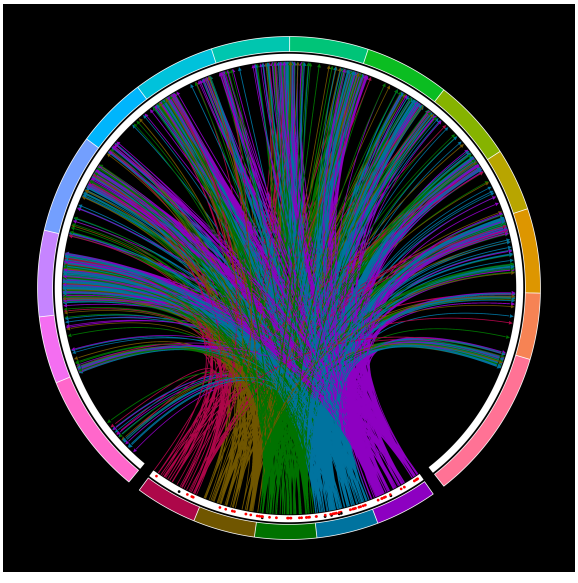
Acta Universitatis Agriculturae Sueciae

Doctoral Thesis No. 2021:14
Faculty of Natural Resources and Agricultural Sciences

Little strokes fell great oaks

– small RNA in potato and *Phytophthora infestans* interactions

Kristian Persson Hodén



Little strokes fell great oaks

– small RNA in potato and *Phytophthora infestans* interactions

Kristian Persson Hodén

Faculty of Natural Resources and Agricultural Sciences

Department of Plant Biology

Uppsala

DOCTORAL THESIS

Uppsala 2021

Acta Universitatis agriculturae Sueciae
2021:14

Cover: “Tree of death”. Illustration of exogenous targeting of *Phytophthora infestans* small RNA in potato (prior removal of false positives).
(photo: Kristian Persson Hodén)

ISSN 1652-6880
ISBN (print version) 978-91-7760-706-9
ISBN (electronic version) 978-91-7760-707-6
© 2021 Kristian Persson Hodén, Swedish University of Agricultural Sciences
Uppsala
Print: SLU Service/Repro, Uppsala 2021

Little strokes fell great oaks. small RNA in potato and *Phytophthora infestans* interactions

Abstract

Small RNAs (sRNAs) are non-coding RNAs of approximately 20-30 nucleotides in length. sRNAs bind to Argonaute proteins (AGOs) an integrated partner of the RISC complex. sRNAs act as templates for RISC to recognize complementary mRNA transcripts which can be cleaved by AGOs and thereby cause gene inactivity. In this study, 14 AGOs were discovered in potato. Phylogenetic analysis separated AGOs from *Solanaceae* and *Brassicaceae* families into three different clades, identifying AGO15 as *Solanaceae*-specific located in an evolutionary early branch in the AGO4 clade.

In previous work, PiAgo1 was categorized to associate with 20-22 nucleotide sRNA. In this study, potato was infected by a *pHAM34:PiAgo1-GFP* strain, followed by co-immunoprecipitation and sRNA sequencing. We found that the proportion of 5'U sRNA increased mostly among the nucleotides during infection. Based on sRNA target predictions mRNAs for resistance proteins were the dominating class. A potato alpha/beta hydrolase-type encoding gene (*StABH1*) was predicted to be cleavage by the single microRNA of *Phytophthora infestans* (miR8788). Cleavage was confirmed by 5' RACE and transient transcription Dual-LUC assays. Transgenic *StABH1* knockout potato lines were significant more diseased, demonstrating the importance of StABH1 in the defence to *P. infestans*.

To further investigate infection-induced sRNA events in the potato and *P. infestan* system, degradome sequencing was performed resulting in more than 30,000 targets, highlighting the need of an improved analytic strategy. Hence, the R package smartPARE was created with functionality to distinguish between true and the false cleavages. smartPARE was based on a deep learning convolutional neural network applying cyclical learning rate and Bayesian optimisation. smartPARE generated a cross-validated accuracy of 100% and identified 4,073 cleavages in potato and 702 in *P. infestans*. Several cascade events were seen either induced by Pi-sRNAs or St-sRNAs. In conclusion, by applying smartPARE to our 10 datasets, a more complex interaction than earlier demonstrated between the two organisms were found. An observation calling for further detailed analysis of precursor and target sites.

Keywords: Argonaute, Degradome, miRNA, small RNA, *Solanum tuberosum*, *Phytophthora infestans*

Author's address: Kristian Persson Hodén, Swedish University of Agricultural Sciences, Department of Plant Biology, Uppsala, Sweden

Liten tuva stjälper ofta stort lass. små RNA i interaktionen mellan potatis och *Phytophthora infestans*

Sammanfattning

små RNA är 20-30 nukleotider långa och är betydelsefulla då de reglerar geners uttryck. Dubbelsträngat RNA binder till Dicer proteinet som klyver RNAt till de korta längderna. I proteinkomplexet RISC associeras den ena RNA-strängen med komplementärt mRNA som bryts ner, det vill säga genen uttrycks inte. Argonauten (AGO) är integrerad i RISC och är ett av de aktiva proteinerna vid klyvningen av mRNAt. I denna studie påträffades 14 AGOs hos potatis (*Solanum tuberosum*, St). Fylogenetisk analys separerade AGOs från familjerna *Solanaceae* och *Brassicacea* i tre olika grupper. AGO15 kunde endast identifieras i *Solanaceae* och påvisades att tidigt i evolutionen ha förgrenats från AGO4-gruppen. Vi har utnyttjat en *pHAM34:PiAgo1-GFP* stam vid infektion av potatis och påföljande sRNA sekvensering för att närmare studera sRNA-processerna i *Phytophthora infestans* (*Pi*) såväl som i värdväxten potatis. Bland resultaten kan nämnas att andelen 5'U sRNA ökade mest jämfört med de andra nukleotiderna under infektion. *P. infestans* har bara ett mikroRNA - miR8788. Bioinformatisk analys föreslog klyvningar i flera möjliga mRNA under infektion. Ett membranprotein i potatis (StABH1) visades klyvas via en rad experimentella analyser. Ytterligare undersökningar gjordes av *StABH1* som visade att genen är viktig för försvaret mot *P. infestans.* För att ytterligare identifiera infektionsinducerade målsekvenser hos potatis och *P. infestans* utfördes en degradomsekvensering. Degradomsekvenseringen pekade ut fler än 30 000 möjliga klyvningar men visualisering av rådata påvisade att flertalet målsökningar var felaktiga. Således konstruerades smartPARE, ett R-paket med funktion att skilja på sanna och falska klyvningar. smartPARE baserades på en modell framtagen genom djupinlärning med konvolutionella neurala nätverk. För att ytterligare förbättra modellen tillämpades metoderna cykliskt inlärande och Bayesiansk optimering. smartPARE resulterade i en exakthet hos korsvaliderat data på 100%. Analys av de ca 30 000 möjliga klyvningarna identifierade 4073 sanna klyvningar i potatis samt 702 i *P. infestans.* Ett flertal sRNA från *P. infestans* visade sig målsöka mRNA i potatis. Några av dessa sRNA härstammade från effektorgener i *P. infestans*, vilket ger effektorerna en möjlig dubbelfunktion, det vill säga, de kan påverka potatis både som protein och sRNA.

Nyckelord: Argonaut, Degradom, mikroRNA, små RNA, *Phytophthora infestans*, *Solanum tuberosum*

Författarens adress: Kristian Persson Hodén, Sveriges lantbruksuniversitet, Institutionen för växtbiologi, Uppsala, Sweden

Dedication

To my family – The root to my life, my crutch to lean on and the flowers on my table

The creation of a thousand forests is in one acorn.

— Ralph Waldo Emerson

Contents

List of publications

This thesis is based on the work contained in the following papers, referred to by Roman numerals in the text:

I. Liao Z*, **Persson Hodén K***, Singh RK, Dixelius C. 2020. Genome-wide identification of Argonautes in Solanaceae with emphasis on potato. *Scientific Reports* **10**: 20577

II. Hu X*, **Persson Hodén K***, Liao Z, Åsman A, Dixelius C. *Phytophthora infestans* Ago1-associated miRNA promotes potato late blight disease. (manuscript)

III. **Persson Hodén K**, Hu X, Martinez-Arias G, Dixelius C smartPARE: an R package for efficient identification of true mRNA cleavage sites. (pending revision)

Papers I is reproduced with the permission of the publisher.

