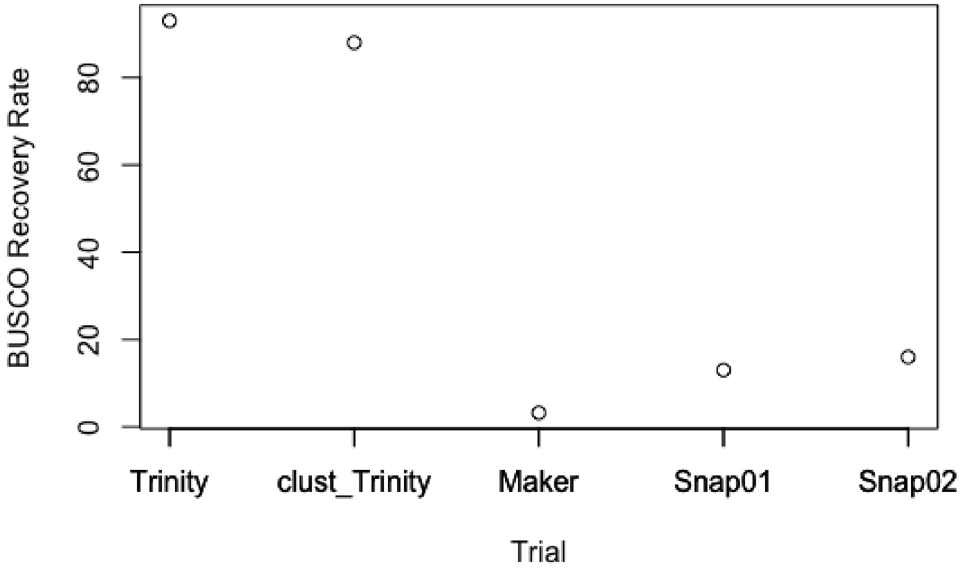

Figure 6. BUSCO scores across the main steps of the transcriptome profiling pipeline. Clustered Trinity transcripts are also shown to demonstrate their presence.

Clustering and Expression-Level Filtering Transcripts for lncRNA

Due to the highly redundant nature of *de novo* assembled transcriptome builds, the Trinity output was clustered for both 'Riesling' accessions using the cd-hit-est algorithm. Here, contigs were clustered based on the default standard of 90% similarity threshold and a default word size of five nucleotides. Clustering in accession 588673 resulted in 48,769 contigs. The N50 value for this transcript set was 1,284 nucleotides with an average contig length of 817 nucleotides. In the accession Ventosa, cd-hit-est generated 110,900 contigs, the N50 value was 1,395 nucleotides with an average contig length of 806 nucleotides.

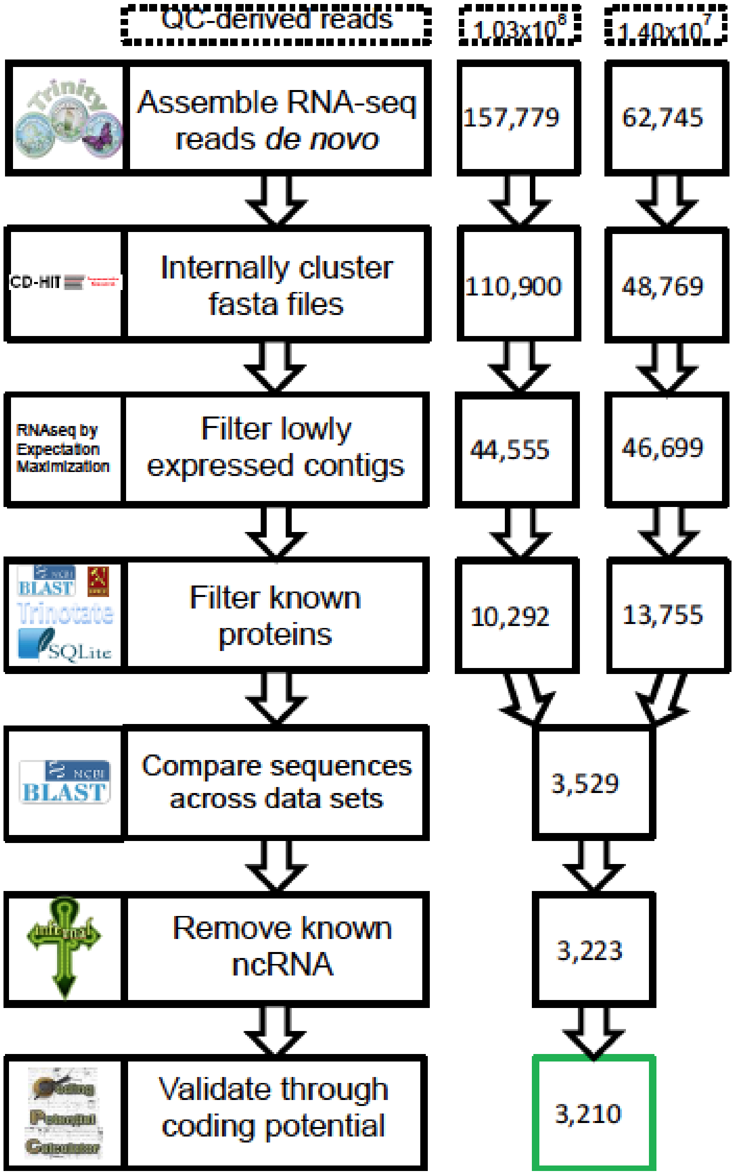
QC-derived reads
1.03x10^8
1.40x10^7
Trinity
Assemble RNA-seq reads *de novo*
157,779
62,745
CD-HIT
Internally cluster fasta files
110,900
48,769
RNAseq by Expectation Maximization
Filter lowly expressed contigs
44,555
46,699
NCBI BLAST
Trinotate
SQLite
Filter known proteins
10,292
13,755
NCBI BLAST
Compare sequences across data sets
3,529
infernal
Remove known ncRNA
3,223
Coding Potential Calculator
Validate through coding potential
3,210

