9-1-2010

Exploratory analysis of pathogen resistance responsible genetic elements in two eudicots from whole genome sequence

Aubrey Bailey

Exploratory Analysis of Pathogen Resistance Responsible Genetic Elements in Two Eudicots from Whole Genome Sequence

Approved: ________________________________
Director of Bioinformatics

____
Head, Department of Biological Sciences

Submitted in partial fulfillment of the requirements for the Master of Science degree in Bioinformatics at the Rochester Institute of Technology.

Aubrey Bailey

September 2010

Abstract

Agricultural products are becoming more homogenous in order to conform to consumer expectations. In doing so, they have become a high-risk target for widespread crop failure due to pathogens or other agricultural maladies. Presented here are a collection of tools and a work-flow for the wholesale detection and characterization of disease resistance-associated proteins from whole-genome sequences. These techniques were also adapted for identifying potential miRNA regulatory sequences.

Putative R-gene sequences identified from the Malus x Domestica 'Golden Delicious' genome were acquired for verification against wild apple species. Resistance gene analogs were PCR amplified from R-gene associated domains (TIR, NBS, and LRR) in wild apple cultivars. Known R-gene sequences were clustered alongside the PCR analogs and putative R-genes from 'Golden Delicious'.

A covariance model approach for the *de novo* detection of 3,187 putative pre-miRNA regulatory sequences is also explored. *Vitis vinifera* sequences used to build this model were retrieved with 38% efficiency.

Computational clusters may be useful in the physical mapping of sequences to the chromosomally duplicated gene clusters characteristic of R-genes. Known R-genes included in computational clusters may also clarify the function of the unknown sequences. A possibility exists for the identification of resistance genes that have been lost in the selective breeding of commercial cultivars.

Thesis Committee Members:

Thesis advisor:
Dr. Sandra Connelly
Assistant Professor,
School of Biological and Medical Sciences,
Rochester Institute of Technology

Thesis committee members:
Dr. Angela Baldo
Computational Biologist
U.S. Department of Agriculture
Cornell University

Dr. Michael Osier
Associate Professor, Director of Bioinformatics
School of Biological and Medical Sciences,
Rochester Institute of Technology

Dr. Gary Skuse
Professor, Department Head
School of Biological and Medical Sciences,
Rochester Institute of Technology

Acknowledgments (in no particular order)

Kevin and Virginia Bailey, Muriel Hykes, and Paul Shuch – thanks for all the fish

Bryn, Devin, Ariel, Curran, Adam, and Erin – for motivating me in your own way

My teachers and professors, especially those who served as advisers and committee members – for keeping me on track and my education up to this point

All of my friends and colleagues, especially my editor, April

Eighth grade ISCS – for teaching me to be comfortable in a laboratory

Camp Dittmer – for housing and feeding me summer 2009

The Internet – for all of its research and distractions

God – for everything. Literally.

Angela Baldo, Gennaro Fazio and the rest of the NYS-ARS and USDA – for a summer of employment, the roots of this thesis and all of my data.

Cornell University – for their journals and IT resources

RIT – for a wonderful education and six years of adventure

Williamsport Area School District – for a good launch at college

The Town of Elimsport, esp. EUMC – for a stable foundation that has never changed

and Haley, my loving fiancée

Without any of which I may not have completed this.

Table of Contents

1. Introduction

1.1 Overview

The field of genomics is one of the core competencies of a bioinformaticist. Complete genome sequences and their associated data are being produced at an exciting rate, and that speed is increasing. As the machinery for sequence production hits its stride in the industry, better and more tailored algorithms are required to find regions of interest within this sea of data. Genomically interesting regions may be genes, gene groups, and sequences regulating the expression of those genes. Increasingly the lesson is being learned that it is not so much the gene that is used, but *how and when* it is used. These genes and regulatory regions have been found to be capable of expressing such facets of organism development as the shape, color, position, number and protruding angle of appendages. (Harfe 2005)

High-level computational techniques such as machine learning should be embraced to fill the role of genomic sequence identification. This category of algorithms can discern a pattern responsible for a gene or its regulation from representative sequences of interest, rather than requiring a researcher to find this pattern first. (Eddy 1998)

1.2 The problem

As our global society becomes more pressed for the biologically affected resources of land, food, and water, the demand for efficiency in biological products stands to increase as well. The most common and domestically visible of biological products are those of the agricultural sector. The U.S. is the single largest national contributor to the global agriculture industry. Plant agriculture alone comprises nearly 1% of the United States GDP. (Dimitri 2005) This position has been cultivated through a combination of industrial and scientific advances which allowed significantly more production per land area than competitors. This edge has been further honed through certain compromises in the industry. Synthetic fertilizers, monoculture and pesticide use have helped to raise production rates dramatically but come with significantly negative ecological impacts. Because of these ecological impacts many of these practices are being deconstructed as “unsustainable” and are being economically penalized through fines, taxes, negative media attention, or even outright banned in vulnerable locations such as those near wetlands.

While synthetic fertilizers and pesticides have their roots in the field of Chemistry, crop monoculture comes as a result of biological and business processes. Monoculture is the technique of focusing complete attention on a single product to the exclusion of all others. This can be likened to the Henry Ford’s assembly line business-model of producing only one type of car. It allows for specialization in quantity far beyond normal methods at the cost of diversity and flexibility.

Fortunately for farmers, consumers in the U.S. hold a certain fondness for the assembly line model and have expressed through their purchasing patterns that certain varieties, or cultivars, are preferred above all others. Iceberg lettuce, Portobello mushrooms, Haas avocados, navel oranges, Golden/Red Delicious apples, Concord grapes, and sweet corn have all benefited from the unflinching loyalty of their consumer base. Because of their popularity among commercial producers these varieties have become known as “commercial cultivars”.

Commercial cultivars are usually selectively bred for traits such as size, shape, texture, taste, color, and ease of preparation (seed-less, cling-free stones, etc.) as these are characteristics on which consumers base their purchases. In a classic case of 'you get what you ask for', the selection process tends to drop traits that aren't selected for. (OMAFRA 1998) These lost traits are ones that are not visible to consumers such as nutrient usage, weather tolerance, and disease resistance. While nutrient use and weather tolerance are modifiable characteristics in plants, with many options available to the farmers, the third choice, disease resistance, is less so.

The real danger of crop monoculture lies in the inherited basis for disease resistance that is more strongly represented in plants than metaphyta. Many of these commercial cultivars are reproduced clonally, by splitting off part of the living plant and rooting it elsewhere. Others are only permitted to self-polinate or to be pollinated by members of their genetic lineage. This leads to an extremely homogenous population, which is good for producer and consumer expectations, but also good for viral and microbial expectations. (Buddenhagen 1977) If all descendants of a commercial cultivar are closely related then it stands to reason that a virus or mold capable of successfully infecting one individual will rapidly infect an entire harvest.

There exists a significant history of sweeping pathogenic destructions in mass-produced crops such as the 1850's Irish potato famine and the Phylloxera plague that nearly eradicated European *Vitis vinifera* (wine grape) cultivars at the end of the 19th century. These are precedents for multi-billion dollar damages to U.S. agriculture from *currently* low impact diseases for which commercial cultivars may widely lack resistance.

As a business practice, monoculture is far too successful to be warned against to any effective measure, but as scientists we can identify disease risks and potentially modify these commercial cultivars to be resistant against such incursions. As the demand for agricultural products and global population continue to rise, so will high yield techniques. Industrial producers are acutely aware of the risk of pathogen-induced crop losses and even using modern techniques it is estimated that disease claims between 5 and 10% of any given agricultural yield per year. (Kansas 2009) For this reason alone it is worth examining the innate ability of plants to resist these losses. Improvements in this field would also help to offset the over 22 million acres of land that were chemically treated as a preventive measure against these diseases in 2007 (2007 USDA Census of Agriculture, pg 59).

Genomic investigation of agriculturally significant organisms for genetic elements and molecular markers is not only economically advantageous and environmentally responsible, but also scientifically feasible for the small (~500Mb) genome sizes in question. Through comparison to the markers identified in commercial cultivars and their wild relatives, selective breeding initiatives can ensure that disease resistances are not lost or are reincorporated in the process. The assessment of inherited risk factors and their appropriate responses must be performed now before such a pandemic can be allowed to occur.

1.3 **Biological Background**

1.3.1 Plant immune response

Plants respond to pathogenic or otherwise stressful stimuli differently from metazoans. Disease resistance and abiotic stress response is inherited through genomically encoded signaling cascades rather than the mobile immune response of t-cells and antibodies common to the animal kingdom. Genes coding for the initial recognition of pathogenic factors and the corresponding catalysis of the plant's response are known as R(resistance)-genes. R-genes usually code for extracellular or transmembrane proteins, usually pattern-recognition receptors, which detect

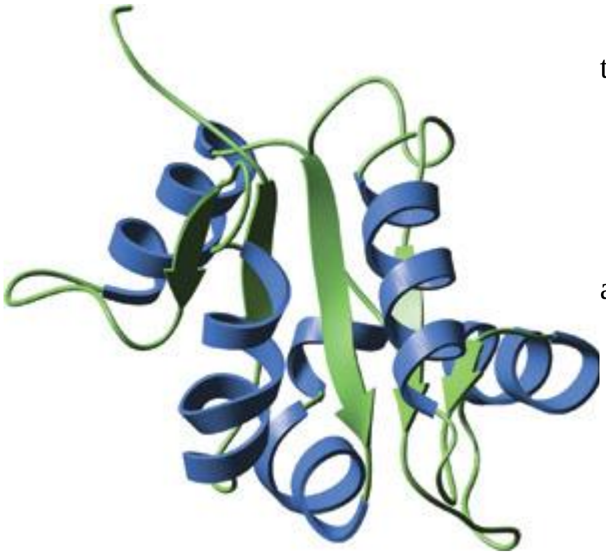
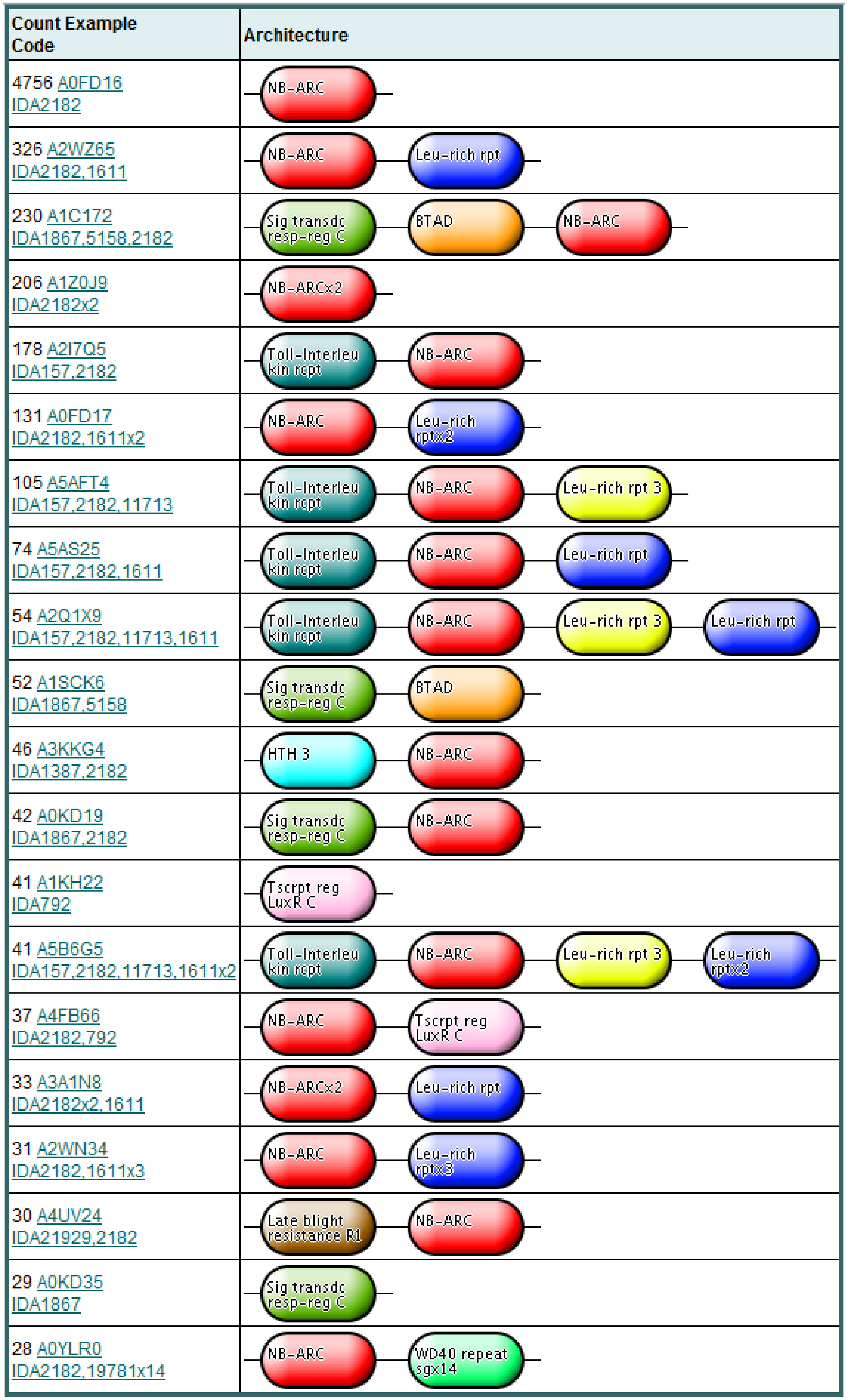