Additional publications

Tzelepis G*, **Persson Hodén K***, Fogelqvist J, Åsman A, Vetukuri RR, Dixelius C. 2020. Dominance of mating type A1 and indication of epigenetic effects during early stages of mating in *Phytophthora infestans*. *Frontiers in Microbiology*. **11**: 252

Liao Z, **Persson Hodén K,** Dixelius C. 2021. Small talk and large impact: the importance of small RNA molecules in the fight of plant diseases. In: RNAi for Plant Improvement and Protection. Chapter 9. Mezzetti B, Sweet JB, Burgos L (eds), CABI, UK

* These authors contributed equally

The contribution of Kristian Persson Hodén to the papers included in this thesis was as follows:

I. Did a majority of the bioinformatics. Wrote the paper together with co-authors.

II. Did all bioinformatics. Performed wet-lab work. Wrote the paper together with co-authors.

III. Participated in design of the project. Did all bioinformatics. Wrote the paper together with co-authors.

Abbreviations

Ago/AGO	Argonaute
AI	Artificial intelligence
ANN	Articial neural networks
CNN	Convolutional neural networks
CRN	CRinkling and Necrosis
DCL	DICER-LIKE
DE	Differential expression
DEGs	Differentially expressed genes
DL	Deep learning
FDR	Benjamini-Hochberg False Discovery Rate
Mb	Megabases
miRNA	MicroRNA
ML	Machine learning
mRNA	Messenger RNA
MSA	Multiple sequence alignment
ncRNA	Non-coding RNA
NGS	Next-generation sequencing
nt	Nucleotides
Pi	*Phytophthora infestans*
phasiRNA	Phased secondary small interfering RNA
RISC	RNA induced silencing complex
sRNA	Small RNA
sRNAseq	Small RNA sequencing
siRNA	Small interfering RNA
St	*Solanum tuberosum*
tasiRNA	Trans-acting small interfering RNA
wt	Wild type

1. Introduction

Bioinformatics and the study of small RNA (sRNA) are both two relatively young fields of research, bioinformatics popularized in the 1990's (Hogeweg, 2011); the same decade small interfering RNAs were discovered (Lee *et al.*, 1993; Fire *et al.*, 1998). The demand for implementation of bioinformatics to study sRNA interactions has increased since the first microRNA (miRNA) expression array (Krichevsky *et al.*, 2003; Gusev & Brackett, 2007), to date comprising analyses of the full spectrum of sRNAs of biological systems (Li *et al.*, 2020). Arduously, many of the methods available (laboratory or bioinformatics) are designed for model organisms and demand adaptation to be applied on crops such as potato or on a pathogen like *Phytophthora infestans* (Pi), having gene silencing pathways not yet fully characterized (Vetukuri *et al.*, 2011; Amar *et al.*, 2014).

When predicting the impact of sRNAs in biological systems, one challenge is the extensive scope of analyses involved. A bioinformatician needs to know how to analyse the origin of the sRNA, sRNA targeting of messenger RNA (mRNA) or other transcripts and effect of the sRNA on the transcript level. Furthermore, if the targeted transcript is of unknown character, analysis might comprise protein prediction studies and phylogenetic analyses in attempts to deduce possible protein functions. It is also important to know the genome content of the studied organism(s), so that corrections for possible biases from the sequencing or specific genomic features, e.g. repetitive sequences, can be accounted for. Furthermore, programming might be necessary to combine results from different large-scale analyses or to construct software that might take the analysis to the next level.

The intention of this thesis introduction is to give a brief overview of the areas a bioinformatician might encounter while working in the plant-pathogen sRNA field, with emphasis on potato-*Phytophtora infestans*. The introduction is hence written to an audience with basic understanding of bioinformatics. The content comprises analysis of: i) sRNA-field related sequencing methods ii) protein analytics methods, iii) different types of sRNA including development and benchmarking of selected software/tools/programs iv) software development v) genome information of potato and *P. infestans*. Due to the wide range of topics discussed, the following text is

biased towards recent findings from key papers. Many reviews are referred to which contain additional information in each research field.

1.1 Sequencing methods

Since the discovery of the three-dimensional structure of DNA (Watson & Crick, 1953), the development of nucleotide sequencing has been progressing, resulting in the first whole nucleic acid sequence (Holley *et al.*, 1965) and the development of Sanger's 'chain-termination' or dideoxy technique (Sanger *et al.*, 1977), known as Sanger sequencing. The second generation of DNA sequencing measuring pyrophosphate synthesis was developed during the 1980's and advanced into the first considerable commercial "next-generation sequencing" (NGS) technology (Nyrén, 1987; Heather & Chain, 2016), licensed to 454 Life Sciences. A method of sequencing of single DNA molecules attached to microspheres was developed in 1997 (Voelkerding *et al.*, 2009) and commercialized in 2006 as the Solexa (later acquired by Illumina) Genome Analyzer, the first sequencing platform of "short reads". The next year SOLiD, a short-read sequencing technology based on ligation, was released. SOLiD is characterized by interrogation of multiple octamer ligations (Buermans & den Dunnen, 2014), each octamer significantly improving read accuracy.

Third-generation DNA sequencing comprises real time sequencing of single DNA molecules, generating reads ranging up to several kilobases in length (Dumschott *et al.*, 2020). Commonly, third-generation sequencing technologies are Pacific Biosciences' (PacBio) single-molecule real-time (SMRT) technology and Oxford Nanopore Technologies' platforms MinION, GridION and PromethION. For a more comprehensive description of the history and theory behind the DNA sequencing methods, please see articles referred to in Voelkerding *et al.* (2009), Heather & Chain (2016) and Dumschott *et al.* (2020).

With the development of new sequencing methods, the tools for their analysis have developed, both in numbers and in complexity. In the following chapters, the focus will be on the description of the sequencing analysis methods of mRNA, sRNA and degradome sequencing, which are important for sRNA analysis and has been relevant for this thesis work.

1.1.1 Quality control and pre-processing of data

Independent of type of sequencing methods applied, it is important to have high quality in-put materials and that the data used for downstream analysis is optimized (Chen *et al.*, 2018b). Sequences might contain adapter contamination, base content biases, over-represented sequences, library preparation errors and sequencing inaccuracies. The quality of FASTQ data can be monitored using FASTQC

(Andrews, 2010), offering per-base and per-sequence quality profiling features. FASTQC can detect overrepresented sequences, including adapters that can be trimmed with an adaptor trimming tool, for instance Cutadapt (Martin, 2011). Cutadapt was designed to trim 454 sequencing data and sRNA data, however it also perform well on Illumina data. Cutadapt can trim or discard adapter-containing reads and discard reads of specified length. Trimmomatic was developed to handle Illumina sequence data, with main algorithmic innovations related to adapter identification and quality pruning (Bolger *et al.*, 2014). Trimmomatic applies two main quality filtering alternatives, sliding window and maximum information quality filtering. Moreover, Trimmomatic also perform: 5' and 3' end-trimming, read cropping, read filtering (dropping a read not meeting user specified criteria) and quality score conversion.

Several other tools are developed to quantify quality control, for instance fastp which is estimated to run two to five times faster than Cutadapt or Trimmomatic (Chen *et al.*, 2018b). MultiQC was developed to integrate metrics from several quality control tools (Ewels *et al.*, 2016), enabling quick identification of global trends and biases. Other pre-processing tools compatible with MultiQC are available, see the MultiQC documentation page https://multiqc.info/docs/#.

1.1.2 RNA sequencing analysis

RNA sequencing (RNAseq) was developed more than 10 years ago (Emrich *et al.*, 2007; Wang *et al.*, 2009a; Raplee *et al.*, 2019), comprising RNA extraction, ribosomal RNA depletion or mRNA enrichment, cDNA synthesis, adaptor ligation of sequencing libraries and sequencing (Stark *et al.*, 2019). RNAseq can be applied to a wide range of applications, commonly coupled with diverse types of biochemical assays (Conesa *et al.*, 2016). Analysis pipelines of the various experimental set-up sequencing data might hence vary. Common steps comprise design of experiment, quality control, sequence mapping, quantification, data visualization, differential gene expression analysis, alternative splicing analysis, functional analysis, gene fusion discovery and expression quantitative trait loci mapping. In this chapter, I will go through the present steps that have been useful for me during work related to sRNA (excluding alternative splicing analysis, gene fusion discovery and expression quantitative trait loci mapping from the steps mentioned above). The RNAseq steps are briefly discussed in this chapter (except for quality control, which was already discussed in chapter 1.1.1). For a more comprehensive summary, please see the review by Conesa *et al.* (2016) and papers referred to in the chapter.

When designing experiments, it is important to consider library type, sequencing depth and number of biological replicates sufficient for achieving the statistical power needed for detection of trends in the system of study (Conesa *et al.*, 2016). Secondly, the sequencing itself needs to be performed to avoid unnecessary biases,

for instance exclusion of ribosomal RNA, generally accounting for more than 90% of the total RNA of the cell.

RNAseq quantification tools comprise two major categories, alignment-based and alignment-free (Jin *et al.*, 2017). Alignment-based quantification comprises estimation of transcript-abundance based on reads mapping to a specific genome or transcriptome. Alignment-free quantification is defined by estimation of transcript abundance via pseudo-alignment in k-mer space. Pseudo-alignment is based on the de Bruijn Graph (Bray *et al.*, 2016), which is represented by all possible combinations of sequences by the defined character or symbol overlaps between sequences of symbols (Compeau *et al.*, 2011). A k-mer based method uses k-mer features of genome sequences, e.g. positions or frequencies to perform the assembly (Han & Cho, 2019).

A recent study was performed by Schaarschmidt *et al.* (2020) comparing the performance of the following six aligners from the different assembly categories: HISAT2 (Kim *et al.*, 2019), CLC Genomics Workbench (https://digitalinsights.qiagen.com), RSEM (Li & Dewey, 2011), kallisto (Bray *et al.*, 2016), STAR (Dobin *et al.*, 2013) and Salmon (Patro *et al.*, 2017). The study was performed on two polymorphic *Arabidopsis* accessions, showing highly similar results. In another recent study (Corchete *et al.*, 2020), HISAT2 and STAR outperformed TopHat2 (Kim *et al.*, 2013), RUM (Grant *et al.*, 2011) and Bowtie2 (Langmead & Salzberg, 2012) regarding unmapped reads. The gene expression quantification level was also compared between the aligners revealing that aligners using any of the following methods: raw reads, effective counts, estimated counts or coverage normalization, achieved poorest ranks.