Figure 7. Computational pipeline for the *de novo* identification of lncRNAs. Values in the right flowchart show the number of transcripts carried forward at each step along the pipeline.

To further reduce the complexity of the data sets, clustered transcript sets were filtered by an expression level threshold of FPKM ≥ 1.50. Accession 588673 resulted in 46,699 contigs with an N50 contig length of 1,270 nucleotides and with an average contig length of 780 nucleotides. Accession Ventosa resulted in 31,103 contigs with an N50 value of 1,925 nucleotides and an average contig length of 1,380 nucleotides.

Removal of Putative Protein-Coding Genes

To identify and remove all transcripts putatively annotated as protein-coding from the clustered and expression-level filtered transcript sets, I, *in silico*, translated all transcripts into predicted polypeptides, and searched the resulting dataset against the reference protein databases UniProt-Swiss-Prot and Uniprot-Uniref90 with an E-value threshold of 1×10^{-20}. Only the top BLAST hit for each sequence in the accessions was accepted based on bit score, E-value, and percent identity. Contigs from each accession were binned into the categories of Viridiplantae proteins, non-Viridiplantae proteins, and proteins for which no homologous hit was found. Results for this analysis can be found in Table 3. Due to repetitive entries in the SQLite database probed for these categories, the sequences that lacked annotations were clustered by name resulting in 13,755 and 10,292 unannotated sequences in accessions 588673 and Ventosa, respectively.

Table 3. Classification of protein by best BLAST hit.

Classification	**Ventosa**	**588673**
Viridiplantae	34321	41686
Non-Viridiplantae	682	3340
No Classification	10292	13755

The Final Set of Non-Coding Transcripts and Their Validation

To identify RNA sequences that occurred in both the 588673 and Ventosa accessions, the transcripts from both were searched against one-another using the blastn algorithm. Only the Ventosa transcripts corresponding to the most closely related (based on bit score, E-value, and percent identity) 588673 transcript with an alignment length of at least 200 nucleotides were carried forward in the analysis. In matches of at least 200 nucleotides, the longest transcript from either 588673 or Ventosa was taken. This resulted in 3,529 sequences.

These 3,529 putatively identified non-coding RNAs then were filtered for the presence of known non-coding RNAs housed in the RFAM v12.0 reference database. Only 196 transcripts matched sequences in the data base at an E-value threshold of 0.01. Upon removal of these known ncRNA sequences, the dataset consisted of 3,223 putative long non-coding RNAs.

To validate the set of putative long non-coding RNAs identified by the pipeline, I used the software Coding Potential Calculator (CPC). CPC served as a pseudo-independent tool for validation because, while its results are based on blastx alignments, the parameters are much less stringent than those used above. Using this tool against the UniProt-Swiss-Prot database, 3,210 sequences were predicted to be non-coding, supporting the classification of 99.60% of the predictions. CPC was also executed against the Uniprot-Uniref90 reference protein database where 3,155 transcripts were predicted to be non-coding, supporting 97.90% of my predictions.

Free Energy Levels of Protein-Coding and Non-Coding Transcripts

It has been hypothesized that the regulatory function of lncRNAs is inherently associated with the higher free energy of their secondary structure [80–82]. To examine the possibility that lncRNAs have a higher free-energy level than mRNAs, I used the RNA free energy calculator and folding algorithm RNAfold. RNAfold was used to calculate the secondary structure and the minimum free energy of all putative lncRNAs and a randomly selected set of 3,225 annotated protein coding transcripts identified above. Sequences representing the highest and the lowest free energies can be found in panels A and C in Figure 8. All free energy values were corrected for the length of the sequence, as the length of the transcript is a key parameter in the minimum free energy of these structures. The corrected minimum free energy distribution of all non-coding RNAs are shown in panel B of Figure 8. The mean length-corrected minimum free energy for annotated protein coding genes was -0.264 kcal/mol/nt with a standard deviation of 0.042 kcal/mol/nt. The mean length-corrected minimum free energy content of the putatively annotated lncRNAs was -0.258 kcal/mol/nt with a standard deviation of 0.030 kcal/mol/nt. These two data sets were found to be significantly different using a two-tailed Welch's t-test (p <<< 0.05). Long non-coding RNAs then were aligned to the reference genome using the map2assembly tool in Maker whereby 3,049 were mapped. Using genome-aligned lncRNAs and the genomic coordinates of gene structures from Maker and SNAP, gene frequencies were mapped across the reference genome on all linkage groups in 1Mbp bins. Figure 9 demonstrates these frequency distributions in the inner and middle tracks. In addition, average raw uncorrected free energy values of protein-coding and non-coding transcripts were plotted across all 1-Mbp bins and are shown in the outer track of Figure 9.