Figure 1: The TIR domain of human TLR2.

responsible states. These proteins then trigger a variety of immune responses dependent on the detected state. (DeYoung & Innes 2006) Plant immune responses are frequently extracellular messages that warn the uninfected cells of the organism to take action. Sometimes the response is the localized, hypersensitive response: apoptosis of the infected cell to prevent further spread of the disease. This response is contained within each somatic cell as detection, rather than through a coordinated mobile defense (t-cells). Plant resistance genes typically occur in clusters and may account for as much as 2% of the protein-coding genes in *Arabidopsis thaliana*. These clusters are thought to be the results of linkage-groups of R-genes during ancestral chromosomal duplication events. (Yin 2004)

Like most classes of genes, R-genes share similar motifs used for pattern recognition and signal transduction. In plants some common elements are the TIR (toll/interleukin receptor), which is similar to NOD (nucleotide-binding oligomerization) in mammals, and LRR (leucine-rich repeat) domains terminal of an NBS (nucleotide-binding-site) domain. (Michelmore 1998) By examining genes that are known to have these domains, and other known R-genes from evolutionarily related species, researchers have produced computational models for the identification of potentially novel R-genes as well as homologous forms of known genes in related organisms (Figure 2).



Count Example Code	Architecture
4756 A0FD16 IDA2182	NB-ARC
326 A2WZ65 IDA2182,1611	NB-ARC — Leu-rich rpt
230 A1C172 IDA1867,5158,2182	Sig transdc resp-reg C — BTAD — NB-ARC
206 A1Z0J9 IDA2182x2	NB-ARCx2
178 A2I7Q5 IDA157,2182	Toll-Interleu kin rcpt — NB-ARC
131 A0FD17 IDA2182,1611x2	NB-ARC — Leu-rich rptx2
105 A5AFT4 IDA157,2182,11713	Toll-Interleu kin rcpt — NB-ARC — Leu-rich rpt 3
74 A5AS25 IDA157,2182,1611	Toll-Interleu kin rcpt — NB-ARC — Leu-rich rpt
54 A2Q1X9 IDA157,2182,11713,1611	Toll-Interleu kin rcpt — NB-ARC — Leu-rich rpt 3 — Leu-rich rpt
52 A1SCK6 IDA1867,5158	Sig transdc resp-reg C — BTAD
46 A3KKG4 IDA1387,2182	HTH 3 — NB-ARC
42 A0KD19 IDA1867,2182	Sig transdc resp-reg C — NB-ARC
41 A1KH22 IDA792	Tscrpt reg LuxR C
41 A5B6G5 IDA157,2182,11713,1611x2	Toll-Interleu kin rcpt — NB-ARC — Leu-rich rpt 3 — Leu-rich rptx2
37 A4FB66 IDA2182,792	NB-ARC — Tscrpt reg LuxR C
33 A3A1N8 IDA2182x2,1611	NB-ARCx2 — Leu-rich rpt
31 A2WN34 IDA2182,1611x3	NB-ARC — Leu-rich rptx3
30 A4UV24 IDA21929,2182	Late blight resistance R1 — NB-ARC
29 A0KD35 IDA1867	Sig transdc resp-reg C
28 A0YLR0 IDA2182,19781x14	NB-ARC — WD40 repeat sgx14

http://www.ebi.ac.uk/interpro/ISpy?ipr=IPR000767&mode=ida

Figure 2: A variety of R-gene associated domain architectures observed in Interpro.

1.3.2 Protein Domains

Protein domains are portions of a protein sequence between 25 and 500 amino acids in length, usually around 200. Domains are usually self-stable despite attachment to the rest of their associated protein. As such, domains mutate, act, and are selected pseudo-independently of the rest of the molecule. The TIR, NBS and LRR domains are distinguished from other domains through their involvement with the plant immune response. (Yin 2004) While the individual roles of each domain are poorly understood, their involvement in the process is well-documented. While all NBS-LRR-class gene products contain an NBS domain, the presence of the TIR domain seems to be optional and the number of LRR repeats is highly variable (Figure 3).

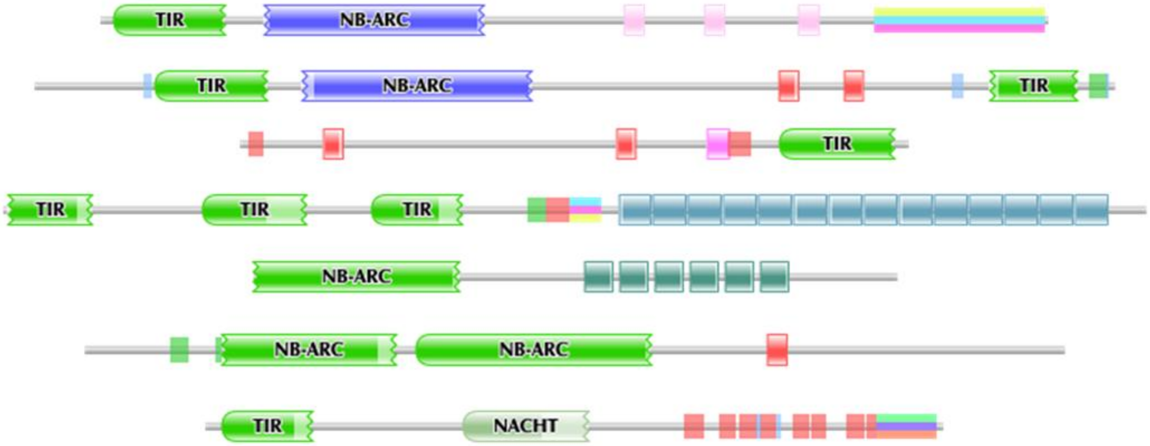


Figure 3: A collection of domains in genes. Unlabeled blocks are different types of LRRs.

The TIR domain (Illustration 1) is part of the Toll-like receptor (TLR) family and functions as a pattern recognition receptor. (Werling 2009) This domain is membrane-spanning and non-catalytic. TIR domain interactions between receptors and adapters play a key role in activating conserved cellular signal transduction and receptor signaling mediation pathways in response to bacterial LPS, microbial and viral pathogens, cytokines and growth factors. (Xu 2000) The central NB-ARC (Figure 4) domain in R-proteins has been proposed to function as a molecular switch that, depending on the nucleotide bound defines the activation state of the R-protein.

The NB subdomain is the catalytic core of a functional ATPase domain, and its nucleotide-binding state is proposed to regulate activity of the R-gene product.

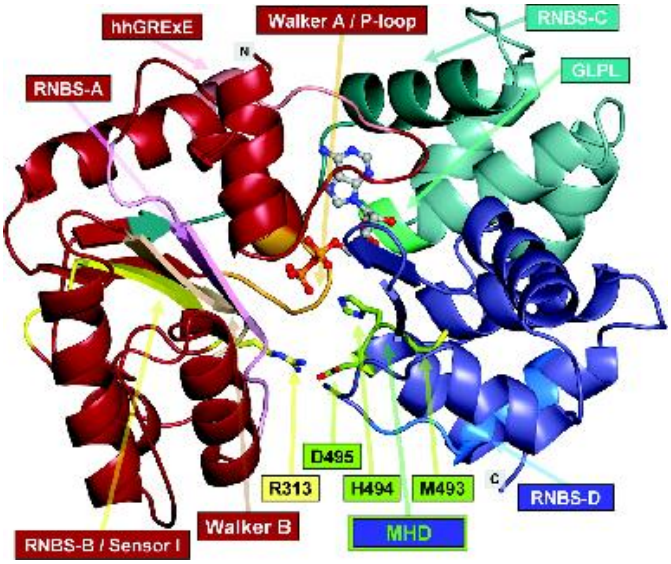


Figure 4: Structural model of the NB-ARC domain of I-2. The central domain of the NBS-LRR type of R-gene (Ooijen 2008)

The LRR is a versatile binding motif that is particularly rich in hydrophobic leucine amino acids. Each LRR repeat is 26-29 residues in length and forms a “horseshoe-bent-spring-shaped” structure of external helices and internal sheets. Collections of these repeats can form structures that are extremely flexible in terms of accommodating rapidly changing and diverse protein-protein interactions. (Hillig, RC. 1999) The TLR domain includes as many as 10 such repeated subdomains consecutively. (Figure 5) The protein family database (Pfam) maintains an annotated collection of such structures, multiple sequence alignments, and profile HMMs for each of these functional domains.

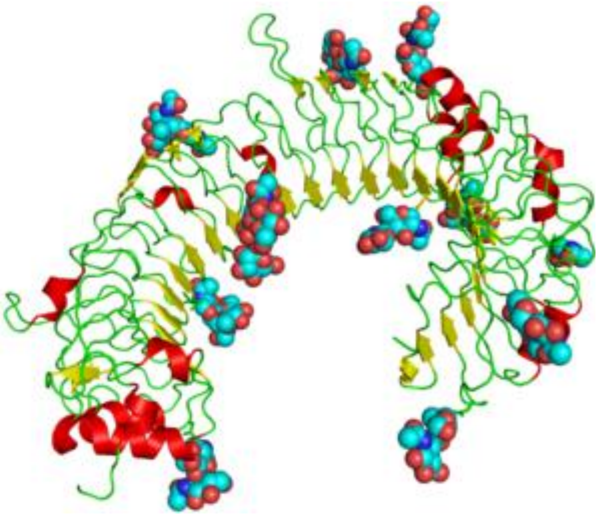


Figure 5: The curved leucine-rich repeat region of Toll-like receptors, represented here by TLR3. A versatile binding motif. http://commons.wikimedia.org/wiki/File:TLR3_structure.png

1.3.3 miRNA regulation

Identifying the presence or absence of an R-gene in the genome of a cultivar is not sufficient to determine its phenotypic expression. A genomically coded, transcribed sequence can be silenced post-transcriptionally by non-coding regulatory elements. For example, many avian species have the genes responsible for the generation of tooth-like structures, but lack teeth due to silencing by untranslated regulatory sequences. (Phillips 2008) To conduct a thorough study of R-genes it is necessary to search genomic sequences for evidence of such post-transcriptional regulatory elements.