In a study of human disease prediction (Tong *et al.*, 2020), the performance of 278 different pipelines was studied, featuring 13 sequence mapping methods, three methods for quantification and seven normalization methods. The assessment in regard to accuracy, precision, and reliability revealed that the higher scoring pipelines were more precise in predicting disease outcome. The most important pipeline factor for variation in performance was the normalization, where median normalization scored best compared to upper quartile (UQ), fragments per million mapped fragments (FPM), fragments per kilobase of gene length per million mapped fragments (FPKM), trimmed mean of M-values (TMM), relative log expression (RLE) and expression index (EIndex) in most combinations of other pipeline factors.

After read alignment, quantification of expression levels for transcripts of each sample is performed (Teng *et al.*, 2016). A benchmarking R/Bioconductor package (http://bioconductor.org/packages/rnaseqcomp) was compiled by Teng *et al*. (2016) in an attempt to (unbiasedly) evaluate RNA-seq quantification pipelines free of bias. They compared Flux Capacitor (Montgomery *et al.*, 2010), Cufflinks (Trapnell *et al.*, 2010), eXpress (Roberts & Pachter, 2013), kallisto, RSEM, Salmon and Sailfish

(Patro *et al.*, 2014) revealing that Flux Capacitor and eXpress clearly were underperforming and that RSEM slightly outperformed the other tools.

Differential gene expression (DGE) analysis is one of the most implemented features in RNAseq analysis (McDermaid *et al.*, 2019). The procedure involves evaluation of differential genes over dataset specific conditions, for example type of treatment or time after treatment. Seventeen DGE methods were reviewed by Corchete *et al.* (2020) under six experimental conditions and three FDR (Benjamini-Hochberg False Discovery Rate) levels (FDR < 0.05, 0.01, 0.001). It was found that limma trend (Ritchie *et al.*, 2015) was the most stable tool for analysis followed by limma voom, NOISeq FPKM (Tarazona *et al.*, 2011), baySeq (Hardcastle & Kelly, 2010) and some of the applications in edgeR (Robinson *et al.*, 2010). The performance of the tools depended to a large extent on the number of differentially expressed genes (DEG) included in the analysis and the FDR level compared.

Visualization of data enables the researcher to detect patterns and issues that they otherwise might oversee with traditional modelling (Rutter *et al.*, 2019), for instance negative binomial modelling or linear regression modelling (Law *et al.*, 2014). Visualization tools can detect designation problems of differentially expressed genes, normalization issues, and pipeline errors of usual character (Rutter *et al.*, 2019). A traditional data visualization tool for viewing of read levels is the Integrative Genomics Viewer (Thorvaldsdóttir *et al.*, 2013). Some DGE tools provide visualization functions e.g. edgeR and DESeq2. The R package "bigPint" can detect designation problems of differentially expressed genes, normalization issues, and pipeline errors of usual character (Rutter *et al.*, 2019). Several RNAseq visualization tools are listed in the article by Nazarie *et al.* (2019).

Functional analysis is often the last part of the RNAseq analysis (Conesa *et al.*, 2016) and comprises determination of functions or identification of molecular pathways of the DEGs. Functional profiling is dependent on available data of functional annotations for the specie(s) analysed. Several tools are available for different comparisons, some of which integrates different RNAseq biases e.g. GOseq (Young *et al.*, 2010) accounts for selection bias and SeqGSEA (Wang & Cairns, 2013) integrates splicing and determines enrichment. A protocol for pathway enrichment analysis was published by Reimand *et al.* (2019) applying g:Profiler (Reimand *et al.*, 2016), GSEA (Subramanian *et al.*, 2005), Cytoscape (Shannon *et al.*, 2003) and EnrichmentMap (Merico *et al.*, 2010). The protocol comprises creation of omics data-based gene list, the establishment of pathways that are statistically enriched, and feature interpretation from visualized data.

1.1.3 Small RNA sequencing analysis

Small RNA sequencing (sRNAseq) analysis resembles RNAseq analysis in many aspects. A sRNAseq pipeline could include quality control (discussed in chapter 1.1.1), normalization, differential expression (DE) analysis of annotated reads and visualization of the expression patterns (Beckers *et al.*, 2017). Often relative normalization is performed on sRNAseq data (Meyer *et al.*, 2010). This is performed by scaling to the dataset size and reported in reads per million (RPM) for each respective dataset. For this normalization to be valid sRNA sub-populations must have equal proportions across the different conditions being analysed, e.g. tissues or mutant backgrounds (Lutzmayer *et al.*, 2017). In a study by Qin *et al.* (2020), comparing the performance of nine normalization tools, trimmed mean of M-values scored best and the median and the upper quantile performed worst. With extensive and asymmetric level of DE none of the involved methods was better than moderately helpful. Because of the poor standard of the normalization methods, it might be worth to mention the sRNA spike-in oligonucleotides developed by Lutzmayer *et al.* (2017), that enable absolute normalization of sRNAseq data, even across independent experiments.

Several tools are developed to analyse DE of sRNA data between conditions of interest. One strategy is to map sRNA reads to a reference genome, count features and statistically evaluate differences between conditions (similar to DE analysis in RNAseq, (Anders *et al.*, 2013). Examples of such tools are sRNAtoolbox (Rueda *et al.*, 2015) and sRNAnalyzer (Wu *et al.*, 2017). Another approach is to perform DE of unique sequences instead of mapping the sRNA reads to genome features (Jeske *et al.*, 2019). This approach overcomes obstacles of short reads with multiple mappings in the genome or mapping to unannotated parts of the genome. One tool capable of this approach is DEUS.

Several RNAseq pipelines available can be configured to process different stages of sRNAseq (Beckers *et al.*, 2017). However, to date there are specific sRNAseq pipelines and platforms with different approaches available e.g. sRNAnalyzer (Wu *et al.*, 2017), UEA sRNA Workbench (Beckers *et al.*, 2017) and sRNAPipe (Pogorelcnik *et al.*, 2018).

A part of the sRNAseq analysis might also be the characterization of the sRNAs in the analysis. This is often in line with the classification of the precursor transcript of the sRNA. Morgado & Johannes (2019) list 60 different tools for plant sRNA categorization. Characterization of selected types of sRNA and other possible parts of the analysis, such as target prediction, will be discussed further in chapter 1.3.

1.1.4 Degradome sequencing analysis

A degradome is used to generate high-throughput information on transcript targeting by sRNAs. Most eukaryotic mRNA possess an m7G (a methylated version of guanosine) cap at the 5' end (Furuichi, 2014, 2015). When an mRNA is cleaved by an AGO, the 3' part of the cleaved mRNA will be uncapped and instead of the m7G cap possess a 5′-phosphate at the 5' end that can be ligated to a 5' adaptor (German *et al.*, 2008). As the 3' end of the truncated RNA is polyadenylated, reverse transcription using oligo(dT) with 3'-adapter sequence can be applied to construct a cDNA with adaptors in both ends. Further details about library preparation can be found in Zhai *et al.* (2014) and Sanz-Carbonell *et al.* (2020).

After sequencing and quality control of the reads, degradome analysis tools can be applied. In a paper by Thody *et al.* (2018), three popular degradome analysis tools CleaveLand4 (Brousse *et al.*, 2014), PAREsnip (Folkes *et al.*, 2012) and sPARTA (Kakrana *et al.*, 2014) were benchmarked against PAREsnip2. PAREsnip2 outperformed the other tools, both in the categories speed and resources. PAREsnip2 also detected more of already validated cleavages than the other tools, especially when the Fahlgren and Carrington (2010) targeting rules were applied. These rules, as opposed to Allen *et al.* (2005) targeting rules, tolerate a mismatch or G:U wobble pairs at positions 10 and 11.

1.1.5 DNA sequencing concepts

If a species already is sequenced and reference genome(s) already are provided (like when working with the potato-*P. infestans* system) working with sRNA interactions demand no extensive knowledge in the field of whole genome sequencing. However, to obtain an understanding for how the sequencing of the potato and the *P. infestans* genome was performed (presented in chapter 1.5) some concepts need to be clarified.

A genome assembly is the process of sorting reads of sequenced DNA in the correct order of the genome (Kalyanaraman, 2011). The genome can be assembled of reads into smaller contiguous overlapping parts called contigs (Batzoglou *et al.*, 2002). The contigs might then be sorted into super-contigs or scaffolds. N50 is a measurement implying that 50% of the nucleotides (nt) are located to contigs of this length or longer.

When performing DNA sequencing some difference apply to the techniques available to date. Maybe the most prominent difference is the sequencing lengths of the output. The length of short read technology reads is commonly only a few hundred bases. The invention of paired-end reads (each read being linked with another read some distance away (Risca & Greenleaf, 2015)) has improved the coverage of short reads techniques. Third generation sequencing methods, e.g. PacBio sequencing, produce reads longer than 10 kb at average (van Dijk *et al.*, 2018), revolutionizing the field. For example, third generation sequencing improves genome resolution, filling of gaps between contigs (Jain *et al.*, 2018).

Computational gene prediction is essential for automatic annotation of large genomes (Wang *et al.*, 2004). *Ab initio* gene prediction methods generally utilize statistical models, e.g. hidden Markov models, neural networks or Support Vector Machines (Wang *et al.*, 2019c; Scalzitti *et al.*, 2020), to combine signal and content sensors. Signal sensors refer to specific patterns and sites namely, promoter and terminator sequences, splicing sites, branch points or polyadenylation signals. Content sensors refer to species-specific patterns of codon usage allowing distinguishability between coding and surrounding non-coding sequences (Wang *et al.*, 2004), e.g. nucleotide composition or exon and intron lengths (Scalzitti *et al.*, 2020).