Long Non-Coding RNA Analysis Across the *Vitis* Genus

In addition to two accessions of *V. vinifera* cv. 'Riesling', two accessions of both *Vitis aestivalis* (1664 and 588626) and *Vitis rupestris* (588160 and 588174) were analyzed using the *de novo* lncRNA identification pipeline. Summary data for these analyses can be found in Table 4. In short, 5,743 putative lncRNA transcripts were identified in *V. aestivalis,* of which 5,718 were validated by use of CPC. In *V. rupestris*, 6,013 transcripts were identified, of which 5,993 were validated with CPC.

Table 4. lncRNA discovery in wild grapevine accessions of *V. aestivalis* and *V. rupestris.*

Pipeline Step	Number of Transcripts Retained			
	Aestivalis.1664	Aestivalis.588626	Rupestris.588160	Rupestris.588574
Trinity	95261	81993	88801	97275
CD-HIT	76028	67073	72810	79223
RSEM	59772	61093	62386	55060
Trinotate	20236	21241	23211	21396
BLAST	6158		6322	
Infernal	5748		6013	
CPC	5718		5993	

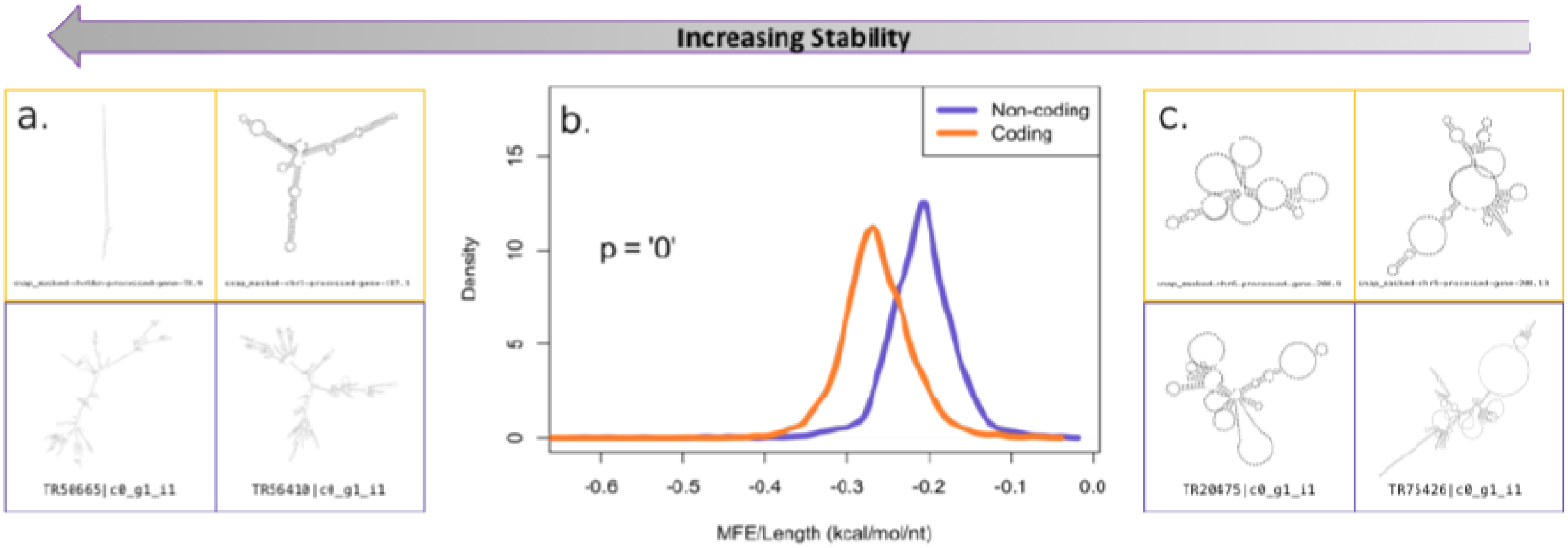

Figure 8. Free energy figures of protein coding genes and lncRNAs. (a). Representative stable structures for both protein-coding genes (orange) and lncRNAs (purple). (b). Length-corrected free energy distributions. (c). Representative unstable structures for both protein-coding genes (orange) and lncRNAs (purple).