At the start of the Human Genome project, it was believed that there would be close to 100,000 protein-coding genes identified at its completion. In the current literature there are now predicted less than 25,000 protein-coding genes. This frustrated researchers for several years into wondering what they were doing wrong. Regions previously discarded for not conforming to the traditional notion of a genetic element are now being reexamined for transcription as regulatory elements such as small interfering RNA (siRNA) and miRNA. (John 2004)

miRNA are small (~22nt), highly conserved, nucleic sequences. These short sequences bind with stringent complementarity to mRNA transcripts, usually in the 3' UTR. Once bound, miRNA affect expression of their target, usually by silencing it via miRNA directed cleavage. Pre-mature, “stem-loop” miRNA are formed from hairpin-like structures created as the result of single stranded base-pairing in primary transcripts. This means that a reverse complementary sequence (or some degenerate form of one) resides up or downstream of the site of a mature miRNA. (Krol 2004)

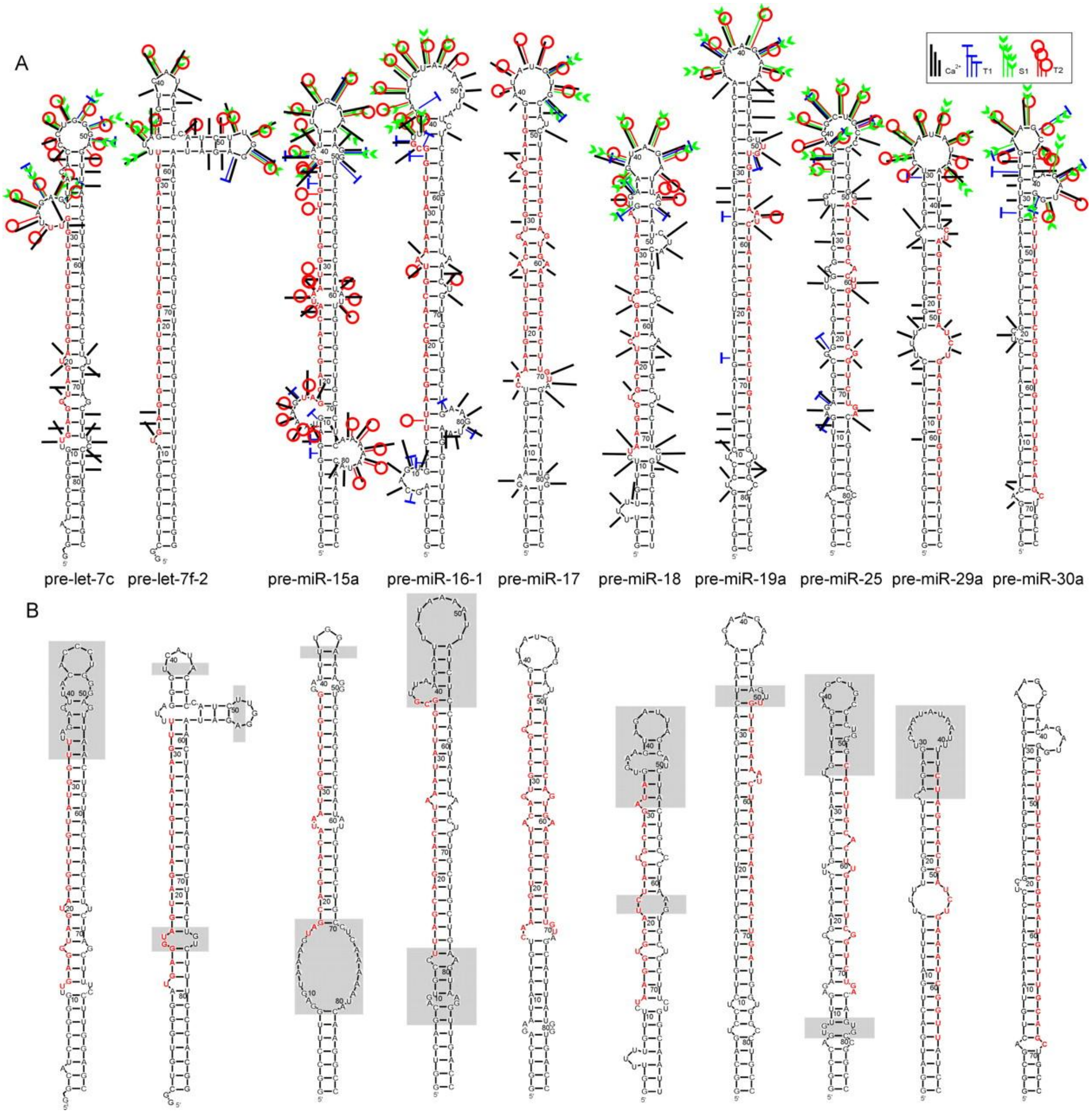


Figure 6: Experimentally determined and predicted structures of 10 microRNA precursors.(Krol 2004)

Most miRNA genes have similar characteristics to genetic elements such as promoters, repressors and intergenic regions. Like most RNAs, miRNA also undergo some form of post-transcriptional splicing in their single-stranded state in order to attain their active geometry. (Figure 6) Unlike most genes and mRNA, miRNA primary transcripts are usually polycistronic, containing multiple discrete loops which will each go on to affect a different target. Enzymes including Dicer and the miRISC play important roles in the maturation of the transcribed miRNA, and the facilitation of interaction between it and its target. The exact mechanisms are of this process are the subject of an astounding amount of research but are currently poorly characterized. (Krol 2004)

Such miRNA systems are being hailed as one of the key elements of eukaryotic genomic regulation and are thought to be the answer to many of biology's unanswered questions regarding organism level control and development. miRNAs play a paramount role in sequence specific intracellular and extracellular expression modulation, both down (silencing) as well as up (silencing a repressor). The physiological effects of miRNA regulation are observed in cellular differentiation, apoptosis, proliferation, metabolism, hematopoiesis, limb morphogenesis and other developmental patterning. (Chen 2004) Plant miRNAs are high-level regulators of gene expression. Mutants impaired in miRNA biogenesis exhibit severe abnormalities across diverse gene-groups. Similarly, plants that over-express particular miRNAs or express miRNA-resistant versions of particular miRNA targets exhibit a wide array of unusual phenotypes. (Jones-Rhoades 2006) The flexibility, specificity, and robustness of microRNA systems are such that they are being recognized for their role in extracellular as well as organism-level modulation, possibly including a method of systemic response to R-gene activation.

1.4 Organisms of interest

Currently eight whole-genome-sequences for plants have been completed and published and more than 50 are in progress. Most of these sequencing efforts are focused on commercial cultivars. Because of these efforts we are given a window into plant biology that makes specific characterization and study of commercial cultivars possible. Identifying the genetic bases for the phenotypes of interest is the first step in understanding and improvement.

1.4.1 Family Rosaceae

The family Rosaceae is a plant family of around 100 genera including a number of agriculturally significant herbs, shrubs, and trees including apples, pears, raspberries, strawberries, peaches, plums, cherries, almonds, and of course roses from which the family draws its name. This family is well represented both agriculturally as well as scientifically in terms of the amount of data available in major repositories from several members of the family. Spirally arranged, serrated leaves, brightly colored flowers, and a wide variety of fruit from which it draws its astounding agricultural representation, characterize the family. (Shulaev 2008) In the U.S. over $7 billion of rosaceous crops are produced yearly with projections to expand especially in the global market. (Slovin 2006) Many crops in this family are reproduced clonally, by grafting. This makes them good targets for study as well as high-risk targets for pathogenic agents. Among these, the apple (*Malus domestica)* holds special representation for its current status of having the most genes of any studied plant genome at 57000. (Velasco 2010) By tons produced, it is also the most widely cultivated of any rosaceae. An Italian agricultural research institute, IASMA, is close to finishing the *Malus domestica x GD* (Golden Delicious apple) genome and has provided some advance R-gene-candidate sequences for this project.

Table 1: Comparison of the apple genome to other sequenced plant genomes
Velasco et al. Nature Genetics (8/2010) | doi:10.1038/ng.654

Name	Genes/gene density	Transposable elements	Transcription factors	miRNAs	Resistance genes	Biosynthetic genes
Soybean	46,430/0.42	50.3	5,671	41 (85)	392 (61)	958
Poplar	45,654/0.94	35.0	2,758	174 (234)	402 (59)	1,034
Arabidopsis	27,228/2.2	18.5	2,437	89 (199)	178 (32)	719
Grape	33,514/0.66	21.5	2,080	130 (137)	341 (57)	1,121
Rice	40,577/0.97	39.5	2,798	140 (447)	535 (89)	910
Brachypodium	25,532/0.94	28.1	2,187	62 (129)	238 (89)	390
Sorghum	34,496/0.47	62.0	2,312	116 (148)	245 (75)	555
Maize	32,540/0.15	84.2	5,246	153 (170)	129 (74)	457

1.4.2 *Vitis vinifera*

Genus *Vitis* is also agriculturally significant. This genus contains all grape species from common edible varieties like 'Concord' and 'Cabernet Sauvignon', to inedible ones such as the experimental 'Pixie' and fox grapes. Most of the grapes grown in the U.S. are table grapes; however the U.S. also exports over $1 billion of wine annually. Globally this is about 6% of market share, but as European countries lose ground and Prohibition-era laws are repealed, that number stands to grow significantly. (Brunke 2003)

Vitis vinifera is the most cultivated and economically important grape species. Grapes are not only consumed fresh but also processed in making juice and wine. Though new cultivars of grapes have been developed over last few decades, their introduction into the market has been difficult because of strict regulations governing the propagation of varietals and their modification. For this reason, study rather than modification of these cultivars is an appropriate course of action so that pathogenic consequences can be understood and prepared against.

V. vinifera is a diploid plant with 2n = 38 chromosomes. It has a comparatively small genome of about 500 Mb which makes genome analysis feasible. Large scale ESTs are already available for *V. vinifera*. The Pinot Noir grape, *Vitis vinifera* L. cv. Pinot Noir has two complete genome sequences that are published and available, one with 8x (Table 2) (Refseq Project ID: 33471) and one with 12x coverage. The 8x sequence is directly comparable to the sequencing techniques of the *Malus* initiative as both were completed by IASMA.

Table 2: Genome statistics for Vitis vinifera L. cv. Pinot Noir (Refseq Project ID: 33471). Although the genome is much larger than A. thaliana there is a comparable number of protein-coding genes. Genome size does not correlate to complexity.

Name	RefSeq	Length (Mbp)	GC content	Proteins	RNAs
Chromosome 1	NC_012007	15.6308	0.0%	968	57
Chromosome 2	NC_012008	17.6034	0.0%	826	61
Chromosome 3	NC_012009	10.1869	0.0%	774	57
Chromosome 4	NC_012010	19.2931	0.0%	1169	54
Chromosome 5	NC_012011	23.4283	0.0%	1259	54
Chromosome 6	NC_012012	24.1489	0.0%	1193	62
Chromosome 7	NC_012013	15.2337	0.0%	1145	53
Chromosome 8	NC_012014	21.5572	0.0%	1440	67
Chromosome 9	NC_012015	16.5322	0.0%	765	52
Chromosome 10	NC_012016	9.64704	0.0%	378	26
Chromosome 11	NC_012017	13.9363	0.0%	803	34
Chromosome 12	NC_012018	18.5408	0.0%	947	58
Chromosome 13	NC_012019	15.1919	0.0%	944	44
Chromosome 14	NC_012020	19.4804	0.0%	1264	86
Chromosome 15	NC_012021	7.69361	0.0%	404	26
Chromosome 16	NC_012022	8.15885	0.0%	415	25
Chromosome 17	NC_012023	13.0591	0.0%	862	46
Chromosome 18	NC_012024	19.6913	0.0%	1397	61
Chromosome 19	NC_012025	14.0718	0.0%	723	55
Master WGS	NZ_CAAP00000000	467.476	32.0%	23335	1374
Plastid Pltd	NC_007957	0.16	37.0%	84	53
Mitochondrion MT	NC_012119	0.773279	44.0%	74	34

1.4.3 Arabidopsis

Arabidopsis thaliana is a small flowering plant of the mustard family, brassicaceae, with the smallest and most investigated plant genomic sequence to date. *A. thaliana* is a diploid plant with 2n = 10 chromosomes. It became the first plant genome to be fully sequenced based on the fact that it has a small genome of ~120 Mb with a simple structure having few repeated sequences. Its short generation time of six weeks from seed germination to seed set and its large number of seeds have facilitated its adoption as *the* model organism of plant molecular biology. Though of no economic importance, it is an invaluable resource to agriculturally important crops, particularly to members of the same family, which includes canola, an important source of vegetable oil. Its full sequence is readily available and is frequently compared in most plant genomic research. Homology between *V. vinifera*, *Malus* sp., and *A. thaliana* R-genes and miRNAs are germane.

Table 3: Arabidopsis thaliana - The Arabidopsis Information Resource (TAIR9) Statistics showing a comparison of well characterized protein and miRNA genes in a small genome.

Chr	chromosome length (bp)	Protein coding	pre-tRNA	rRNA	snRNA	snoRNA	miRNA	Other RNA	Pseudo genes	TE genes	Total
1	34,964,571	7,054	240	0	2	18	52	107	242	683	8,398
2	22,037,565	4,237	96	2	0	15	29	75	218	825	5,497
3	25,499,034	5,436	93	2	7	15	29	67	201	878	6,728
4	20,862,711	4,214	79	0	0	11	28	48	121	711	5,122
5	31,270,811	6,318	123	0	4	12	36	53	144	804	7,494
All	**134,634,692**	**27,169**	**631**	**4**	**13**	**71**	**174**	**350**	**926**	**3,901**	**33,239**

The theory of evolution posits that organisms share common ancestors. It follows that these organisms also share common genes. For this reason research in one organism is at least partially relevant to any related organism. The process of locating these shared sequences is a discipline known as 'homology searching', so-called for the technique's assumption of evolutionary relatedness between species. Known R-genes can be located across classes and families of agricultural significance in seconds using this technique, but a set of caveats comes with such extrapolations.