1.2 Protein analyses

When working with sRNA data analyses, investigation into protein analyses might be necessary to gain information about target genes in case there are no annotations available, or further information about the protein is desired. The following chapter is hence dealing with methods that have been useful during my PhD education.

1.2.1 Protein function and domain predictions

To acquire information about a protein of which the amino acid sequence is available there are two main concepts which can be followed (Eisenhaber, 2013). The first concept represents function heritage from a mutual predecessor gene, comprising homology searches such as BLAST or BLAST+ (Altschul *et al.*, 1994; Camacho *et al.*, 2009). The second concept comprises lexical analysis, interpretation of physical properties and sequence motif-function association (Eisenhaber, 2013). Segment-based analysis is the core of protein studies (Eisenhaber *et al.*, 2016). In general terms, proteins consist of two types of segments, namely globular domains and non-globular segments. Eisenhaber *et al.* (2013) present a strategy of six steps for protein sequence analysis comprising linguistic analysis, subcellular motifs, post-translational modification motifs, structural preference of nonglobular regions, families of globular domains, searches in sequence databases, sequence analytics and molecular function synthesis.

The InterPro database determines the protein families of sequences and predicts relevant sites and domains of the proteins (Mitchell *et al.*, 2019). InterPro integrates results from 14 different databases specialized in different areas of prediction. Among these 14 databases, several are found in the protein sequence analysis strategy by Eisenhaber *et al.* (2013) mentioned above. Uniprot is the largest resource of predicted sequence annotations in UniProt Knowledgebase (The UniProt Consortium, 2017).

1.2.2 Phylogenetic analysis

Phylogenetic analyses can be applied to determine evolutionary relationships between proteins, genes or species (Som, 2015). To compare sequence data in phylogenetic trees, a multiple sequence alignment (MSA) of the sequences must first be performed (Ashkenazy *et al.*, 2019). The accuracy of the phylogenetic tree hence relies on the correctness of the MSA (Kemena & Notredame, 2009). PASTA is an example of an MSA method that has scored well on larger protein datasets at reduced computational effort (Collins & Warnow, 2018).

Previous studies have concluded that trimming of unreliable MSA regions can enhance the accuracy of the phylogenetic analyses (Kück *et al.*, 2010; Di Franco *et al.*, 2019). However, other evaluations have indicated that filtering might exclude

phylogenetically informative sites (Chang *et al.*, 2014). The challenge with filtering algorithms is to remove phylogenetically uninformative sites (Tan *et al.*, 2015). Introduction of weights to the columns in the MSA has also been proposed to improve phylogenetic tree reconstruction (Chang *et al.*, 2014). Furthermore, a method based on concatenation of a large set of MSA into a single SuperMSA demonstrated to perform better than unfiltered MSA and single weight-based MSA (Ashkenazy *et al.*, 2019) .

Several different types of algorithms can be inferred to produce the phylogenetic tree. When comparing tree reconstruction accuracy of common phylogenetic algorithms, Bayesian and Maximum likelihood algorithms have been proposed to outperform maximum parsimony and neighbor joining algorithms (Ogden & Rosenberg, 2006). Beerli (2006) compared the inference of Bayesian and maximum-likelihood using the same sampling algorithm. In general, the Bayesian method performed better in accuracy and coverage, although for some comparisons both methods scored the same. In a review by Yang & Rannala (2012) it is claimed that Bayesian and Maximum likelihood inference belong to opposing statistical philosophies. Hence, a feature of one of the inferences might be considered an advantage or a limitation depending on the philosophy. More on strengths and weaknesses among major methods of phylogenetic analysis can be read about in this review.

Four of the most common fast maximum likelihood-based phylogenetic tools RAxML/ExaML (Stamatakis, 2014; Kozlov *et al.*, 2015), PhyML (Guindon *et al.*, 2010), IQ-TREE (Nguyen *et al.*, 2015), and FastTree (Price *et al.*, 2010) were compared applying 19 extensive phylogenomic datasets, comprising hundreds to thousands of genes (Zhou *et al.*, 2018). Slower approaches comprising ten tree searches per alignment exceeded faster approaches of one search per alignment applying PhyML, RAxML, or IQ-TREE. IQ-TREE scored the best-observed likelihoods for all concatenation-based species trees, with RAxML/ExaML scoring almost as well. Another investigation testing the inferences on bacterial genomes concluded that RAxML and IQ-TREE performed equally (Lees *et al.*, 2018).

1.3 Small non-coding RNA analyses

Non-coding RNAs (ncRNA) are defined as RNAs that are transcribed from DNA but not translated into proteins (Mattick & Makunin, 2006). There are several classes of ncRNA, however the most diverse range of ncRNAs belongs to the category of small ncRNAs (sRNAs). sRNAs are approximately 20–30 nucleotides long and include, *inter alia*, small interfering RNAs (siRNAs), miRNAs and phased secondary siRNAs (Borges & Martienssen, 2015). Just as their biogenesis differ (Fig. 1) (Brant & Budak, 2018) so do their functions, even if some features overlap. The differences between them make them separable and predictable using sequence-based algorithms.

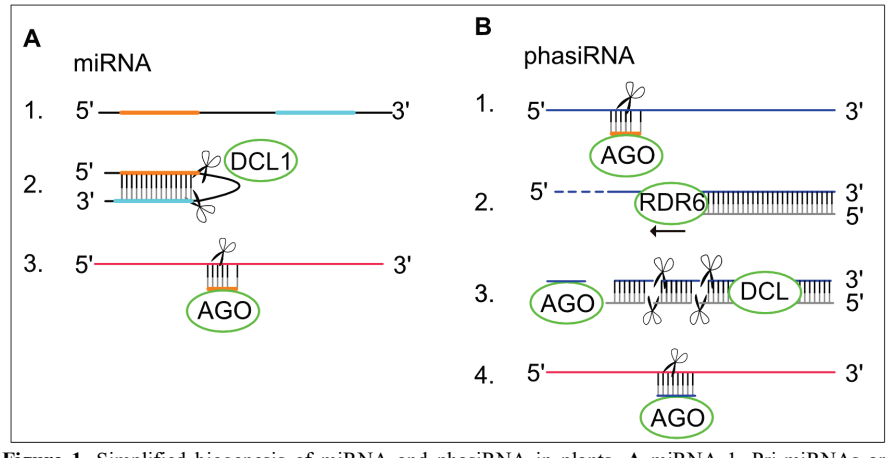


Figure 1. Simplified biogenesis of miRNA and phasiRNA in plants. **A** miRNA 1. Pri-miRNAs are transcribed from miRNA genes and processed into pre-miRNAs. 2. The pre-miRNAs are further processed by DICER-LIKE1 (DCL1) into an RNA duplex consisting of a 5p and a 3p miRNA. Any of the miRNAs might be incorporated into AGOs involved in the RISC complex and target single stranded transcripts through base-pairing. **B** phasiRNA biogenesis. 1. The phasiRNA biogenesis is activated by an AGO RISC complex, generally loaded with a 22 nt miRNA cleaving the phasiRNA precursor transcript. 2. The 5' fragment of the cleaved transcript is degraded whereas the 3' fragment is transformed into a double-stranded RNA by the RNA-DEPENDENT RNA POLYMERASE6 (RDR6). 3. The double-stranded RNA is cleaved by a DCL yielding phasiRNAs. 4. The phasiRNAs might be incorporated into AGOs in the RISC complex to target single stranded transcripts through base-pairing.

1.3.1 Argonautes

Argonaute proteins (AGOs) are involved in the RNA induced silencing complex (RISC) (Qi *et al.*, 2005), incorporating sRNA to target mRNA or long-ncRNA by sequence complementarity (Hamilton & Baulcombe, 1999; Chi *et al.*, 2009). Eukaryotic AGOs are characterized by the four domains: PAZ (Piwi-Argonaute-Zwille), MID, PIWI (P-element-induced wimpy testis), and N-terminal (Höck & Meister, 2008). The PAZ domain is anchoring the 3' end of the sRNA in eukaryotes (Ma *et al.*, 2004). The mid domain contains a pocket responsible for sorting of the 5′ end nucleotide of sRNA (Ma *et al.*, 2005), resulting in different abundances of sRNAs with different 5' end-nt in different plant AGOs (Mi *et al.*, 2008). Recognition of base-pairing at position 15 of miRNA duplexes is performed by a QF-V motif within the PIWI domain in *Arabidopsis* (Zhang *et al.*, 2014). The PIWI domain also contains a catalytic tetrad important for the slicing function of AGOs (Faehnle *et al.*, 2013; Arribas-Hernández *et al.*, 2016), together with the N-terminal domain. The N-terminal domain is the least characterized AGO domain (Miyoshi *et al.*, 2016). Base-pairing between the sRNA and the target strand was prevented in the bacterium *Thermus thermophilus* by the N-terminal domain (Wang *et al.*, 2009b). In humans, unwinding of the sRNA/target duplex is enabled by the N-terminal domain during RISC assembly (Kwak & Tomari, 2012).

The number of AGOs differ between species, for example the fission yeast *Schizosaccharomyces pombe* contains one AGO (Höck & Meister, 2008), *Arabidopsis* contain 10 AGOs (Morel *et al.*, 2002) and the nematode *Caenorhabditis elegans* 27 AGOs (Höck & Meister, 2008). Phylogenetic analyses have revealed that plant AGOs can be grouped into three major clades based on *Arabidopsis* AGO nomenclature: AGO1/5/10, AGO2/3/7, and AGO4/6/8/9 (Fang & Qi, 2016). Individual functions of plant AGOs can be found in the review by Zhang *et al.*, (2015).