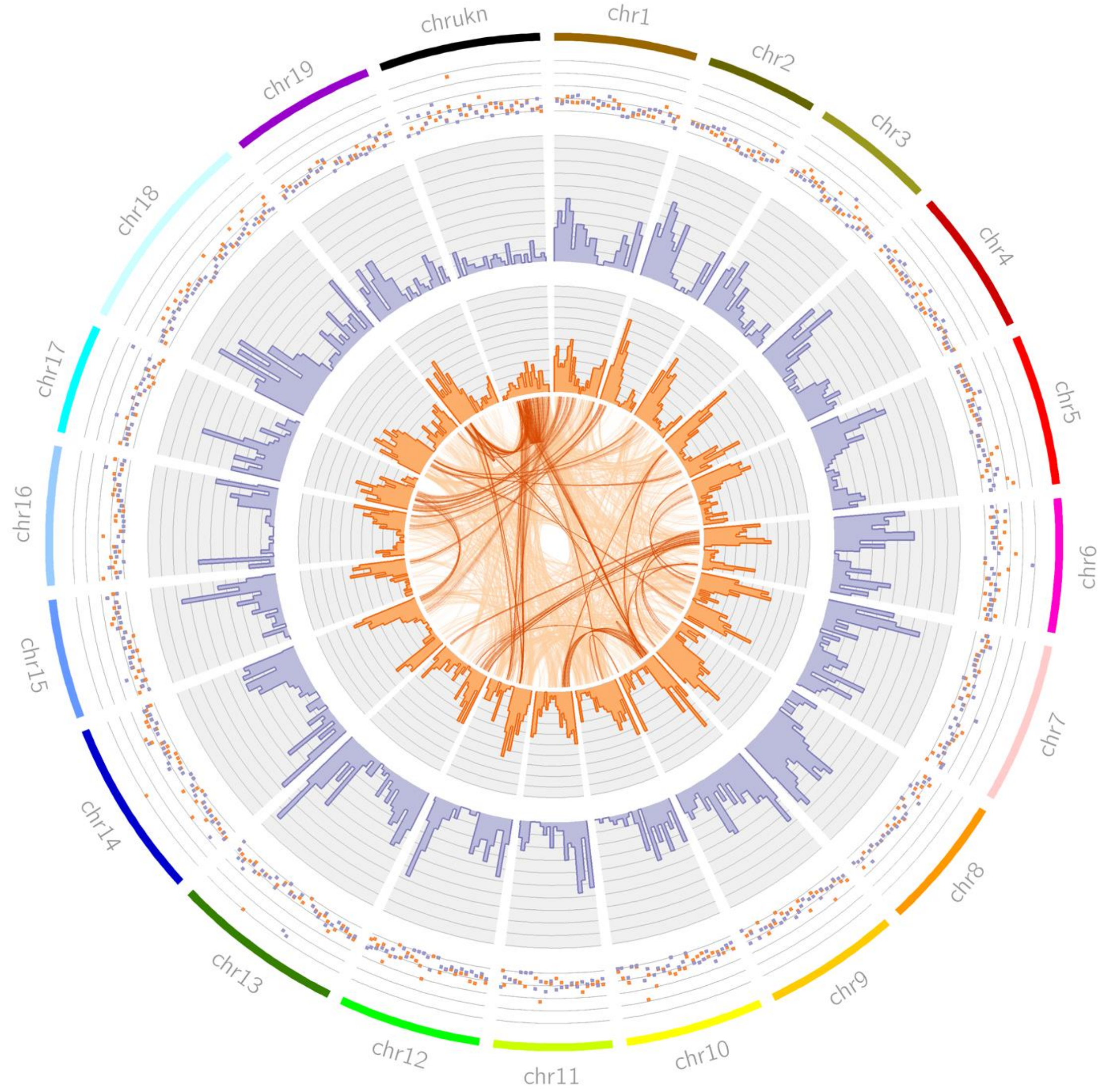

Figure 9. Gene duplication, gene frequency distribution, and average raw free energy of transcripts in grapevine. The center track of the Circos plot [83] shows evidence of gene duplication at two significance thresholds: E-value = 1×10^{-5} (light orange) and E-value = 1×10^{-20} (dark orange). The orange histogram shows gene frequency distributions across all linkage groups on a scale of 0 to 30 genes for 1Mbp bins. The purple histogram shows lncRNA frequency distributions across all linkage groups on a scale from 0 to 30 transcripts. The scatter plot shows average raw free energy for each 1 Mbp bin for both protein coding genes (orange) and lncRNAs (purple) ranging from 25 to 750 –kcal/mol.

DISCUSSION

Biological research in the post-genomic era requires massive data sets that can only be generated and compiled by the collective effort of communities of scientists who share a species or taxon of common interest. For structural and functional genomics, the grape research community has relied on the *V. vinifera* reference genome sequence constructed from the DNA of a 'Pinot Noir'-derived grapevine [59]. Since the public release of this reference genome sequence in 2007, most grapevine transcripts and transcriptomes have been based on its gene prediction models. Genome-guided transcriptomes, however, are limited in that non-predicted transcripts and splice variants in the genome sequence will remain undetected or erroneously assembled. Furthermore, errors in sequence assembly and gene model predictions in the refence genome are carried over to the transcript information and can lead to assembly artifacts. The high-level heterozygosity in grapevine further complicates transcript assembly by potentially misidentifying allelic variants as paralogous transcripts. This problem can now be remedied with the application of genome-independent assembly of RNA-seq reads into consensus transcripts.

In this thesis, I present a reference transcriptome for *Vitis vinifera* cv. 'Riesling' using a coupled *de novo* assembly and genome-guided prediction algorithm. This approach takes into consideration the high-level heterozygosity of grapevine by use of *de novo* methods and takes advantage of the 'Pinot Noir' genome sequence without incorporating its potentially problematic gene prediction models. The Maker-generated transcriptome I present in this work offers a more complete picture of the grapevine transcriptional landscape and will facilitate the study of transcript-level variation among grapevine cultivars and species.