1.5 Computational Background

Gene sizes are measured in hundreds to thousands of bases, while genomes are measured in hundreds of millions to billions. *A. thaliana*, the current smallest green plant genome still has in excess of 100 million nucleotide pairs in its genome.

Traditional computational techniques were suited to this scale of data, but not complex enough to address the irregular and seemingly whimsical nature of biological features. To address these problems brand new classes of algorithms were created using probabilistic models and non-linear mathematics. These solutions rely on high quality biological observations and evolutionary estimations that had been previously calculated in order to extrapolate back into the evolutionary past. One of the most preeminent of these algorithms is a method for finding an inexact match of one sequence inside of a second longer sequence.

1.5.1 BLAST

Computers have been easily able find an *exact* match between two strings of characters; however, genomic sequences are known to change over time. This means that related sequences will seldom match one another perfectly. The current solution is known as a 'local alignment', and the *de facto* standard is the Smith Waterman algorithm, most famous for its implementation in BLAST, the Basic Local Alignment and Search Tool. BLAST can pull any number of needle-like objects from a genomic haystack within a matter of seconds and some iterations even include rules for the likelihood that different types of changes occurred evolutionarily. It presents the user with a ranking of how closely each “hit” matches the queried sequence. Because genomes are so large there exists a chance that arbitrary sequences will bear coincidental resemblance to a shared sequence that in reality is not derived from a common ancestor. This is controlled for through a statistic known as an expected-value (e-value) that represents the chance that an arbitrary match of the queried sequence was presented as a result. (McGinnis S., & Madden T.L. 2004) Even using these controls there is no guarantee that finding a sequence similar to a known one is meaningful. No sequence is completely independent of other biological features, such as morphology, which contribute substantial knowledge to conservation.

1.5.2 T-COFFEE

Once multiple hits have been established, comparing those sequences side by side is a logical step. A second class of algorithms, known as MSA, or multiple sequence alignment, fills this role. The currently favored implementation is known as T-COFFEE. T- COFFEE compares sequences as expected, but begins to take exponentially longer as more sequences are added to an alignment. If the sequences that are being aligned were likely to return many hits, such as searching for all sequences in a genome that have loose resemblance to the TIR domain, then the alignment might have thousands of sequences to work with and require months to complete.

1.5.3 Profile Hidden Markov Models

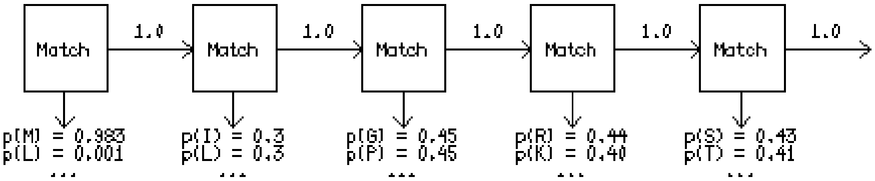


Figure 7: Basic form of a Markov probability chain showing transition probabilities between the various states of statistical recsidence.. Hidden Markov models make no assumption of current state. (Eddy, SR. 1998)

One of the crowning achievements of computational biology steps in here. Profile Hidden Markov models (Figure 7) are a statistical construct bordering on machine learning. Profiles are constructed from consensus among a set of representative sequences, such as a small handful of properly aligned sequences bearing a domain of interest. From this data the method models the statistical probability of changing from one state (amino acid residue) to the next in a chain of events, or sequence. Part of each of the two major HMM sequence suites, HMMER (Eddy SR. 1998) and SAM, (Hughey, R. and Krogh, A. 1995) is the ability to align sequences to a model. This is significantly faster (arithmetic over geometric) than aligning the sequences to each other because there is only ever one sequence at a time being aligned to the model. Such models can also be used to score entire chromosomes for regions conforming to their profile. This has become a standard for the prediction of previously uncharacterized genes or other sequences of interest.

1.5.4 Hierarchical Clustering

Even with a quality MSA, composing sequences may be located many sequences away from their closest relatives. A phylogenetic estimation can infer evolutionary relationships from differences between aligned sequences. The fewer the number of permutations, the more related the sequences are likely to be. These same techniques can be used as a form of biologically sensitive clustering. This is desirable if sequences of a known function are included as part of the distance estimation because their collocation in clusters with unknowns predicates similar sequence and hence a likely similar function. A package of tools for garnering phylogenetic inferences is Phylip. (Felsenstein, J. 2005)

One of the most time consuming and robust features of the Phylip package is its ability to jumble a data set and perform the same calculation multiple times. This is akin to performing hundreds of the same wet-laboratory experiment. While a significantly different outcome is not expected, it closes the door on irregular circumstances. Phylip attains this through a method known as 'bootstrapping'. The multiple data sets generated in this fashion have the option to carry their results through each of several associated programs; distance matrix calculation, phylogeny estimation, and tree construction. Ultimately the assembled clusters produced bear significant, high-quality evidence of their intra-relation. (Wang 1998)

1.6 Hypotheses

By identifying R-gene associated sequences we can create markers for selective breeding or transfect commercial agricultural products with the missing or reactivated versions of genes that their more hardy relatives possess without having to alter existing processes or expectations.

Here it is hypothesized that some NBS-domain-bearing RGAs identified in wild cultivars will cluster more closely with NBS-domain bearing sequences from the IASMA identifications than other arbitrary matches within the same organism. Furthermore, it is conjectured that there is at least one miRNA responsible for regulating R-gene translation in *V. vinifera*. This miRNA may be found through homology searching of known RGAs against genomic sequence or ab initio prediction of many putative miRNAs from genomic sequence and their subsequent comparison to *V. vinifera* R-genes.

2. Materials and Methods

Presented here are the tools and techniques used to design *in silico* experiments for the detection and characterization of large quantities of unknown sequences with a high likelihood of being involved in agricultural crop disease resistance and abiotic stress response. Through the combined use of these methods, clusters of sequences have been elucidated from genomic sequence.

2.1 Overview

All of the work that follows extends from at least one of the five (Geneva, Rose, IASMA, *V. vinifera*, and miRNA) sets of data described herein. The first set, collectively known under the topographical moniker 'Geneva', was a collection of degenerate PCR primer results from wild apple germplasm at the Geneva, NY ARS. This set represented empirically derived *Malus* R-gene analogs (RGAs). This set was conceptually translated into AAs and later became a subset of a larger collection of RGAs gathered from the NCBI representing the entire family Rosaceae, 'Rose'. The purpose of gathering the first two data sets was for their use in verifying, aligning and clustering alongside a set of computationally predicted R-genes from the Italian IASMA. These sequences were shared in collaboration between IASMA and the Geneva ARS in an effort to validate the predicted genes. Since the full *Malus* genome was not yet available, a comparable genome, *Vitis vinifera*, was drawn from the NCBI's refSeq. This provided a compatible sequence platform for the development of a miRNA prediction method.

2.2 Sequence Resources

2.2.1 IASMA

IASMA, the Agricultural Institute of San Michele all'Adige is a research institution in northern Italy that specializes in agricultural genomics, mapping, and phenotyping. They have previously published an 8x WGS of Pinot Noir (*Vitis vinifera* L. cv. pinot noir) and are currently leading the Golden Delicious apple (*Malus x domestica* 'Golden Delicious') sequencing project. IASMA has collaborated in the past with the USDA Agricultural Research Service (Geneva, NY) for computational biology resources. For this reason it was possible to obtain access to unpublished, conceptually translated, R-gene predictions from their Golden Delicious apple sequencing project. IASMA researchers selected these sequences for being predicted genes using 'FgenesH', 'Twinscan', 'GlimmerHMM[58]' and 'GeneWise'. These sequences were also selected for their conformation to the Pfam-defined NBS domain profile. Since genomic sequence was not available for the *Malus* initiative yet, it was decided that miRNA gene finding would done in the published *V. vinifera* sequence in order to maintain comparable sequencing technology and techniques.

2.2.2 USDA-ARS

The US Department of Agriculture, Agricultural Research Service (USDA-ARS; Geneva, NY, USA) was the source for the 'Geneva' and 'Rose' data sets. PCR amplification from degenerate RGA-specific-domain primers also produced a small library (Geneva-set) of *Malus* sequences from wild relatives of the genus.   The 'Geneva' set consisted of a set of 788 apple RGAs generated using four pairs of degenerate PCR primers based on the NBS domain described by Baldi et al. 2004. Specimens were taken from the USDA's apple germplasm collection in Geneva, New York. These accessions included 28 *Malus* species (*Malus x* *domestica, asiatica, baccata, florentina, floribunda, fusca, halliana, honanensis, hupehensis, kansuensis, micromalus, ombrophila, orientalis, prattii, prunifolia, pumila, robusta, sargentii, sieboldii, sieversii, sikkimensis, sublobata, sylvestris, transitoria, zumi, yunnanensis,* and *zhaojiaoensis),* hybrid White Angel, and the rootstock Geneva 41. Roughly 30% of these NBS identified sequences also included the TIR domain.

The second, 'Rose' set from the NYS-ARS was the product of previous RGA studies on *Rosaceous*-related genera. This set includes 75 RGAs from *Rubus*, 236 from *Prunus*, 18 from *Fragaria*, 151 from *Rosa*, 42 from *Pyrus,* and 252 from various wild apple species. More than 500 *Rosaceae* sp., 350 *Populus tritocharpa*, 390 *V. vinifera*, and 163 *A. thaliana* confirmed R-gene sequences were retrieved from the NCBI and contributed to this set so as to provide a more comprehensive collection of R-genes and their analogs. The combined set of these sequences provide a foundation from which the R-gene candidates can be identified and verified through sequence similarity.

2.2.3 *Vitis vinifera* and miRNA

In lieu of *Malus* nucleotide sequence, it was necessary to have a comparable substrate for the development of a workflow for miRNA-gene prediction. Because of IASMA's role in both the *V. vinifera* and *Malus* sequencing efforts, and the reasonable evolutionary distance between the two genera, the NCBI's RefSeq version of *V. vinifera* was chosen for development.

There existed a second *V. vinifera* sequencing project at the time of this analysis. Genoscope, the French National Sequencing Center, was working on a 12x run of inbred *Vitis vinifera* L. cv. pinot noir. This sequence was designed to be significantly less heterozygous than the IASMA iteration. These data are less likely representative of the cultivated grape, and less directly comparable and portable to the apple sequences and RGA libraries. Unfortunately, at the time of this research the Genoscope sequence was still at 8x coverage and had not yet been integrated into RefSeq. Because the mode of accession of this sequence was unknown and still volatile, and to preserve the future reproducibility of the results herein, accession numbers NC_012007.2 – NC_012025.2 from the IASMA effort were used. miRNA sequences for this prediction were gathered from *V. vinifera* stem-loop sequences registered in the Plant miRNA database (PMRD)(Zhang 2010) and miRBase (Griffiths-Jones S. 2004). These sequences in combination with Rfam (S. Griffiths-Jones 2003) models and alignments were used to construct the model for miRNA gene prediction.

2.3 Tools

The following tools were used in multiple ways throughout this investigation. What follows is a brief description of each tool, their role/function, and version.

2.3.1 COMPUTER – PIECHERRY

Any large-scale computational analysis requires a good computer. All of the *in silico* calculation was performed on a USDA server hosted at Cornell. This computer ran 16 Itanium cores and 82GB of memory to perform these analyses. The operating system was Red Hat Enterprise version 5.5.

2.3.2 BLAST

BLAST, The Basic Local Alignment and Search Tool (Zhang 2000) can quickly search large sequences for pattern matches that normal text searches miss. It is also attentive to biology specific details such as forward and reverse strands, nucleotide base-pairing, insertions/deletions, and amino acid substitution frequencies. In this situation BLAST can be used to rapidly find sequences close to known miRNA or RGAs for closer examination. Version 2.2.4 was used online at the NCBI and version 2.2.18 was installed on the server.

2.3.3 SAM/HMMER

SAM (Hughey, R. and Krogh, A 2009) and HMMER (Eddy SR 1998) are two Hidden Markov model sequence alignment and modeling suites. These programs use slightly different algorithms to construct profile HMMs from input sequences. The models are useful for scoring large sequences for their fit to the model as well as aligning sequences in arithmetic time. HMMs have proven to be extremely useful in the field of gene prediction. SAM was used for the construction of the full-scale alignments and HMMER was used early in the exploration of a HMM method of miRNA detection. HMMer version 3.0 and SAM v3.5 were installed on the server.