1.3.2 microRNA prediction

miRNAs are small non-coding RNAs, with a usual length of 19-24 nucleotides (Ling *et al.*, 2013), transcribed from miRNA genes into primary miRNAs (pri-miRNAs, Xie *et al.*, 2005) by RNA polymerase II and coactivators (Fig. 1A) (Xie *et al.*, 2005; Wang *et al.*, 2013). Plant pri-miRNAs are processed into precursor miRNAs (pre-miRNAs) with a length of 49-900 nt (Bologna & Voinnet, 2014). The pre-miRNAs are further processed by RNase III enzyme DICER-LIKE1 (DCL1) into a RNA duplex consisting of mature miRNA and the complementary miRNA* (Kurihara & Watanabe, 2004). The mature miRNAs can be incorporated in AGOs involved in the RISC complex (Baumberger & Baulcombe, 2005), which recognize and target the mRNA via base-pairing between the mature miRNA and the mRNA. Generally, the miRNA* is regarded as a functionless "passenger strand", hence degraded (Eamens *et al.*, 2009). However, an accumulation of functions among miRNA* sequences has been discovered over the years (Devers *et al.*, 2011; Peng *et al.*, 2011; Aceto *et al.*, 2014). The association with miRNA* being functionless does no longer apply, why the miRNA/miRNA* nomenclature can be deceptive and preferably is written 5p/3p (referring to the positions in the pre-miRNA) (Desvignes *et al.*, 2015). Further details about plant miRNA biogenesis can be found in recent reviews (Wang *et al.*, 2019a; Gao *et al.*, 2020)

miRNAs can be identified using four different approaches (Mishra *et al.*, 2015): Conservation-based strategies, machine-learning strategies, high throughput techniques (including next-generation sequencing (NGS) strategies) and homology-based strategies.

Conservation-based strategies originate from the homology shown between miRNAs in different species (Mishra *et al.*, 2015). However, the levels of conservation differ between miRNAs through phylogenetic analyses. It has been demonstrated that the most plant miRNAs are inherited from ancestral embryophytes and spermatophytes (Taylor *et al.*, 2014). It was also proposed that more ancient miRNAs are not more conserved than younger miRNAs, which earlier was suggested. The early approach of conservation-based strategies was to predict miRNAs located to intergenic regions that were conserved between related species (Mishra *et al.*, 2015). Furthermore, the candidate regions should have predicted secondary structures folding into stem-loops. Along with increased identification of miRNA genes, homology-based search methods were developed, focusing on the similarity to known mature miRNA sequences. Genomic features of the matches to known mature miRNAs are extracted and aligned with their miRNA families. Tool-specific-criteria are then applied to construct the final list of the candidates (Artzi *et al.*, 2008; Lucas & Budak, 2012; Mishra *et al.*, 2015).

The machine-learning strategies utilize already confirmed miRNA hairpins as a positive training dataset and a negative training dataset with other hairpin structures, containing for example mRNAs, tRNAs and rRNAs. Properties distinguishing the

datasets are then generated by the machine learning tools (Mendes *et al.*, 2009). The current machine-learning strategies and future perspectives are discussed in the review by Schäfer & Ciaudo (2020).

High throughput techniques can be applied after miRNA sequencing with for example NGS technology (Motameny *et al.*, 2010). High throughput techniques are characterized by two steps; a filtering step and a modelling step (Mishra *et al.*, 2015). The filtering step comprises mapping of the sequenced reads to the genome of the sequenced species or mapping to sequences as similar to the genome as possible, e.g. a genome of a related species or a transcriptome of the query species. Unwanted reads, for instance those mapping to tRNA, are commonly discarded. Reads mapping to miRNA databases are identified as miRNA candidates. In the modelling the remaining uncharacterized reads are utilized to identify miRNA transcripts based on miRNA features. An example of a high throughput technical tool is ShortStack (Axtell, 2013). ShortStack is an example of attempts to characterize full pre-miRNA (hairpin) from the read data, (including mature miRNA and miRNA*). ShortStack annotations were hence proven highly specific with a very low number of false positives. A review handling trends in the field of the miRNA bioinformatics tools is presented by Chen *et al.* (2019).

1.3.3 phasiRNA detection

Phased secondary small interfering RNAs (phasiRNAs) are sRNA with a reported length of 18-29 nt in plants (Zhao *et al.*, 2020), with a characteristic phased configuration (Liu *et al.*, 2020). A pattern of regularly spaced siRNAs is generated by an endonucleolytic cleavage, which can be detected by mapping of siRNAs to the precursor transcript. The phasiRNA biogenesis is activated by the AGO1 RISC complex loaded with a 22 nt miRNA in *Arabidopsis* (Cuperus *et al.*, 2010) (Fig. 1B). The 22 nt miRNA is produced by DCL1 when an asymmetric bulge is present at the complementation of the miRNA-5p with the miRNA-3p (Chen *et al.*, 2010; Manavella *et al.*, 2012b). The 22 nt miRNA RISC complex cleavage generates a 5' fragment which is degraded by a 3′-5′ exonucleolytic complex, for instance the SKI2-3-8 complex (Branscheid *et al.*, 2015). The 3' fragment is transformed into a double-stranded RNA by the RNA-Dependent RNA Polymerase 6 (RDR6) in plants (Liu *et al.*, 2020), possibly recruited by AGO1-RISC or AGO7-RISC. It is speculated that a one-hit RISC directed target might recruit RDR6 to the 3' end of the transcript and a two-hit target might recruit RDR6 to the 5' end. Furthermore, the double-stranded RNA produced by RDR6 is cleaved by one of at least three different Dicer family members yielding different phasiRNA lengths.

PhasiRNA are derived from *PHAS* loci which can be mRNA of protein-coding genes or long noncoding RNAs in plants (Peragine *et al.*, 2004; Johnson *et al.*, 2009). *PHAS* loci within protein-coding genes is the largest group of known *PHAS* loci, including the large group of nucleotide binding leucine-rich repeat (NLR) resistance

genes (Zhai *et al.*, 2011; Fei *et al.*, 2015; Liu *et al.*, 2020). PhasiRNA from protein-coding genes are also involved in plant development (Marin *et al.*, 2010), plant parasitism (Shahid *et al.*, 2018) and seed germination (Guo *et al.*, 2018). One group of phasiRNAs from the long, non-coding RNAs are trans-acting and are hence called trans-acting siRNA (tasiRNA), originating from *TAS* loci (Fei *et al.*, 2013).

There are different approaches for computational detection of phasiRNA. Historically, one approach was to use ShortStack (Axtell, 2013). ShortStack detects phasiRNAs based on a user-defined size range and threshold for number of sRNAs. Firstly, sRNAs are clustered depending on alignment localization of the reads. sRNAs of the cluster within the user-defined size range and threshold are annotated as Dicer-derived. The primary sRNA size of the cluster is denoted DicerCall. A phase score is calculated for the phase size of the DicerCall to evaluate the level of phasing of the loci (Axtell, 2020). Additional functions are available in the ShortStack package such as miRNA annotation (as mentioned in chapter 1.3.2), estimation of RNA size distributions, repetitiveness, strandedness (specification of strand) and hairpin-association (Axtell, 2013).

PhaseTank was released with improved average sensitivity (77.9%) compared to ShortStack (26.9%) (Guo *et al.*, 2015). PhaseTank defines phasiRNA clusters as a region containing at least four phased reads with maximal separation of 84 nt. PhaseTank applies an algorithm that estimates the relative sRNA production (RSRP) of each cluster. It keeps the clusters with top 5% RSRP value. The abundances are then estimated for every potential 21-nt phasing of the clusters, after which the clusters are filtered and given a phased score. The different functions of PhaseTank comprise phasiRNA detection, phasiRNA cluster determination, triggered miRNA prediction and identification of phasiRNAs/tasiRNAs functional cascades.

unitas was published by Gebert *et al.* (2017), scoring equally well as PhaseTank on artificial (non-natural) datasets, consisting of only phasiRNA. However, unitas keeps the sensitivity with increased numbers of non-phased sequences when PhaseTank loses sensitivity. unitas calculates the probability to observe more than a defined number of phased reads within a sliding window (default = 1 kb) based on binomial distribution. If the Bonferroni corrected p value is below the significant level (default = 0.05) to reduce the false discovery rate unitas utilizes the following thresholds: i) There must at least be the same number of mapped phased reads as there is un-phased reads within a sliding window. ii) The phasiRNA has to be derived from at least five different loci. iii) At least 10% of the phased reads need to map to each genomic strand.