In development of the transcriptome, I attempted to anchor the new transcripts to reference transcripts previously identified in 'Pinot Noir' data [10]. Anchoring was performed using two methods: MCScanX, a program for collinearity detection, and a newly developed reciprocal best hit (RBH) analysis. In both cases, I could anchor a relatively low percentage of transcripts (49% and 55%, respectively). In the case of MCScanX, the low percentage of anchoring may be attributed to annotations on different versions of the genome. Between the publication of the reference transcriptome and our 'Riesling' transcriptome, an updated reference sequence was released to reflect new data obtained by more advanced sequencing methods. On the computational level, anchoring based on an updated version of the reference genome could produce results that appear as chromosomal rearrangements. Most notably, segments from an unplaced linkage group (chrUkn) are sorted into their likely locations in the genome map. Figure 5 demonstrates apparent chromosomal rearrangement even beyond the placement of ChrUkn segments, suggesting that collinearity-based approaches may be insufficient in homolog identification.

These data mirror the findings of Venturini *et al.* (2013) who characterized a transcriptome for *V. vinifera* cv. 'Corvina', and were able to recover only 51% of the v1 reference transcriptome's 29,971 transcripts [61]. These data are further supported by results of a recent analysis by Chin *et al.* (2016) who assembled the genome of *V. vinifera* cv. 'Cabernet Sauvignon' and were able to align only 16,981 (57%) of the reference transcriptome to their new genome [62]. I attempted to anchor the new annotation to both the v1 and v2.1 reference transcriptome using a RBH-driven approach, and I was able to anchor only 47% and 46% of the transcripts, respectively. Errors in the v2.1 reference transcriptome are easily attributable to methodology as Vitulo *et al.* (2014) combined RNA-seq data from *V. vinifera* cv. 'Cabernet Sauvignon', and two rootstocks derived from four separate species, (*V. vinifera, Vitis berlandieri, V. rupestris,* and *Vitis riparia),* making the assumption that interspecific variation was insignificant [10]. This assumption may, however, be incorrect, and the introduction of data from non-vinifera grapevines could have profound impacts on the quality of the transcriptome from the perspective of cultivar variation.

Several quantitative features of the Riesling transcriptome were similar to those of other plants. For example, of the genes predicted, 13.3% (3,471 genes) occurred in tandemly duplicated gene arrays, a percentage consistent with the percent of gene arrays in *Arabidopsis thaliana* (16.6%) [84] and even such distantly related organisms as humans, mice, and rats (14-17%) [85]. Further, nearly 65% of these arrays contain only two genes, a percentage consistent with *A. thaliana,* where 75% of its tandem arrays contains only two genes [84].

I also predictively annotated the transcriptome for function through computational means, namely using the software HMMER and BLAST. Both programs returned unexpectedly few predicted functional annotations (57% and 65% respectively). These numbers are similar to the results from the reference transcriptome, indicating that the inability to functionally annotate (via computational means) is a feature that is characteristic to such non-model species as grapevine. For example, a recent analysis of rose-scented geranium was able to identify GO terms in only 33% of assembled transcripts and found protein homology in reference databases in only 66% transcripts using blastx [86]. Over time, as more proteins are characterized and experimentally annotated for function, these numbers are expected to increase.

The inability to call many single-copy orthologs indicates that Maker may not have provided a complete annotation as implemented. This is particularly evidenced by the fact that the raw Trinity output had a nearly 94% call rate of single copy orthologs. With this, we would expect the final annotation to contain this many, if not more, through the use of machine learning. This approach seems to be relatively common, but many analyses go one or more steps further in the complete annotation of genomes. For example a recent analysis of strawberry trained multiple machine learning gene predictors and completed the analysis with a final alignment guided by these predictors [87]. Current analyses are accounting for these approaches to improve the current model. Further evidence for a poor annotation is presented in the form of Figure 5, where a high degree of chromosomal rearrangements is predicted. If this were the case, we would expect 'Riesling' to be unable to generate fertile progeny when crossed with other *V. vinifera* cultivars. However, *V. vinifera* cultivars are inter-fertile, indicating that the apparent chromosomal rearrangements depicted in Figure 5 most likely reflect an inability of the software to generate accurate predictions.

As our insight into the regulation of protein-coding genes improves, there is mounting evidence that long non-coding RNAs (lncRNAs) play an important role in those processes (reviewed in [57, 88, 89]). Identification of these transcripts is paramount for a true understanding of the role of these RNA species. I present a standardized computational pipeline for the identification of lncRNAs, that promises to be particularly useful in non-model species. This pipeline represents a logical sequence of processes for removing known protein-coding genes and other non-coding RNAs using the best methods available to date. The pipeline predicts lncRNAs and then attempts to validate them using a pseudo-independent software, CPC. I consider this validation pseudo-independent, because both the pipeline and CPC incorporate BLAST results, albeit to varying levels of confidence. These transcripts are, at best, predictions, and only experimental evidence will validate their true function. Nonetheless, our predicted lncRNAs were found distinctly different from the protein-coding annotations in several key parameters. For example, lncRNAs tended to be enriched for lower GC content and, surprisingly, longer transcripts. These two factors are both consistent with the idea that lncRNAs have differential stability as compared to protein-coding transcripts. This quantitation may be useful in machine learning prediction algorithms as research on lncRNAs moves forward.