2.3.4 T-COFFEE

T-COFFEE (Notredame 2000) is a tool for pairwise, *de novo*, multiple sequence alignment. It is more flexible than its predecessor, ClustalW, because of its ability to combine protein, secondary structure, and DNA inputs in the same alignment. Although T-COFFEE is slower than an HMM alignment, HMMs require quality aligned sequences for their input and as such, cannot be used *ab initio*. T-COFFEE Version 8.93 was used for these alignments online at the Swiss Institute for Bioinformatics servers.

2.3.5 Phylip

Phylip, (Felsenstein 2005) the phylogeny inference package, is a package for computational phylogenetics. It is favored for its end-to-end pipeline of such useful and highly integrated programs as 'protdist', 'fitch', 'protpars' and 'neighbor'. Phylip requires a quality multiple sequence alignment as input. The various programs can then be used to compute distance matrices between the sequences at each character-location. It clusters them using this distance matrix via one of a variety of algorithms and estimates hierarchical clustering (or a phylogeny) based on this data. Two of these algorithms are Fitch and Kitch. These are “maximum-likelihood”-based programs that allow a tree to be rearranged after each sequence addition. Another algorithm, Neighbor Joining is significantly faster but does not allow for rearrangement of the tree. Kitch differs from Fitch in its assumption of a constant rate of evolutionary change throughout the tree. Fitch allows for different rates of change down each branch. Fitch was selected as the tree-estimation tool for the hierarchical clustering analysis. Phylip package version 3.6 was installed on the server.

2.3.6 INFERNAL/CMs

Hidden Markov Models are limited in their analysis of problems that involve multiple data that fluctuate in correlation with one another. Because of this they are limited to estimating transition probabilities in linear data sets. If a problem requires a second dimension in order to make sense of the first, a covariance model, or CM, is required. This is valuable in cases where secondary structure information, rather than primary sequence structure, is the conserved factor. Such is the case with noncoding RNAs. CMs calculate transition probabilities in much the same way as pHMMs but also measure the impact of changes in one variable on the other. INFERNAL (Nawrocki 2009) is a recently developed CM-based analog of HMMER. It uses the familiar interface of HMMER to build models using a combination of sequence consensus (profile) and RNA secondary structure consensus. Similarly to HMMER, these models can be used to search large sequences quickly and score regions for their fit to the model. INFERNAL version 1.0.2 was installed on the server.

2.3.7 Vienna Package

The Vienna Package (Gruber 2008) is the definitive RNA secondary structure prediction package. It has become so ubiquitous for the analysis of RNA that many other tools list it as a dependency for their installation. One program, 'RNAfold', uses three kinds of dynamic programming algorithms in order to predict the most likely folded conformation; minimum free energy (MFE), the suboptimal folding algorithm, and an RNA base pairing-sensitive iteration of the partition function (Figure 8). Another program in the package, LocARNA, performs secondary-structure sensitive MSAs. A standalone 'RNAfold' version 1.8.2 was installed on a local computer. The Vienna package version 1.8.2 was installed on the server and the University of Vienna's web-server running 1.8.2 was also used.

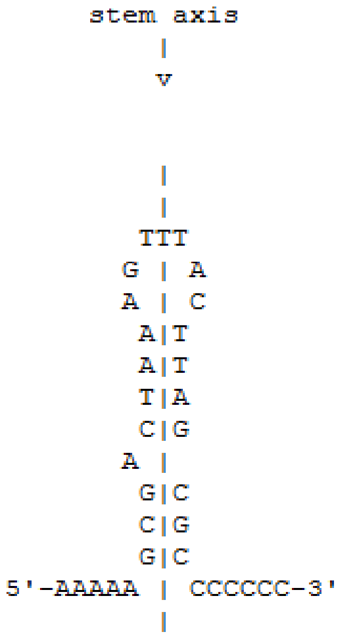


*Figure 8: The partition function - showing a partition along which an RNA may fold as a secondary structure. (*Mokrejš 2009)

2.3.8 Perl/Bioperl

Perl is an extremely powerful computational scripting language for scanning and altering text files. A set of bioinformatics modules known as BioPerl (Stajich 2002) provides a framework for managing common biological data types such as sequences and alignments. It also integrates into BLAST and other common programs to lighten the burden of code that must be written from scratch. Perl version 5.8.8 and Bioperl 1.0.0.5 were installed on the server.

2.4 Work Items

2.4.1 Verification of IASMA seq

Because the IASMA sequences had been conceptually translated, they were slightly less useful from a research point of view. Some quality assurances that the translations were done with attention to the underlying biology needed to be made. Strictly computational translations can overlook or misrepresent features that are biologically significant in the actual polypeptide, such as frame-shifts.

Simple BLAST commands lack the ability to account for frame shifts. A Perl script was written to check the sequences for frame-shifts and other confounding characteristics. It then sorted the resultant hits by percent sequence identity. This program was written in Perl, using the Bioperl BLAST and sequence handling modules and is fairly straightforward in its implementation. These sequences were also screened for vector and low-complexity/repeatable elements such as retrotransposons via BLAST against Univec and Repbase respectively.

Instead of attempting to convert IASMA's conceptually translated sequences back into their frame-ambiguous nucleotide sequence, the nucleotide PCR reads were translated into amino acid sequences. A translating iteration of BLAST (tBLASTx) was used against the IASMA sequences. The PCR RGAs in the Geneva-set and the externally acquired Rose-set could be six-frame-translated and locally aligned to the IASMA translated sequences. Geneva sequences bearing >90% identity to some translated sequence were very likely to be the correct translation and could then be kept for use in hierarchical clustering later.

2.4.2 Domain alignment

The full-sequences of resulting query/hit pairs were retrieved and aligned again using T-COFFEE. This aligned the entire sequences and not just the 'hit' area. The highest scoring pairs were then aligned as a larger multiple sequence alignment with T-COFFEE again. Aligning all of the sequences in this way would have been time prohibitive because of the aforementioned geometric run-time requirements of MSA programs.

The resulting alignment was used to build a profile HMM using SAM. SAM's 'align2model' was then used to align all of the input sequences to the model at arithmetic (as opposed to geometric) speeds. The resulting alignment organized all of the domains as co-linearly as possible across all of the sequences.

IASMA sequences that bore similarity to 'Geneva' RGAs should present themselves proximately to one another when clustered. Similarly, RGA candidates with orthologous targets, form, or function should align along the same motifs and, later, be clustered accordingly.

This was repeated with 6 different permutations of the model being built: Rose, Italy, a combined Rose+Italy, Rose+A.th., Italy+A.th., and Rose+Italy+A.th. The most informative of these alignments, 'Rose' with *A.thaliana*, was taken to the Phylip package for clustering and further analysis.

A second full-scale alignment was made using the domain specific models available from Pfam. By concatenating these models using the -A switch (Eddy 1998) in HMMER, the models for TIR, NBS-ARC, and LRR1/2/3 were synthesized into a single model consistent with the domains considered to represent the majority of R-genes in plants. Although this model was drawn from more evolutionarily distant organisms, it was considered to be relevant enough to warrant exploration. This model was concatenated to the empirical model to tailor the sensitivity to organism-level specific details.

The HMMER suite contains a program known as 'hmmpfam' to search for Pfam domains. It functions much the same way as hmmscore. This program rapidly scans large amounts of sequence data, returning probabilistically quantified sequence fit and the range over which each is supported. This program was used to determine and quantify the presence of each of the domains in the IASMA sequences. This further secured their validity as TIR-NBS-LRR domain-bearing R-genes.

A negative control was also constructed to ensure that the IASMA sequences were not giving false positives for the TIR-NBS-LRR domains. Three arbitrary domains, (C1q, Haemagg_act, and MA-Mit) were chosen from Pfam to construct another composite model that would likely have nothing in common with the IASMA sequences. The IASMA sequences were scored again with this model using hmmpfam.

In each alignment a handful of sequences were responsible for expansive sections wherein no other sequence aligned. To prepare the file for a more expedient calculation of evolutionary distances, these sequences were trimmed along the sites that were not informative to the rest of the alignment. (Budd 2007) Where there is little or no co-aligned sequence there is less information. Huge tracts of blank space require time and disk space.

To remove these uninformative sequences, a Perl script was written using the BioPerl alignment modules. This script indexed the alignment, counted the number of nucleotides in each column and output them in human-readable format. From there researcher discretion was used to distinguish between a short run of unmatched sequence and a longer, removable insertion. These removed sequences were kept for examination of similarity to known transposable elements from Repbase (Kapitonov 2005) or other sequence via BLAST's bl2seq.

2.4.3 Phylip clustering

A clustering tree was constructed from the trimmed, aligned, IASMA R-gene candidates. This tree illustrated the clustering and putative evolutionary distances between these sequences and the 'Rose' RGAs. By virtue of association, these clusters allowed the known sequences to lend their own characterization to the IASMA sequences. Association in this manner supports the hypothesis that those sequences actually encode proteins with the function of disease resistance. This association also presents clues as to which functional family of resistance they may belong to. It may also provide insight into historic adversaries that the organisms once shared. The location of a cluster without any of the unknown sequences as members may indicate that R-gene families from the ancestral varieties have been lost in 'Golden Delicious'.

The Phylip suite was used to construct these clusters. The alignments were bootstrapped with 'seqboot' to form 300 jumbled data sets for the further programs. A random seed of 7 was given. The protein distance matrices were calculated using 'protdist' and the Jones-Taylor-Thornton scoring method. The resulting distance matrices were each rotationally added to form 300 cluster-trees using 'fitch'. The input data were jumbled twice using a random seed of 7 and executed following the Fitch-Margoliash, weighted least squares, method. Due to time (80k+ minutes) and data-storage constraints, the entire analysis was reduced from the intended 500 bootstraps down to the performed 300. These 300 trees were combined into a single consensus tree using 'consense' and formatted with 'drawtree'. This entire process was carried out on both the Pfam and the PCR based models. The resultant clusters are presented and discussed in their appropriate sections.

2.4.4 The problem of the miRNA gene finder

Plant miRNAs have primarily been discovered through direct cloning and sequencing of small cellular RNAs or through comparative genomics. Such efforts have fallen short of the rapid pace of characterization afforded to other branches of genomic investigation. This is due in part to the lack of a broadly accepted method for their *de novo* detection.

Although primary miRNA transcripts are polyadenylated, and hence theoretically identifiable in EST collection efforts, their representation in EST libraries is close to 0.01%. Most estimates suggest that miRNA should account for ~1% of expressed sequences. Researchers believe that the preparation of cells for transcript collection is too hostile, or not sensitive enough to preserve these sequences. (Lazzari 2009)

Plant miRNAs tend to occur in gene families encoding identical or nearly identical mature miRNAs. There are at least 20 such miRNA families broadly conserved among flowering plants. These miRNA loci are highly conserved against changes in sequence or genomic positions. Only a handful of mutations at miRNA loci have been identified in genetic screens. (Jones-Rhoades 2006)

Although a number of methods were explored for the prediction of pre-miRNA genes, the only method that produced significantly meaningful results used a combination of secondary structure information alongside the statistical framework of a pHMM.

The CM (covariance model)-based suite INFERNAL was used to examine both of these data in tandem.

Drawing from the 'miRNA' set of known *Vitis vinifera* sequences from PMRD and miRBase, an RNA secondary-structure-sensitive alignment was made using the 'locaRNA' web

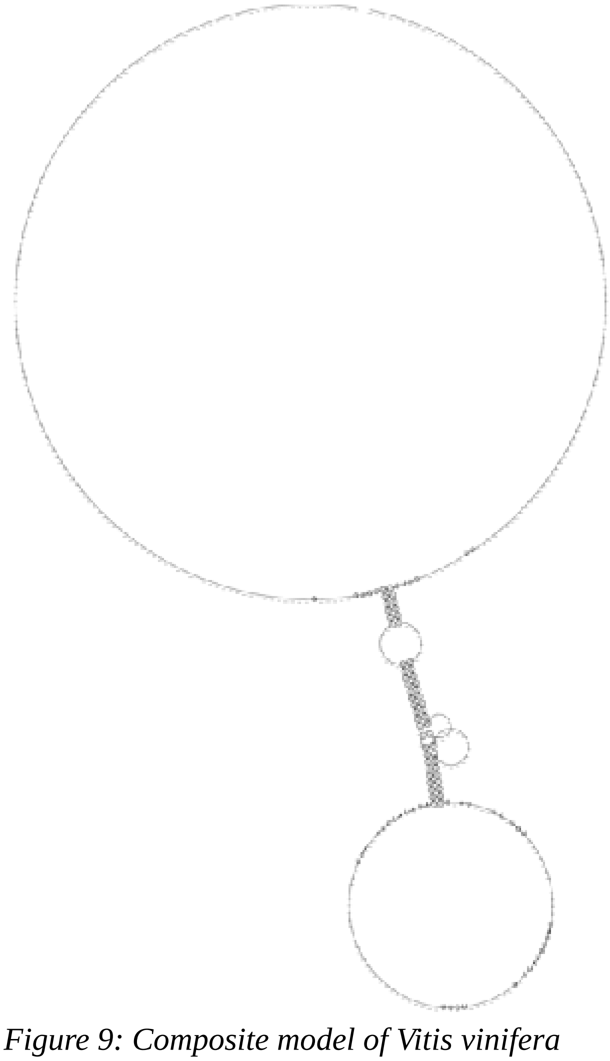


Figure 9: Composite model of Vitis vinifera miRNA stem-loop structures

application. It was feasible to perform an MSA of this size because there were so few (144) stem-loop sequences currently available for this organism. Following methods intentionally similar to those of HMMER, a CM was constructed from the alignment. This CM was used to search genomic sequences at arithmetic speeds for regions matching the composite structure and sequence (Figure 9) of the known miRNA stem-loop structures. The predicted folding of the sequences was calculated internally to INFERNAL using the Vienna package's 'mfold'.