1.3.4 small RNA target prediction

Argonaute-incorporated sRNAs guide the RISC complex to mRNAs that base pair with the sRNA, degrading the mRNA through cleavage (Martinez *et al.*, 2002) or translational inhibition (Doench *et al.*, 2003). Because of the wide conservation in plants, miRNA is believed to be the most functionally important sRNA and hence most studied (Chen *et al.*, 2018a). Animal miRNA target sites are mainly located within the 3′-untranslated region of the targeted mRNA (Bartel, 2009). In plants, the most confirmed target sites are located in open reading frames (Liu *et al.*, 2014). Plant miRNA require a greater complementarity to the target mRNA than animal miRNAs. One region of the miRNA, called the "seed", is of greater importance to complement with the target site. In animals, this region is often detected from nucleotides number two to eight from the 5' end. However, a more centred seed region is also observed from position four to 15, counting from the 5' end of the sRNA (Shin *et al.*, 2010). In plants, complementation allows up to 5 mismatches (Liu *et al.*, 2014) and the seed region ranges from nucleotide two to 13, with base-pairing at position ten and eleven (close to the Argonaute-catalysed slicing site) being more critical for target repression (Schwab *et al.*, 2005). In the absence of base-pairing at sites nine to eleven, slicing is inhibited (Wang *et al.*, 2015), which occurs in natural target-mimic sites that inhibit regulation of miRNA targets (Franco-Zorrilla *et al.*, 2007).

Because of differences among animal and plant miRNA target sites, different tools are needed to predict the target sites. A comparison was performed of eleven plant miRNA prediction tools (Srivastava *et al.*, 2014), concluding that a combination of psRNATarget (Dai & Zhao, 2011) and Targetfinder (Fahlgren *et al.*, 2007) delivers high true positive coverage. Additionally, intersection of psRNATarget and Tapirhybrid (Bonnet *et al.*, 2010) provided highly precise predictions. The comparison conducted by Srivastava (2014) also concluded that many tools were optimized towards *Arabidopsis*. Other tools with different approaches have been developed, e.g. comTar (Chorostecki & Palatnik, 2014) which is focused on predicting targets conserved across species. Many recent tools involve the inclusion of a degradome, which is discussed more in detail in chapter 1.1.4.

1.4 Programming

Programming is defined as the action of writing a sequence of coded instructions to a computer to process data (Blackwell, 2002). A certain level of programming skills is desirable when working with bioinformatic analyses. Programming is useful to concatenate results from different tools into pipelines. It can also aid with preparation of data for visualization.

One can separate programming languages depending on their levels of abstraction into ranges from low-level to higher level languages (Kahanwal, 2013). In computer science, abstraction refers to the closeness of the language to the computer's own language (Machine language, comprised of binary digits, i.e. zeroes and ones). Popular bioinformatics programming languages are high-level language Python (Van Rossum & Drake, 2009) and R, a language and environment adopted for graphics and statistical computing (Oliphant, 2007; Bayón *et al.*, 2016; R Core Team, 2017). High-level programming languages are independent of any architecture, hence portable across various platforms (Watt, 2004). The first high-level programming language was Plankalkül, designed in 1945 (Bauer & Wössner, 1972; Rojas *et al.*, 2000). Freely available "Bio-toolkits" are compiled for several programming languages that make customization of pipelines or analyses easier (Mangalam, 2002). Examples of such are BioPython for Python or Bioconductor for R (Gentleman *et al.*, 2004).

1.4.1 Artificial intelligence

Artificial intelligence (AI) is the capability of a computer or computer-controlled robot to execute an assignment usually associated with rational creatures (Copeland, 2020). AI has problem solving, decision making, and pattern recognition capacities (Du *et al.*, 2020). Simplified, AI can be divided into two main categories: strong AI and weak AI. Strong AI refers to a programmed computer possessing a mind being able to understand and have cognitive states (Searle, 1980). Weak AI gives us powerful tools, empowering humans to develop and test hypotheses more accurately. Strong AI does not yet exist (Du *et al.*, 2020).

The expression "Artificial intelligence" was launched in 1956 (Brunette *et al.*, 2009). Early AI development resulted in two main approaches: the "top down" approach, comprising implementation of higher-level functions, and the "bottom up" approach, simulating neurons to create higher level functions. For further details in the history of AI, please see the review by Brunette *et al*. (2009).

1.4.2 Machine learning

Machine learning (ML) is a subset of AI which directs its attention to the ability of computers to learn from received data, while organizing processed information by manipulation of algorithms (Du *et al.*, 2020). Algorithmic models are trained on input data to recognize patterns and based on the patterns perform specific tasks. There are several types of ML algorithms available, however the two main methods are supervised (the most common) and unsupervised (Moore *et al.*, 2019). In supervised learning, the algorithmic model is provided with a training dataset consisting of labelled input examples and preferred output (Paeglis *et al.*, 2018). The aim of the algorithm is to create a function that links the input variable with the output variable so that the function can predict the output variable as correctly as possible for each new input variable. In unsupervised learning, no labels are provided to the model propelling it to create the input-output linking from unstructured data.

1.4.3 Deep learning

Deep learning (DL) refers to a specific type of ML that applies artificial neural networks or ANNs (Fig. 2A, Hogarty *et al.*, 2020). DL is capable of using infinite number of layers, each layer being able to learn distinct features of the training dataset. Different weighing for different stimuli allows adaptation to accomplish complex tasks (Hogarty *et al.*, 2018). Except for the multiple layer similarity between biological neural networks and ANNs, a resemblance of some ANNs lies in the Heaviside function. Similar to nerve firing, the Heaviside function returns an all-or-nothing response. Improvements in several fields of analysis, comprising image recognition, has been achieved applying DL algorithms (LeCun *et al.*, 2015). DL algorithms are sometimes referred to as “black boxes” (Lu *et al.*, 2018), because the ANN-generated features are of too high dimensionality for the human mind to interpret. Present DL algorithms comprises deep Boltzmann machines (Salakhutdinov & Hinton, 2012), long-term and short-term memory (LeCun *et al.*, 2015) deep kernel machines (Nikhitha *et al.*, 2020), deep recurrent neural networks (Pascanu *et al.*, 2013), and convolutional neural networks (CNN) (LeCun *et al.*, 2015). CNN have good advantages in the field of image classification predominantly because extraction of multi-level images features is possible in the CNN architecture (Mo *et al.*, 2019). Characteristic for CNN architecture is the transformation of simple features (e.g. lines and edges) of the input images into complex features (e.g. shapes and colours, Lu *et al.*, 2018). The transformation is done in what is referred to as hierarchical feature maps (Fig. 2B) built by multiple convolutional layers. Further, some layers can merge similar features to reduce dimensionality.

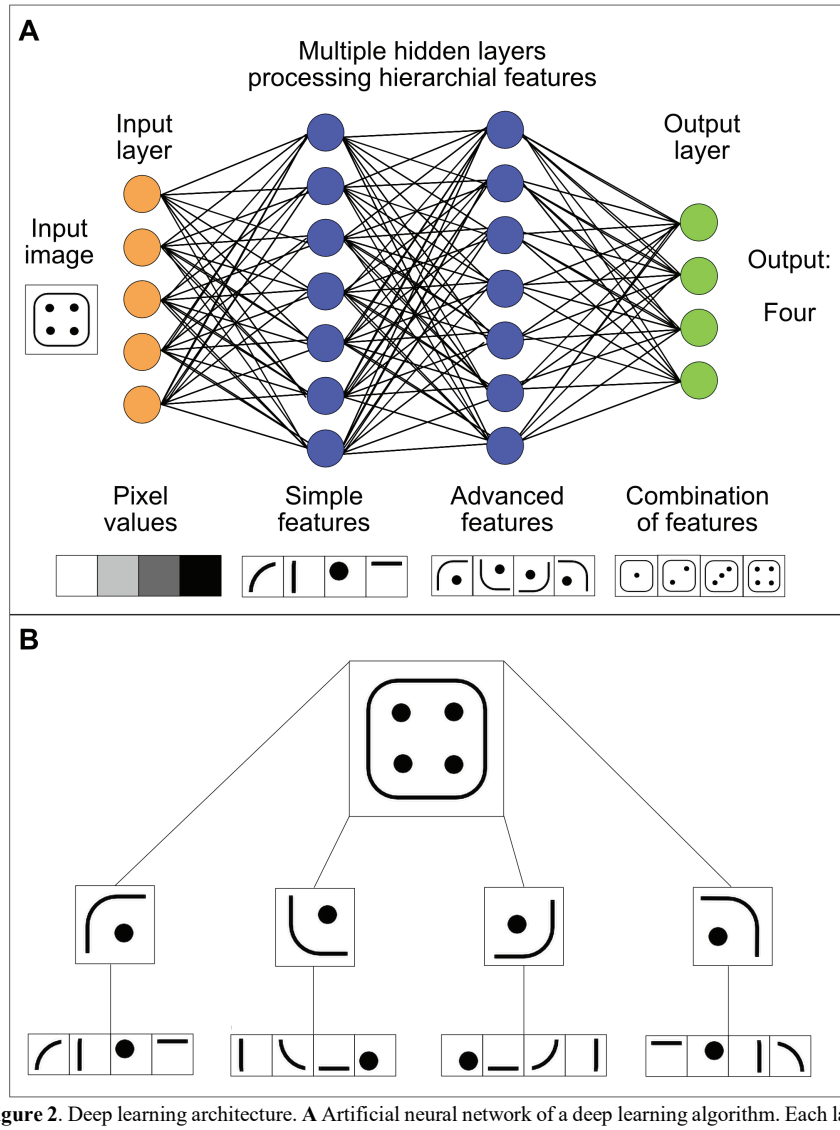


Figure 2. Deep learning architecture. **A** Artificial neural network of a deep learning algorithm. Each layer processing image features hierarchically. **B** Simplified hierarchical feature map of layers in a convolutional neural network. Each descendant layer consists of separated features of the ascendant layer. Concepts collected from Mukkulainen (1990) and Waldrop (2019).