While functional annotation and experimental validation of these transcripts are beyond the scope of this study, many aspects of this pipeline provide evidence for the existence of these transcripts. For example, filtering for moderate and high expression levels reduces the chance of finding ‘accidental’ transcripts. Further, the requirement for multiple discoveries reduces the risk of false positive lncRNA transcripts. Moreover, this pipeline works independent of coding potential calculations. Many pipelines use tools such as CPC to serve as a filtering mechanism, but lack any power to validate their results. This study provides both a pipeline for identification and evidence that the pipeline reliably gleans long non-coding transcripts. This validation demonstrates the pipeline’s efficacy in both cultivated and wild grapevine RNA-seq data. Further, results from CPC seem to indicate successful identification using both Uniprot-Swiss-Prot (a small, experimentally validated protein database) and Uniprot-Uniref90 (a larger, non-experimentally validated protein database). These results indicate that Uniprot-Swiss-Prot alone could be used for CPC validation, which greatly reduce the requirement for computational resources.

The results presented here, taken together with results of previous grape transcriptome assembly projects, suggest that the RNA-seq and predictive method-based genome annotation will be improved greatly by the availability of cultivar-specific genome sequences [61, 62, 90]. This is necessary for the development of cultivar-specific gene models and inter-cultivar analyses of variations. Furthermore, to more accurately pinpoint differences in the transcriptome of 'Riesling' and 'Pinot Noir', data from more cultivars need to be include in the analysis. Only then will we be able to speak of true varietal differences in grapevine at the transcriptional level. This will then need to be tied to functional annotation. Work on lncRNAs will need to expand in two directions. First, more efficient and accurate methods for their identification are crucial. While logical, this process is time consuming and computationally intensive. Better prediction methods will find the same or similar sets of lncRNA while being more efficient. Second, lncRNAs need to be validated in both presence and function. Both facets of this research will be critical in moving the viticulture industry forward in the face of a changing climate and changing agricultural practice.

REFERENCES

1. Jacob F, Monod J. On the Regulation of Gene Activity. Cold Spr Harb Symp Quant Biol. 1961;26:193–211.

2. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012;7:562–78.

3. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol. 2013;31:46–53.

4. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.

5. Trapnell C, Pachter L, Salzberg SL. TopHat: Discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25:1105–11.

6. Trapnell C, Williams B a, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010;28:511–5.

7. Eckalbar WL, Hutchins ED, Markov GJ, Allen AN, Corneveaux JJ, Lindblad-Toh K, et al. Genome reannotation of the lizard Anolis carolinensis based on 14 adult and embryonic deep transcriptomes. BMC Genomics. 2013;14:49.

8. Haas BJ, Wortman JR, Ronning CM, Hannick LI, Smith RK, Maiti R, et al. Complete reannotation of the Arabidopsis genome: methods, tools, protocols and the final release. BMC Biol. 2005;3:7.

9. Wang G, Wang L, Cui Y, Yu M, Dang C, Wang H, et al. RNA-seq analysis of Brachypodium distachyon responses to Barley stripe mosaic virus infection. Crop J. 2016.

10. Vitulo N, Forcato C, Carpinelli EC, Telatin A, Campagna D, D'Angelo M, et al. A deep survey of alternative splicing in grape reveals changes in the splicing machinery related to tissue, stress condition and genotype. BMC Plant Biol. 2014;14:99.

11. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 2013;8:1494–512

.

12. Xie Y, Wu G, Tang J, Luo R, Patterson J, Liu S, et al. SOAPdenovo-Trans: De novo

transcriptome assembly with short RNA-Seq reads. Bioinformatics. 2014;30:1660–6.

13. Chang Z, Li G, Liu J, Zhang Y, Ashby C, Liu D, et al. Bridger: a new framework for de novo transcriptome assembly using RNA-seq data. Genome Biol. 2015;16:30.

14. Schulz MH, Zerbino DR, Vingron M, Birney E. Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics. 2012;28:1086–92.

15. Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, et al. De novo assembly and analysis of RNA-seq data. Nat Methods. 2010;7:909–12.

16. Chopra R, Burow G, Farmer A, Mudge J, Simpson CE, Burow MD. Comparisons of de novo transcriptome assemblers in diploid and polyploid species using peanut (Arachis spp.) RNA-Seq data. PLoS One. 2014;9:e115055.

17. Rana SB, Zadlock FJ, Zhang Z, Murphy WR, Bentivegna CS. Comparison of De Novo Transcriptome Assemblers and k-mer Strategies Using the Killifish, Fundulus heteroclitus. PLoS One. 2016;11:e0153104.

18. Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, et al. MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 2008;18:188–96.

19. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics. 2011;12:491.

20. Korf I. Gene finding in novel genomes. BMC Bioinformatics. 2004;5:59.
21. Stanke M, Steinkamp R, Waack S, Morgenstern B. AUGUSTUS: A web server for gene finding in eukaryotes. Nucleic Acids Res. 2004;32 WEB SERVER ISS.:309–12.

22. Sanger F, Coulson AR, Friedmann T, Air GM, Barrell BG, Brown NL, et al. The nucleotide sequence of bacteriophage ΦX174. J Mol Biol. 1978;125:225–46.

23. Neale DB, Wegrzyn JL, Stevens KA, Zimin A V, Puiu D, Crepeau MW, et al. Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. Genome Biol. 2014;15:R59.