A second model was drawn from a composite of three Rfam (RNA analog of Pfam) families mir-399, mir-395, and mir-172 (Figure 10) and used to score sequences alongside the *V. vinifera* derived model for comparison. These 'mir' families were chosen from among the hundreds of other miRNA families registered in Rfam for their significance in literature as being associated with plant specificity. These models were scored against the *V. vinifera* chromosomes

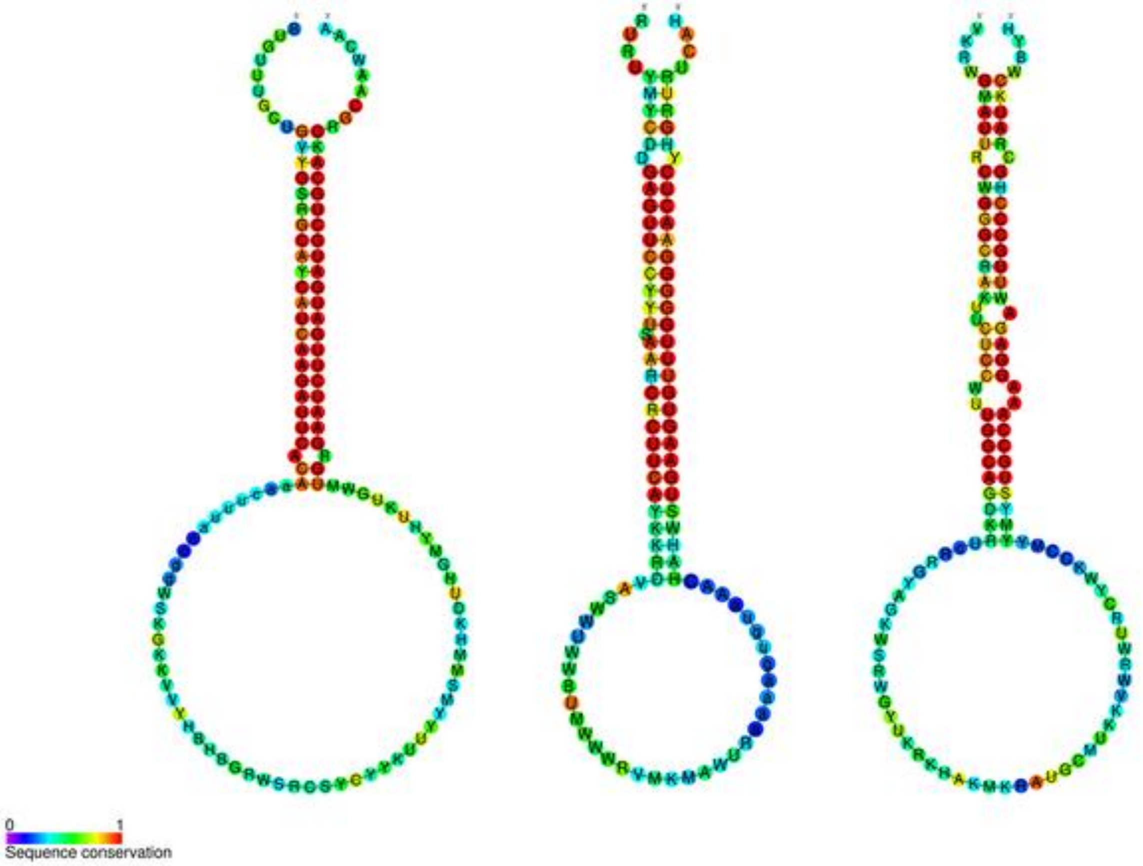


Figure 10: from left to right: the composite secondary structure of mir-172, mir-395, and mir-399

in parallel. The resulting hits from each chromosome were parsed via Perl, from their tabular output format into FASTA format and then concatenated into a single file.

Each putative miRNA was searched against the NCBI EST DB. Sequences with 50 or more nucleotide identities were removed from the final set. Non-coding RNA are usually lost in the EST identification process and therefore any sequences already identified as part of the EST DB have a high probability of being coding RNAs that happen to have hairpin-like structures.

Sequences were also eliminated from candidacy for having predicted secondary-structures with an MFE greater than -28kcal/mol as predicted by the Vienna package's 'mfold'. Other disqualifying characteristics included sequences with fewer than 45 bases, less than 33% and greater than 66% G/C content, more than 25 contiguous, ambiguous nucleotides, and duplicate identification, in which case, only one was kept.

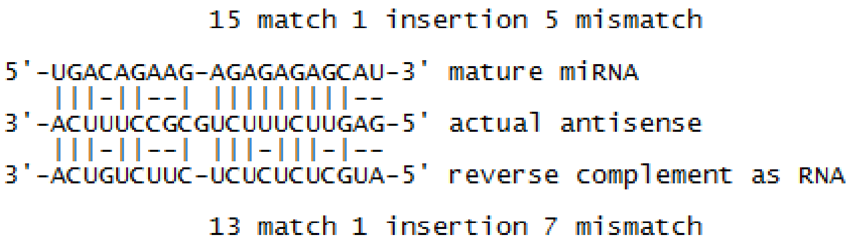
2.5 Failed Attempts and their setup

The study of non-coding RNAs is significantly less resolved than that of traditional genes and their products. Where solid solutions for their detection or examination have cropped up there exist deep but poorly connected wells of literature for each method. For this reason a significant bit more exploration was required to find a successful process by which miRNA genes could be elucidated. The most successful one has already been described; however the other attempts are no less valid from an exploratory point of view.

It has been strongly established that among plants, the primary sequences of pre-miRNA genes lack the evolutionary conservation seen among metaphyta. The information that is conserved is the folded, single-stranded structure of the domains processed by the miRISC. For this reason traditional sequence homology searching techniques would be insufficient in the prediction of novel V. *vinifera* miRNAs.

2.5.1 Homology

A search of the genome was performed using BLAST. This search retrieved mature-miRNA target sites from similarity to targets in other organisms. Genes responsible for coding products are still subject to the normal rules for genetic conservation. Naturally the mature portion (Figure 11) of the miRNA, which is responsible for binding this region, should be conserved as well. A Perl script was written to search 500nt up and downstream of each 'hit' region using BLAST to check for sequences having loose reverse-complementarities to the target. This script was tested on a manufactured data-set to ensure that it was capable of of locating the reverse-complement before being turned onto genomic sequences.



```
               15 match 1 insertion 5 mismatch

5'-UGACAGAAG-AGAGAGAGCAU-3' mature miRNA
   |||-||--| |||||||||--
3'-ACUUUCCGCGUCUUUCUUGAG-5' actual antisense
   |||-||--| |||-|||-|--
3'-ACUGUCUUC-UCUCUCUCGUA-5' reverse complement as RNA

               13 match 1 insertion 7 mismatch
```

Figure 11: Typical difficulty encountered in using BLAST to find the antisense of miRNA vvi-MIR156h

2.5.2 HMM

A second attempt used the pHMM approach taken from the previous effort in RGA domain scoring. A collection of related, immature miRNA stem-loop sequences was taken from the PMRD and miRBase repositories. A small number of representative sequences were taken from previously registered *V. vinifera* miRNAs and aligned using T-COFFEE. This alignment was then used to build a model with SAM. The rest of the sequences were then aligned to this model with 'align2model'. These aligned sequences were appended to the original model such that all of the downloaded plant miRNAs were part of the model used to score the *V. vinifera* genomic sequences.

A second model was constructed in the same way, but using letter-encoded (converted from the parenthetical coding used by 'mfold') structure data derived from the Vienna package's predictions. Because HMMER and SAM are not configured to translate coded structure data from the genomic sequences, they too were converted using 'mfold'. Each chromosome was segmented into 200nt pieces with a 10% overlapping the previous sequence and then translated by 'mfold'. The model was then used to score each of the 200-nucleotide segments. This method was loosely based on the methods described by Oulas et al. (2009).

Attempts to combine the sequence and structure models led to the discovery of alternate MSA formats that incorporated consensus secondary structures such as STOCKHOLM and SELEX, eventually leading to the use of the third method, INFERNAL.

3. Results

3.1 Sequence Cleaning and Verification

A thorough job of cleaning and preparing the sequences for analysis was done to to the RGA candidates before their use in this work.

All but five of IASMA's 943 (99.5%) conceptual translations were successfully matched to the Geneva PCR sequences using tBLASTx searching. The function of IASMA sequences was supported by linking to *in vitro* PCR evidence. Consequently, linkage to IASMA's conceptual translations indicated potential locations of the PCR-RGAs within the their respective genomic sequence.

While there were a few statistically relevant (e-value) hits among the conceptually translated RGA candidates from IASMA, none of them were considered to be vector. Vector sequences are not subject to degradation over evolutionary time and should have manifested as perfect matches if they were present. Likewise, no sequences bearing statistically significant resemblance to transposable or repeat elements were found.

While relating the 'Rose' RGAs to the IASMA sequences, 8 sequences out of the entire IASMA set exhibited possible frame-shifts or alternative splicing. Similarities that contained multiple fractionated hits that could have been interpreted as one of these features were few enough to examine manually. No definitive frame-shift was isolated. Regardless of the exact nature of these sequences, they were removed from further analysis so as to standardize workflow and increase confidence in further results.

Several sequences from the NYS-ARS germplasm library corresponded to the same conceptual protein and appeared as redundant sequences. This can be explained through the mechanism of alternative splicing, and was expected. Because the focus of this experiment was to verify the likelihood that each sequence was correctly translated and corresponded to a known RGA, these redundant sequences were eliminated.

3.2 Alignment of Sequences

The 1,741 sequences from the combined sets were aligned using the described methods. The best alignment from each of the two methods described was kept for clustering. [Supplemental File 1 and 2] The trimmed, empirically derived model's alignment was 1,893 nucleotides in length, compared to the trimmed Pfam-domain-model derived alignment at 4,658 nt.

3.2.1 Informative Site Trimming

The informative site trimming process excised 222,278 nucleotides (17%) from the purely empirical alignment. Removing these sites eliminated 193.5 billion time-consuming distance comparisons during the following clustering. The combined model eliminated nearly as many.