1.5 Genomes

1.5.1 Potato

Potato (*Solanum tuberosum*, St) belongs to the family Solanaceae. Potato is, with a worldwide production of 368 million tons, the fourth most produced crop in the world (2018, http://faostat.fao.org/), being the most produced non-grain food crop. Potato origins from South America, where 151 wild species have been discovered (2020, http://cipotato.org). The wild species contain extremely large genetic diversity (Machida-Hirano, 2015). However, breeding with too few parental lines is proposed to have resulted in a small genetic variation in most present cultivars of the world (Salimi *et al.*, 2016; Wang *et al.*, 2019b). Cultivated potato is generally autotetraploid (2n = 4x = 48) and heterozygous (Manrique-Carpintero *et al.*, 2018), however some varieties are also diploid (2n = 2x = 24), triploid (2n = 3x = 36), or pentaploid (2n = 5x = 60) (Machida-Hirano, 2015).

The assembly of the first potato genome was released ten years ago (The potato genome sequencing initiative, 2011), comprising 86% (723 megabases, Mb) of the estimated full genome (844 Mb). The assembly consisted of 39,031 protein coding genes and was constructed from a homozygous double-monoploid potato clone. The assembly comprised 66,254 super-scaffolds. However, the 443 largest super-scaffolds were larger than 349 kb, together corresponding for 90% of the assembly.

A reference chromosome-scale genome (v4.03) was later constructed for potato using 951 of the super-scaffolds (Sharma *et al.*, 2013), comprising 674 Mb (~93%) of the 723 Mb assembly and ~96% (37,482) of the predicted genes (Fig. 3). 674 Mb corresponds to ~80% of the estimated 844 Mb genome. The v4.03 reference genome was assembled using *in silico* anchoring approaches with physical and genetic maps from a diploid potato genotype and tomato, resulting in a sorting of the super-scaffolds into 12 chromosomal “pseudomolecules” and one pseudomolecule consisting of unanchored super-scaffolds. A study of monoploid and double monoploid clones expanded the genome with an additional pseudomolecule of unanchored super-scaffolds (Hardigan *et al.*, 2016, v4.04). Copy number variations were discovered, affecting 219.8 Mb (30.2%) of the genome. With almost 30% of the genes partially duplicated or deleted, this study revealed a heterogeneous nature of the potato genome. Recently, an updated version of the genome was released (Pham *et al.*, 2020, v6.1), based on Oxford Nanopore Technologies long reads coupled with proximity-by-ligation scaffolding (Hi-C). The potato genome v6.1 assembly comprises 741.6 Mb (87.8%) of the 844 Mb genome, 731.2 Mb anchored to the 12 chromosomes. A 99% reduction in the number of contigs and an increase in N50 scaffold size by 44 times resulted in that 741.5 Mb of the assembly was non-gapped and the discovery of 32,917 high-confidence protein-coding genes.

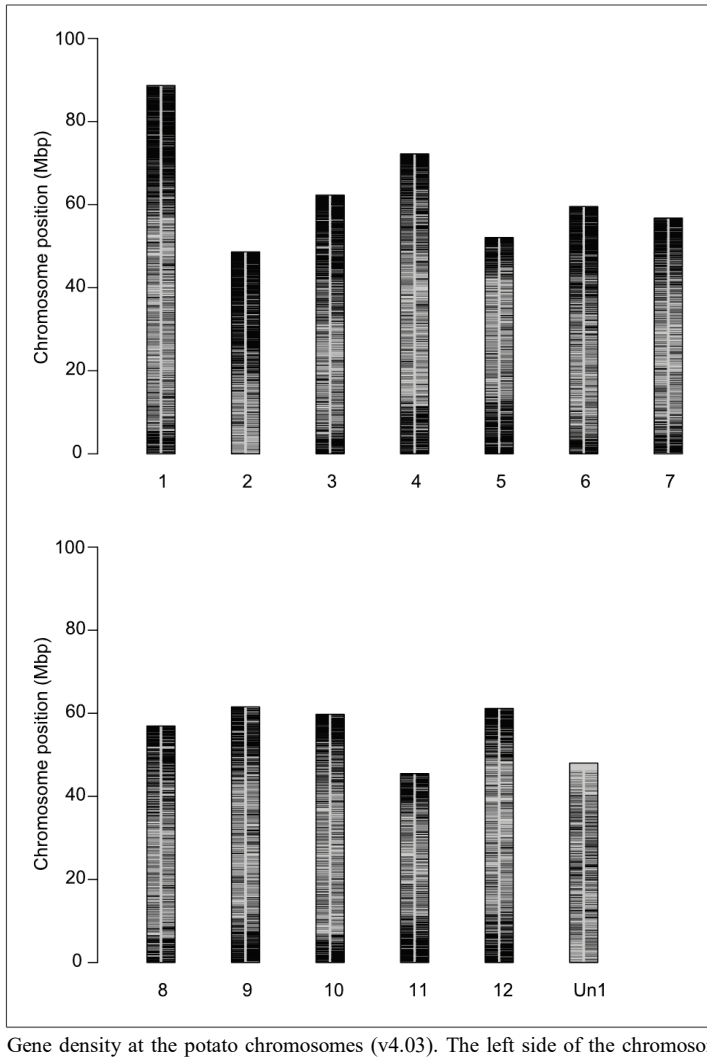


Figure 3. Gene density at the potato chromosomes (v4.03). The left side of the chromosomes depicts genes at the 5' strand and the right side depicts genes at the 3' strand. Chromosome positions for each gene are extracted from the Potato Genome Sequencing Consortium (PGSC) General Feature Format (GFF) file (v4.03) (The potato genome sequencing consortium, 2011; Sharma *et al.*, 2013).

1.5.2 *Phytophthora infestans*

Phytophthora infestans is an oomycete in the Peronosporaceae family and causes the potato and tomato late blight. The potato losses were estimated to M€4800 in the world, representing 15% of the total value of the grown potato (Haverkort *et al.*, 2008). *P. infestans* overcomes host-based resistance and fungicides effectively (Leesutthiphonchai *et al.*, 2018). Hence, improving defence against *P. infestans* at minimal resource expenses is of great importance. To improve control strategies against *P. infestans*, developing deeper understanding of the genetic complexity of the pathogen is necessary.

The *P. infestans* genome sequencing of strain T30-4 revealed a size of ~240 Mb (Haas *et al.*, 2009), still the largest *Phytophthora* genome sequenced (Vetukuri *et al.*, 2018). Repetitive DNA accounted for 74% of the genome sequence. In course of comparisons with other *Phytophthora* species *ab initio* and expressed sequence tag homology, 17,797 genes were identified. Simultaneously, 563 RXLR and 196 Crinklier (CRN for crinkling and necrosis) effectors were identified, although later updated to 557 RXLR and 129 CRNs (Cano *et al.*, 2019). Effectors are described in a review by Sharpee & Dean (2016) as pathogen secreted molecules that change plant processes promoting host colonization. The *P. infestans* genome comprised 4,921 scaffolds based on 18,288 contigs. Strain T30-4 is an F1 progeny of 80029 and 88133 strains (Lee *et al.*, 2001). The majority of dominant asexual *P. infestans* strains are found triploid (Knaus *et al.*, 2016; Li *et al.*, 2017; Tzelepis *et al.*, 2020). However, sexually reproducing *P. infestans* strains so far studied are diploid. *P. infestans* has two mating types, referred to as A1 and A2, both necessary for sexual reproduction and formation of oospores (Drenth *et al.*, 1994).

Oxford Nanopore and Illumina Nextseq were applied to produce two improved genome sequences of *P. infestans* (Lee *et al.*, 2020). The strains used to create the genomes originated from the Republic of Korea with different mating types (KR_1_A1 and KR_2_A2). The number of contigs was reduced to 1,510 and 3,344 in the A1 and the A2 strain, respectively. The A1 genome (201 Mb) was detected shorter than the A2 version (231 Mb). Both genome versions contained almost as much repeat-sequence as the T30-4 genome, approximately 72% each. More genes were discovered in these genomes, 20,172 in KR_1_A1 and 23,771 in KR_2_A2. A prediction of effectors was performed using EffectR and SignalP5, accounting for 433 and 310 RXLRs as well as 40 and 50 CRNs in each genome. When applying the same methods to the T30-4 genome, 306 RXLRs and 54 CRNs were discovered. A decrease of spanned gaps was also detected in the recent *P. infestans* genomes, from 38,410,029 in T30-4 to 0 in KR_1_A1 and 1700 in KR_2_A2.

2. Aims of the study

Computational analyses can support RNA biology research in several ways. Reliable predictions and analyses of biological patterns can aid the scientist in developing hypotheses and deducing possible mechanisms, which can be followed up *in vivo*. Development of better performing analysis tools will hence imply more reliable *in silico* results, leading to less time and costs spent searching for true knowledge *in vivo*. The long-term aim of this project was to deepen the knowledge on plant immunity and in this context elucidate the role of small RNA-associated activities in an important crop system. In this case, potato and the late blight pathogen *P. infestans*.

The aims of my PhD education are the following:

- Clarify numbers of Argonautes in potato and their evolution
- Examine events in potato when infected by an Ago1-GFP tagged *P. infestans* strain
- Examine events in potato during *P. infestans* infection with emphasis on potato AGO1
- Develop strategies to decipher sRNA cleavages based on degradome sequencing data during interaction between potato and *P. infestans*

3. Results and Discussion

3.1 Insights into Argonautes in the Solanaceae family (Paper I)

The Solanaceae family comprises more than 3000 species (Gebhardt, 2016) and contains species human civilization treat as crops, ornamentals and drugs. To this study we mined for AGOs present in databases at that time and focused the investigation into certain species more closely related to potato, which was our main focus. However, also Brassicaceae species were involved in the study including *Arabidopsis*. Plant Argonautes (AGOs) vary in both function and number between species (Fang & Qi, 2016). Most plant AGO functions are studied in *Arabidopsis* and conclusions could hence be made from comparing differences and similarities of additional species in Solanaceae and Brassicaceae.