24. Smit A, Hubley RH, Green P. RepeatMasker. http://www.repeatmasker.org/.
25. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 2015;6:11.

26. Wheeler TJ, Clements J, Eddy SR, Hubley R, Jones TA, Jurka J, et al. Dfam: A database of repetitive DNA based on profile hidden Markov models. Nucleic Acids Res. 2013;41:70–82.

27. Aken BL, Achuthan P, Akanni W, Amode MR, Bernsdorff F, Bhai J, et al. Ensembl 2017. Nucleic Acids Res. 2016;45:D635–42.

28. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 2007;35 SUPPL. 1:501–4.

29. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: The reference human genome annotation for the ENCODE project. Genome Res. 2012;22:1760–74.

30. Lee RC. The C . elegans Heterochronic Gene lin-4 Encodes Small RNAs with Antisense Complementarity to & II-14. 1993;75:843–54.

31. Lagos-Quintana M, Rauhut R, Lendeckel W, Tuschl T. Identification of novel genes coding for small expressed RNAs. Science. 2001;294:853–8.

32. Agrawal N, Dasaradhi PVN, Mohmmed A, Malhotra P, Bhatnagar RK, Mukherjee SK. RNA interference: biology, mechanism, and applications. Microbiol Mol Biol Rev. 2003;67:657–85.

33. Hadjiolov A, Venkov P, Tsanev R. Ribonucleic Acids Fractionation by Density-Gradient Centrifugation and by AGar Gel Electrophoresis: A Comparison. 1965;1962:159–63.

34. Darzacq X, Jady BE, Verheggen C, Kiss AM, Bertrand E, Kiss T. Cajal body-specific small nuclear RNAs: A novel class of 2'-O-methylation and pseudouridylation guide RNAs. EMBO J. 2002;21:2746–56.

35. Siomi MC, Sato K, Pezic D, Aravin A a. PIWI-interacting small RNAs: the vanguard of genome defence. Nat Rev Mol Cell Biol. 2011;12:246–58.

36. Kapranov P, Cheng J, Dike S, Nix D, Duttagupta R, Willingham A, et al. RNA Maps Reveal New RNA Classes and a Possible Function for Pervasive Transcription. Science (80-). 2007;316:1484–8.

37. Auyeung VC, Ulitsky I, McGeary SE, Bartel DP. Beyond secondary structure: Primary-sequence determinants license Pri-miRNA hairpins for processing. Cell. 2013;152:844–58.

38. Lee Y, Ahn C, Han J, Choi H, Kim J, Yim J, et al. The nuclear RNase III Drosha initiates microRNA processing. Nature. 2003;425:415–9.

39. Cai X, Hagedorn CH, Cullen BR. Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. RNA. 2004;10:1957–66.

40. Hamilton AJ, Baulcombe DC. A Species of Small Antisense RNA in Posttranscriptional Gene Silencing in Plants. 1999;213:1997–2000.

41. Jia H, Osak M, Bogu GK, Stanton LW, Johnson R, Lipovich L. Genome-wide computational identification and manual annotation of human long noncoding RNA genes. RNA. 2010;16:1478–87.

42. Paytuví Gallart A, Hermoso Pulido A, Anzar Martínez de Lagrán I, Sanseverino W, Aiese Cigliano R. GREENC: a Wiki-based database of plant lncRNAs. Nucleic Acids Res. 2015;:1–6.

43. Musacchia F, Basu S, Petrosino G, Salvemini M, Sanges R. Annocript: A flexible pipeline for the annotation of transcriptomes able to identify putative long noncoding RNAs. Bioinformatics. 2015;31:2199–201.

44. Legeai F, Derrien T. Identification of long non-coding RNAs in insects genomes. Curr Opin Insect Sci. 2015;7:37–44.

45. Al-Tobasei R, Paneru B, Salem M. Genome-Wide Discovery of Long Non-Coding RNAs in Rainbow Trout. PLoS One. 2016;11:e0148940.

46. Kong L, Zhang Y, Ye ZQ, Liu XQ, Zhao SQ, Wei L, et al. CPC: Assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Res. 2007;35 SUPPL.2:345–9.

47. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST plus : architecture and applications. BMC Bioinformatics. 2009;10.

48. Lin MF, Jungreis I, Kellis M. PhyloCSF: A comparative genomics method to distinguish protein coding and non-coding regions. Bioinformatics. 2011;27:275–82.

49. Li A, Zhang J, Zhou Z. PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme. BMC Bioinformatics. 2014;15:311.

50. Sun L, Luo H, Bu D, Zhao G, Yu K, Zhang C, et al. Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. Nucleic Acids Res. 2013;41:1–8.

51. Eddy SR. A new generation of homology search tools based on probabilistic inference. Genome Inform. 2009;23:205–11.

52. Bateman A. The Pfam protein families database. Nucleic Acids Res. 2004;32:138D–141.

53. Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics. 2015;31:926–32.

54. Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature. 2009;458:223–7.

55. Cabili M, Trapnell C, Goff L, Kaziol M, Tazon-Vega B, Regev A, et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. 2012;447:1915–27.

56. Laurent G St., Vyatkin Y, Antonets D, Ri M, Qi Y, Saik O, et al. Functional annotation of the vlinc class of non-coding RNAs using systems biology approach. Nucleic Acids Res. 2016;:1–20.