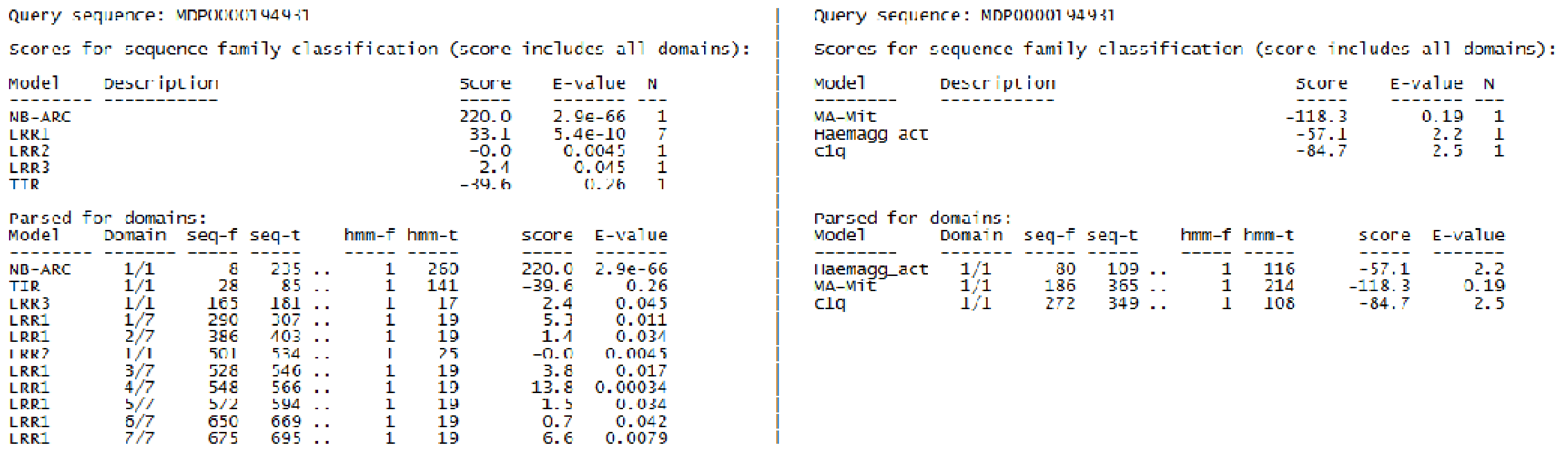
3.2.2 Transposon search of excised sequences

Of the excised sequences, only three met the statistical cutoff (e-value 0.01) for consideration as transposons or other repeatable elements. These sequences were only loosely matched; their similarity was probably coincidental and not investigated further.

3.2.3 HMMpfam Search Results

The initial 'hmmpfam' scores were so overwhelming that they prompted the construction of the negative control model. Almost every sequence in the IASMA set conformed strongly to the Pfam domain model, usually in 2 or more of the included domains. Due to their shorter length, the PCR-RGA sequences had fewer scorable regions causing proportionally lower score statistics than their IASMA counterparts. Scores of this level likely resulted from the IASMA researchers' use of an NB-ARC-domain model derived from Pfam alignments as a large part of their qualification criteria. The PCR-RGA primers were also designed to target the TIR and NBS domains. This still provides evidence for the relatedness of the sequences to one another, and also for the computational models' ability to isolate the same domains as the laboratory techniques. The highest scoring 70 domain-bearing sequences (blended equally for each domain) were parsed from the output file as representatives of the domains [Supplemental Files 3, 4, 5]. These sequences were collected for possible use in flagging clusters as strongly representative of a domain.

Table 4: Comparison of TIR-NBS-LRR model to negative control. Higher scores and smaller E-values indicate less likelihood of an arbitrary match.



```
Query sequence: MDP0000194937

Scores for sequence family classification (score includes all domains):

Model    Description                                   Score    E-value  N
-------- -----------                                   -----    ------- ---
NB-ARC                                                 220.0    2.9e-66   1
LRR1                                                    33.1    5.4e-10   7
LRR2                                                    -0.0     0.0045   1
LRR3                                                     2.4      0.045   1
TIR                                                    -39.6       0.26   1

Parsed for domains:
Model    Domain  seq-f seq-t    hmm-f hmm-t      score  E-value
-------- ------- ----- -----    ----- -----      -----  -------
NB-ARC     1/1       8   235 ..     1   260      220.0  2.9e-66
TIR        1/1      28    85 ..     1   141      -39.6     0.26
LRR3       1/1     165   181 ..     1    17        2.4    0.045
LRR1       1/7     290   307 ..     1    19        5.3    0.011
LRR1       2/7     386   403 ..     1    19        1.4    0.034
LRR2       1/1     501   534 ..     1    25       -0.0   0.0045
LRR1       3/7     528   546 ..     1    19        3.8    0.017
LRR1       4/7     548   566 ..     1    19       13.8  0.00034
LRR1       5/7     572   594 ..     1    19        1.5    0.034
LRR1       6/7     650   669 ..     1    19        0.7    0.042
LRR1       7/7     675   695 ..     1    19        6.6   0.0079
```

```
Query sequence: MDP0000194931

Scores for sequence family classification (score includes all domains):

Model       Description                                  Score    E-value  N
--------    -----------                                  -----    ------- ---
MA-Mit                                                  -118.3       0.19   1
Haemagg act                                              -57.1        2.2   1
C1q                                                      -84.7        2.5   1


Parsed for domains:
Model       Domain  seq-f seq-t    hmm-f hmm-t      score  E-value
--------    ------- ----- -----    ----- -----      -----  -------
Haemagg_act   1/1      80   109 ..     1   116      -57.1      2.2
MA-Mit        1/1     186   365 ..     1   214     -118.3     0.19
C1q           1/1     272   349 ..     1   108      -84.7      2.5
```

3.3 Heirarchial Clustering

The hierarchical clustering was performed using the Phylip package, according to the methods described. The 'Rose' set was compared with the R-gene candidates found in the genome of *Malus x domestica.* A tree of 75 clusters was generated based on the calculated distance matricies. [Supplemental File 6] 'Rose' RGAs are located in a few, but not most of the clusters in the cluster-tree. This result could be due to specificity in the design of the degenerate primers for some specific cluster of NBS-coding genes.

Preliminary analysis shows that some NBS-coding sequences from the 'Rose' set cluster together without any homologues in *Malus* x *domestica*. These may be R-genes for non-apple pathogens. If they include wild apple species they may also be R-genes lost during the selection of 'Golden Delicious'. These clusters may be new valuable sources of markers or resistance genes for apple breeding.

Many of the 75 clades consisted of only IASMA identified sequences. These clusters may be NBS-domain bearing sequences that correspond to other functions. They also may be R-genes that have yet to be identified in the organisms of the ‘Rose’ set.

A general feature of the cluster tree is that most clades are comprised of sequences from two or more species. This implies that the sequences were multiply representative of the same regions across Rosaceae, *i.e,* shared by the lineages through a common ancestor. One particularly interesting clade consists of only Fragaria and Rosa RGAs. This clade may be representative of a completely novel class of R-genes hitherto unseen in the rest of Rosaceae.

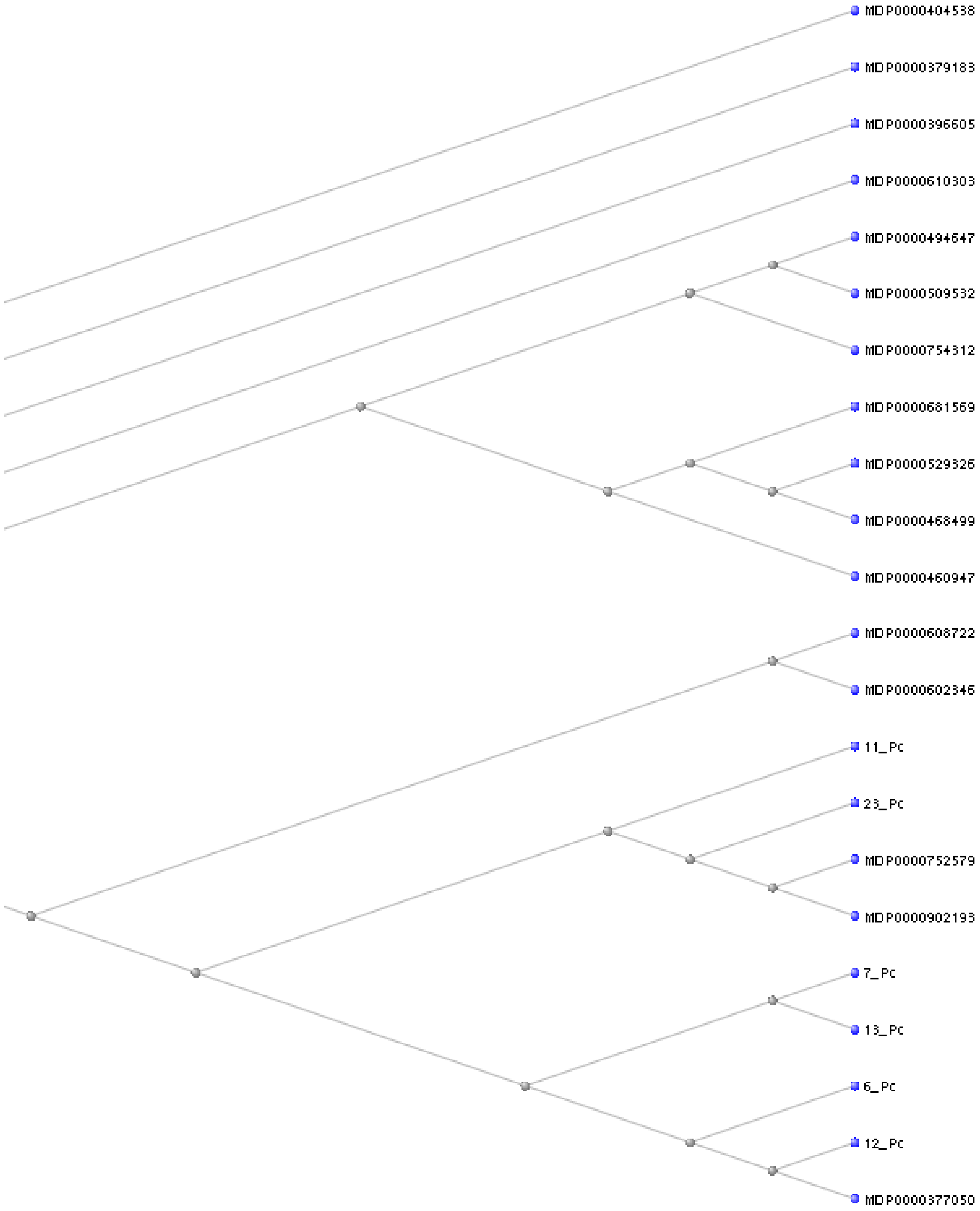


Figure 12: Clusters from a branch of the Pfam-model derived alignment. Some Malus x domestica 'Golden Delicious' (MDP#) sequences cluster more closely to Pyrus communis (#_Pc) than to other 'Golden Delicious' sequences.

3.4 miRNA Prediction

Of the multiple attempts at miRNA precursor gene prediction, only the covariance model-based approach produced meaningful results. By scoring the Rfam and *V. vinifera*-derived models against each refseq chromosome, 4,603 and 2,394 sequences were generated respectively. These sequences were pared down to 1617 (Suplemental File 7) and 1570 (Supplemental File 8) putative pre-miRNAs with the application of secondary filtering criteria. Of these 3187 combined sequences, 8 were identically shared between the two models and removed from the combined set to avoid redundancy.

Of the 144 sequences used to construct the *V. vinifera*-specific model, 64 (44%) were re-identified from genomic sequence between the two models. Of the 64, 21 of the input sequences were identified by both models, 28 were unique to the *V. vinifera* model and 15 were unique to the Rfam model. The 21 co-identified input sequences were not truly identical because of start/stop shifts due to variations in the model profiles.

A handful of the sequences used to build the *V. vinifera* model were detected by the initial model search but were disqualified by later metrics. In a cursory search, the lowest scoring input sequences still detectable by the models had a score of 20 and an e-value of 0.49, well below the threshold of statistical significance. This illustrates that even low-scoring sequences may have known positives among them.

Unfortunately, the web server responsible for hosting the results of Lazarri et al. (2009) was no longer functioning. No genomic location or sequence comparisons could be made between the two attempts because of this.

3.5 Associated Results

An associated manuscript by Baldo et al. (in prep.) describes the genomic relationships of the RGA-families. They report that previous phylogenies constructed from NBS genes distinguished between TIR-NBS-LRR(TNL) and non-TNL subfamilies. Further they assert that RGAs tend to occur in clusters, often mapping to major resistance genes or QTLs. These two subfamilies were also identified in the hierarchical clustering presented here.

Baldo et al. posit that the TIR-NBS clade can be divided into six subclades and the non-TNL clade can be subdivided into five subclades. Chromosomes were unevenly distributed with TIR and non-TIR clades, and subclades of RGAs. Furthermore, some subclades are more strongly represented in a given chromosome than others. Only two chromosomes contain NBS encoding genes from every subclade. Genes located in these clades reside proximately on the cluster-map [Supplementary File 9].

4. Discussion

4.1 Clustering of RGA domains

The hierarchical clustering successfully linked large quantities of known and unknown genes into clades. Multiple analyses have produced results provide sufficient evidence to conclude the relatedness of intra-cluster RGAs. Future research will reveal the exact function and class of patterns described by each cluster. These RGA clusters can help to guide and focus that research. The resistance profiles of the *Malus* cultivars can be used in association with the clustering of their RGA markers to draw inferences as to which diseases are denoted by which markers. For example, if three *Malus* species that are particularly cold-hardy sub-cluster distinctly within a group of other apple RGAs, it may indicate that a mutation within that sub-cluster is responsible.