57. Heo JB, Lee Y-S. Molecular functions of long noncoding transcripts in plants. J Plant Biol. 2015;58:361–5.

58. Franco-Zorrilla JM, Valli A, Todesco M, Mateos I, Puga MI, Rubio-Somoza I, et al. Target mimicry provides a new mechanism for regulation of microRNA activity. Nat Genet. 2007;39:1033–7.

59. Jaillon O, Aury J-M, Noel B, Policriti A, Clepet C, Casagrande A, et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature. 2007;449:463–7.

60. Forcato C. Gene prediction and functional annotation in the Vitis vinifera genome. 2010.

61. Venturini L, Ferrarini A, Zenoni S, Tornielli GB, Fasoli M, Dal Santo S, et al. De novo transcriptome characterization of Vitis vinifera cv. Corvina unveils varietal diversity. BMC Genomics. 2013;14:41.

62. Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, et al. Phased Diploid Genome Assembly with Single Molecule Real-Time Sequencing. Nat Methods. 2016;13:56887.

63. Vilsak T. 2012 Census of Agriculture. 2013.

64. World Wine Production by Country. 2016. http://www.wineinstitute.org/files/World_Wine_Production_by_Country_2015.pdf. Accessed 1 Jan 2017.

65. Pairfq. https://github.com/sestaton/Pairfq.

66. MAKER Tutorial for GMOD Online Training 2014.
http://weatherby.genetics.utah.edu/MAKER/wiki/index.php/MAKER_Tutorial_for_GMOD_Online_Training_2014. Accessed 8 Feb 2017.

67. Gene T, Consortium O. The Gene Ontology project in 2008. Nucleic Acids Res. 2008;36 Database issue:D440-4.

68. Eddy SR. A probabilistic model of local sequence alignment that simplifies statistical significance estimation. PLoS Comput Biol. 2008;4:e1000069.

69. Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, et al. MCScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. 2012;40:1–14.

70. Li W, Godzik A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22:1658–9.

71. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;12:323.

72. TransDecoder (Find Coding Regions Within Transcripts). trinotate.github.io. Accessed 3 Apr 2017.

73. Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR, et al. Rfam 12.0: Updates to the RNA families database. Nucleic Acids Res. 2015;43:D130–7.

74. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics. 2013;29:2933–5.

75. Lorenz R, Bernhart SH, zu Siederdissen C, Tafer H, Flamm C, Stadler PF, et al. {ViennaRNA} Package 2.0. Algorithms Mol Biol. 2011;6:26.

76. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.

77. University of Padua. Unive. http://www.cribi.unipd.it/.

78. Uniprot-Swiss-Prot. http://www.uniprot.org/uniprot/. Accessed 8 Feb 2017.

79. Uniprot-Uniref90. http://www.uniprot.org/uniref/. Accessed 8 Feb 2017.

80. Mu C, Wang R, Li T, Li Y, Tian M, Jiao W, et al. Long Non-Coding RNAs (lncRNAs) of Sea Cucumber: Large-Scale Prediction, Expression Profiling, Non-Coding Network Construction, and lncRNA-microRNA-Gene Interaction Analysis of lncRNAs in Apostichopus japonicus and Holothuria glaberrima During LPS Challeng. Mar Biotechnol. 2016.

81. Mohammadin S, Edger PP, Pires JC, Schranz ME. Positionally-conserved but sequence-diverged: identification of long non-coding RNAs in the Brassicaceae and Cleomaceae. BMC Plant Biol. 2015;15:217.

82. Yang JR, Zhang J. Human long noncoding RNAs are substantially less folded than messenger RNAs. Mol Biol Evol. 2015;32:970–7.

83. Krzywinski M et al. Circos: an Information Aesthetic for Comparative Genomics. Genome Res. 2009;19:1639–45.

84. Rizzon C, Ponger L, Gaut BS. Striking similarities in the genomic distribution of tandemly arrayed genes in Arabidopsis and rice. PLoS Comput Biol. 2006;2:0989–1000.

85. Shoja V, Zhang L. A roadmap of tandemly arrayed genes in the genomes of human, mouse, and rat. Mol Biol Evol. 2006;23:2134–41.

86. Narnoliya LK, Kaushal G, Singh SP, Sangwan RS. De novo transcriptome analysis of rose-scented geranium provides insights into the metabolic specificity of terpene and tartaric acid biosynthesis. BMC Genomics. 2017;18:74.

87. Darwish O, Shahan R, Liu Z, Slovin JP, Alkharouf NW. Re-annotation of the woodland strawberry (Fragaria vesca) genome. BMC Genomics. 2015;16:29.

88. Bhat SA, Ahmad SM, Mumtaz PT, Malik AA, Dar MA, Urwat U, et al. Long non-coding RNAs: Mechanism of action and functional utility. Non-coding RNA Res. 2016.

89. Bai Y, Dai X, Harrison AP, Chen M. RNA regulatory networks in animals and plants: A long noncoding RNA perspective. Brief Funct Genomics. 2015;14:91–101.

90. Da Silva C, Zamperin G, Ferrarini a., Minio a., Dal Molin A, Venturini L, et al. The High Polyphenol Content of Grapevine Cultivar Tannat Berries Is Conferred Primarily by Genes That Are Not Shared with the Reference Genome. Plant Cell. 2013;25:4777–88.