As the Golden Delicious sequences are mapped to loci in the genome, it is expected that more R-genes will be discovered, and others confirmed. The RGA clusters presented here will likely silhouette the physical clusters that are characteristic of R-gene duplication events.

4.2 miRNA Prediction Comparison

A total of 3,187 putative pre-miRNA, stem-loop forming sequences were identified here. Of these sequences, 62 (2%) contained a mature miRNA already registered with miRBase. While at first this number may seem disappointing, it should be noted that these 62 miRNAs account for 38% (163 at the time of this writing) of all *V. vinifera* sequences in miRBase.

With such a large portion of existing *V. vinifera* miRNAs represented by only 2% of these putative sequences, questions are raised as to the validity of the remaining 98%. In comparison to the current 163 experimentally verified mature sequences, 3,187 appears to be a gross overshot. Held against the 5,778 predicted from the same genome by Lazarri et al., this seems to be a conservative number. If genome size were directly proportional to the number of genes (which it is not) then *A. thaliana's* 1,400+ miRNA would put forward the expectation of 5,800+ mature miRNAs in *V. vinifera*. Regardless, putative miRNAs should be verified through homology or isolation and sequencing in a laboratory.

One further observation of interest was the lack of overlap between stem-loops identified by the *V. vinifera*-derived model and the Rfam-derived model. Each one identified roughly the same number of sequences but only eight were the same. Both produced pronounced stem-loop-forming sequences, but almost independent sets of them. One possible explanation is that the Rfam model was less organism-specific and may have identified more ancestral sequences while the other was tailored to genus-specific duplications.

4.3 Covariance Models in Gene Prediction

Covariance Models have shown immense promise both here and in other *de novo* prediction attempts. The arithmetic speeds at which primary and secondary structures can be searched make CMs an invaluable tool. That so many secondary structures were correctly predicted from genomic sequence at arithmetic speeds makes them an invaluable tool for both protein and RNA structure-sensitive investigation. Hairpin stem-loop structures are a fairly common fold for an RNA transcript to take arbitrarily. Unfortunately, these structures can appear in a sequence without it being targeted by miRNA processing enzymes.

Model-based searching methods have tradeoffs to methods such as homology. Homology methods can only identify sequences close to those that are already known, placing an upper limit on the amount of sequences identifiable in plants at around 1,500. These methods have broader acceptance in the scientific community. On the other hand, model-based methods produce a greater volume of unconfirmed but putative miRNAs. Considering the range of systems and the number of protein-coding genes modulated by non-coding RNA, more sequences are likely to be elucidated.

Searching attempts based on sequence similarity are also limited in their sensitivity to stem-loop structures. The degeneracy of antisense sequences is beyond the least stringent settings for BLAST. Frequently an antisense strand will have an insertion and half as many mismatches as matches. Consequently, traditional sequence similarity searches require the antisense strand to identify an orthologous hairpin.

pHMMs are another well accepted method for gene prediction. The HMMER documentation states that HMMER is less effective on nucleotide data because of the smaller range of transitionable states. Traditional sequence-only pHMMs cannot account for the secondary structures that comprise most of the conserved data in pre-miRNAs. Although there is room for improvement in CMs, they represent a powerful tool for analysis when carefully constructed.

5. Conclusion

The described tools and workflow can be used to quickly and accurately identify and cluster genes of interest from genomic sequence. The RGA clusters identified are highly likely to correspond to R-gene products, and will likely form closely around R-gene loci in the physical map. INFERNAL has proven to be an effective method of predicting stem-loop structures that may be targeted by the miRISC. Laboratory verification of the putative miRNA genes should still be undertaken to verify their processing by relevant enzymes. A full characterization of disease resistance-associated genes and their regulatory regions needs to be undertaken. This requires more genomic sequence information to be made available, especially in clonally reproduced crops.

6. Supplemental Materials

6.1 Appendix A – Perl Modules Utilized

PrimarySeq.pm
Seq.pm
SeqIO.pm
BlastResult.pm
ResultI.pm
Fasta.pm
Result.pm
SearchIO.pm
GenericHit.pm
blast.pm
blastxml.pm
genericHit.pm

6.2 Appendix B – Index of Figures and Terms

Figures:

Works Cited

Jon A. Appel, Erick DeWolf, William W. Bockus, and Timothy Todd; “Preliminary 2009 Kansas Wheat Disease Loss Estimates“Kansas Cooperative Plant Disease Survey Report August 11, 2009

Angela Baldo, Giulia Malacarne , Michele Perazzolli, Laura Righetti, Aubrey Bailey, Alessandro Cestaro, Marco Moretto, Silvio Salvi, Robert Viola, Riccardo Velasco, Mickael Malnoy “Genome-wide identification of NBS-encoding resistance genes in Malus X domestica.” In prep (2010)

Aidan Budd; “Gibson Team course pages at EMBL. – Phylogenies and Alignments” Practical course given at Genetics Deparment, Cambridge University, in March 2007 http://www.embl.de/~seqanal/MSAcambridgeGenetics2007/PhylogeniesAndAlignments/PhylogeniesAndAlignments.html

Henrich Brunke, Min Chang, Crystel Stanford “Commodity Profile: Wines and Wine Grapes” Agricultural Issues Center University of California Davis; AgMRC 2003

Buddenhagen, I. W.;“Resistance and vulnerability of tropical crops in relation to their evolution and breeding” Ann. New York Acad. Sci. 1977 287:309-326

Chang-Zheng Chen, Ling Li, Harvey F. Lodish, David P. Bartel; “MicroRNAs Modulate Hematopoietic Lineage Differentiation” Science 2 January 2004: Vol. 303. no. 5654, pp. 83 - 86

Brody J DeYoung & Roger W Innes; “Plant NBS-LRR proteins in pathogen sensing and host defense” Nature Immunology 7, 1243 – 1249; 2006 doi:10.1038/ni1410

Yin T-M, DiFazio SP, Gunter LE, Jawdy SS, Boerjan W, Tuskan GA; “Genetic and physical mapping of Melampsora rust resistance genes in Populus and characterization of linkage disequilibrium and flanking genomic sequence.” New Phytol 2004 164: 95-105.

Carolyn Dimitri, Anne Effland, and Neilson Conklin; "The 20th Century Transformation of U.S. Agriculture and Farm Policy" Electronic Information Bulletin Number 3, June 2005

Brian D. Harfe, Michael T. McManus, Jennifer H. Mansfield, Eran Hornstein, and Clifford J. Tabin "The RNaseIII enzyme Dicer is required formorphogenesis but not patterning of thevertebrate limb" Proc Natl. Acad. Sci. USA. 2005 August 2; 102(31): 10898–10903.

Eddy, Sean; "HMMER User's Guide" Biological sequence analysis using profile hidden Markov models.” Version 2.1.1; December 1998.

Eddy SR.; “Profile hidden Markov models” Bioinformatics. 1998;14(9):755-63. C. Notredame, D. Higgins, J. Heringa; "T-Coffee: A novel method for multiple sequence alignments" Journal of Molecular Biol. 2000, 302, 205-217

Felsenstein, J.; “PHYLIP (Phylogeny Inference Package) version 3.6” 2005. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Griffiths-Jones S.;“The microRNA Registry.”; Nucleic Acids Res., 2004, 32, Database Issue, D109-D111

S. Griffiths-Jones, A. Bateman, M. Marshall, A. Khanna, S.R. Eddy; “Rfam: an RNA family database” Nucleic Acids Research 2003 31(1):p439-441.

Gruber, A. R. R.; Lorenz, R.; Bernhart, S. H. H.; Neuböck, R.; Hofacker, I. L. L. “The Vienna RNA Websuite” Nucleic Acids Res. 2008 July 1; 36(Web Server issue): W70 W74.

Hillig RC, Renault L, Vetter IR, Drell T 4th, Wittinghofer A, Becker J.; “The crystal structure of rna1p: a new fold for a GTPase-activating protein.” Mol. Cell 3(6): 781-791.

Hughey, R. and Krogh, A.; “Sam: SEQUENCE ALIGNMENT and MODELING SOFTWARE SYSTEM.” Technical Report 2009. UMI Order Number: UCSC-CRL-95-07., University of California at Santa Cruz.

Matthew W. Jones-Rhoades, David P. Bartel, Bonnie Bartel.; “MicroRNAs and Their Regulatory Roles in Plants.“ Annual Review of Plant Biology Volume 57, Page 19-53, 2006 doi: 10.1146

John B, Enright AJ, Aravin A, Tuschl T, Sander C, et al. “Human MicroRNA Targets.”PLoS Biol 2004 2(11): e363. doi:10.1371/journal.pbio.0020363

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J; “Repbase Update, a database of eukaryotic repetitive elements.” Cytogenet Genome Res 2005;110:462-467

Barbara Lazzari, Andrea Caprera, Alessandro Cestaro, Ivan Merelli, Marcello Del Corvo, Paolo Fontana, Luciano Milanesi, Riccardo Velasco, Alessandra Stella; “Ontology-oriented retrieval of putative microRNAs in Vitis vinifera via GrapeMiRNA: a web database of de novo predicted grape microRNAs” BMC Plant Biology 2009, 9:82 doi:10.1186/1471-2229-9-82

Martin Mokrejš, Tomáš Mašek, Václav Vopálenský, Petr Hlubucek, Philippe Delbos, andMartin Pospíšek; “IRESite—a tool for the examination of viral and cellular internal ribosome entry sites” Nucleic Acids Research 10/2009

McGinnis S., & Madden T.L.; "BLAST: at the core of a powerful and diverse set of sequence analysis tools."Nucleic Acids Res. 2004, 32:W20-W25.

E. P. Nawrocki, D. L. Kolbe, and S. R. Eddy; “Infernal 1.0: Inference of RNA alignments” , Bioinformatics 25:1335-1337 (2009).

Anastasis Oulas, Alexandra Boutla,Katerina Gkirtzou, Martin Reczko, Kriton Kalantidis, Panayiota Poirazi.“Prediction of novel microRNA genes in cancer-associated genomic regions—a combined computational and experimental approach.” Nucl. Acids Res. (2009) 37 (10): 3276-

Phillips, T.; “Small non-coding RNA and gene expression.” Nature Education 2008 1:1

NCBI/Quetier F., Genoscope - Centre National de Sequencage; “Vitis vinifera” Refseq Project ID: 33471

Vladimir Shulaev, Schuyler S. Korban, Bryon Sosinski, Albert G. Abbott; “Multiple Models for Rosaceae Genomics” Plant Physiology 147:985-1003 (2008)

Slovin, J.P., Davis, T., Rabinowicz, P., Shulaev, V.; “Fragaria vesca, a reference plant for the rosaceae family.” USDA Meeting Abstract. p. 44 2006.

Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, et al. “The Bioperl toolkit: Perl modules for the life sciences.” Genome Res 2002 , 12:1611-1618.

Gerben van Ooijen, Gabriele Mayr, Mobien M. A. Kasiem, Mario Albrecht, Ben J. C. Cornelissen, and Frank L. W. Takken; "Structure–function analysis of the NB-ARC domain of plant disease resistance proteins". J. Exp. Bot. 2008;59:1383-1397

HC Wang, J Dopazo, JM Carazo “Self-organizing tree growing network for classifying amino acids” Bioinformatics 1998, 14 (4): 376-377. doi: 10.1093

Dirk Werling, Oliver C. Jann, Victoria Offord, Elizabeth J. Glass and Tracey J. Coffey; “Variation matters: TLR structure and species-specific pathogen recognition” Trends in Immunology, Volume 30, Issue 3, 124-130, 10 February 2009

Ken R. Wilson, John Zandstra; ”Disease-Resistant Apple Cultivars” Agri-Food Canada; Ridgetown College/University of Guelph (April 1998); omafra.gov.on.ca/english/crops/facts/98-013.htm

Xu Y, Tao X, Shen B, Horng T, Medzhitov R, Manley JL, Tong L.; “Structural basis for signal transduction by the Toll/interleukin-1 receptor domains.” Nature. 2000 Nov 2;408(6808):111-5.

Zheng Zhang, Scott Schwartz, Lukas Wagner, and Webb Miller; "A greedy algorithm for aligning DNA sequences", J Comput Biol 2000; 7(1-2):203-14.

Zhenhai Zhang, Jingyin Yu, Daofeng Li, Zuyong Zhang, Fengxia Liu, Xin Zhou, Tao Wang, Yi Ling, and Zhen Su; “PMRD: plant microRNA database“ Nucleic Acids Research, 2010, Vol. 38, Database issue D806-D813