3.1.1 Solanaceous *AGO* gene evolutionary events

To infer *AGO* gene evolutionary events, a Solanaceae *AGO* gene family tree was reconciled with a species tree constructed in the NCBI taxonomy browser. For replication confidence, the procedure was repeated three times using the different outgroups *Arabidopsis*, *Erythranthe guttata* and *Vitis vinifera*. The analysis revealed six gene duplication events before the diversification of *Petunia* and the remaining Solanaceae species. After the *Petunia* split, four duplications and two losses were detected before the speciation processes of *Nicotiana* and *Solanum* lineages. Among the species analysed the number of *AGOs* varied between ten *AGOs* in *N. obtusifolia* and 17 in *N. tabacum*.

3.1.2 Potato contains 14 AGOs

Mining for potato *AGO* sequences resulted in 14 unique *AGO* homologs (Fig. 4), all containing PAZ and PIWI domains. Orthologs of *Arabidopsis AGO1, AGO2, AGO3, AGO4, AGO5, AGO6, AGO7* and *AGO10* were present among the potato AGOs. *AGO1, AGO2* and *AGO10* were present as two orthologs each and three orthologs were detected for *AGO4*. The *AGO* phylogenetic clades comprise the *AGO1* clade (*AGO1/5/10*, the *AGO2* clade (*AGO2/3/7*), and the *AGO4* clade (*AGO4/6/8/9*, Fang & Qi, 2016). The *AGO1* and *AGO4* clades consisted of five potato homologs each and the *AGO2* clade of four.

Alignment of the potato *AGO* genes to the potato chromosomes revealed that *AGO2a*, *AGO2b*, and *AGO3* were located close to each other on chromosome 2. The fact that AGO2a, AGO2b, and AGO3 are closely related homologs, amongst others seen in the phylogenies of the paper and Fig. 4, indicated that they have been exposed to gene duplication.

Two-hundred three Solanaceae *AGO* homologs and 99 Brassicaceae homologs were used to reconstruct a maximum likelihood phylogeny, revealing that *AGO10* occurred in an ancestor before the divergence of Solanaceae and Brassicaceae. The duplication of the *AGO1* gene likely occurred after the split between the plant families. Solanaceae *AGO4* diverged into two groups, one partitioning with the Brassicaceae *AGO8/AGO9* subclade.

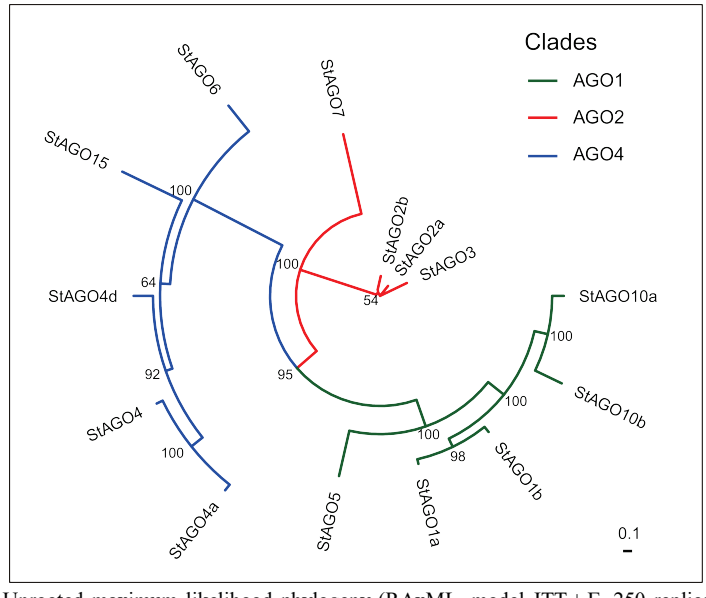


Figure 4. Unrooted maximum likelihood phylogeny (RAxML, model JTT + Γ, 250 replicates) of the Argonaute (AGO) family in potato. Branches are coloured according to clade identity. Bootstrap values are indicated at the branch forks. Bar = number of substitutions per site.

3.1.3 Solanaceae AGO15 diverged from the AGO4 clade

In the Solanaceae and Brassicaceae family tree, Solanaceae AGO15 was detected unique for the family, diverging early in evolution from the AGO4 clade. For further analysis the potato and rice AGOs were compared, revealing that no other AGOs from potato or rice clustered with StAGO15. Alignment between amino acids surrounding the proposed catalytic tetrads of Solanaceae AGO15 and AGO1 variants displayed divergence. StAGO15 catalytic tetrad indicated the motif G-E-Q-R instead of D-E-D-H/D. The residues of the nucleotide specificity loop (NSL), responsible for 5'specificity regulation in *Arabidopsis,* differed between the Solanaceae AGO1 and AGO15 sequences. Electrophoretic mobility shift assays (EMSA) were run to test for sRNA affinity of StAGO1a, StAGO10a, StAGO10b and StAGO15 (Fig. 5). The four StAGOs have basic isoelectric points (pI), (StAGO15 pI = 9.48, StAGO1a pI = 9.46, StAGO10a pI = 9.28 and StAGO10b pI = 9.24) making them positively charged. The positive charge interfered with their migration under the gel running conditions, making the experiments inconclusive (unpublished).

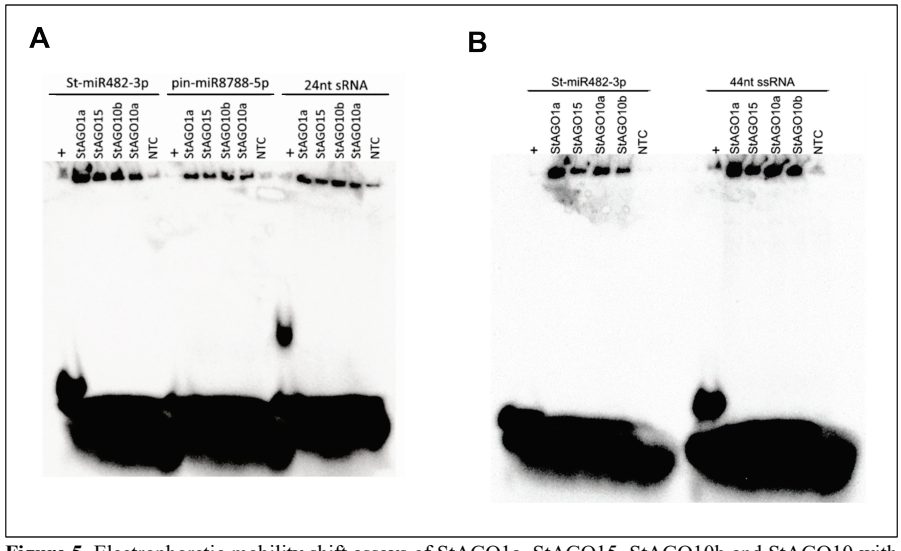


Figure 5. Electrophoretic mobility shift assays of StAGO1a, StAGO15, StAGO10b and StAGO10 with **A** ^{32}P-labeled St-miR482-3p (left), pin-miR8788-5p (middle) and 24nt sRNA (right) and **B** ^{32}P-labelled St-miR482-3p (left), and a 44nt single-stranded non-sRNA control (44nt ssRNA, right). + = ^{32}P-labeled sRNAs alone. NTC = non-template control of *in vitro* translation. (Photo: Zhen Liao).

3.1.4 Differential expression of RNA interference involved genes in potato during infection (Unpublished)

The complex responsible for miRNA processing consists in plants of DCL1, and cofactors Hyponastic Leaves 1 (HYL1) and the zinc finger protein Serrate (Song *et al.*, 2007; Dong *et al.*, 2008; Yang *et al.*, 2010; Manavella *et al.*, 2012a). A DE analysis of six *DCL* genes (excluding a potential pseudogene) discovered by Esposito *et al.* (2018) was performed on *P. infestans* and H_2O inoculated leaves. The DE analysis revealed a slight increase (<2 fold) in *DCL1*, *DCL2a*, *DCL3* and *DCL4* and a slight decrease (<2 fold) in *DCL2b* and *DCL2d* (Fig. 6A). The analysis included genes from the International Tomato Annotation Group (ITAG) annotation of the potato genome (Sato *et al.*, 2012). In the PGSC annotation (v4.03) of the potato genome, also transcripts PGSC0003DMT400019213, PGSC0003DMT400020650, PGSC0003DMT400020673, PGSC0003DMT400053499 and PGSC0003DMT400001805 were predicted to encode DCLs by the PANTHER tool (Thomas *et al.*, 2003). The corresponding proteins lack characteristic domains such as DExD-helicase, helicase-C, Duf283, PAZ, RNaseIII and double stranded RNA-binding (dsRB) domains (Margis *et al.*, 2006) when applying the Pfam domain prediction tool (El-Gebali *et al.*, 2019).

Only one HYL1 could be detected in potato (StHYL1, Fig. 6B), however *StHYL1* only was slightly up-regulated (log2 fold change ~ 0.4) upon infection. The slight

difference in *HYL1* expression indicates that the total production of miRNAs might be relatively steady between uninfected and infection state of potato.

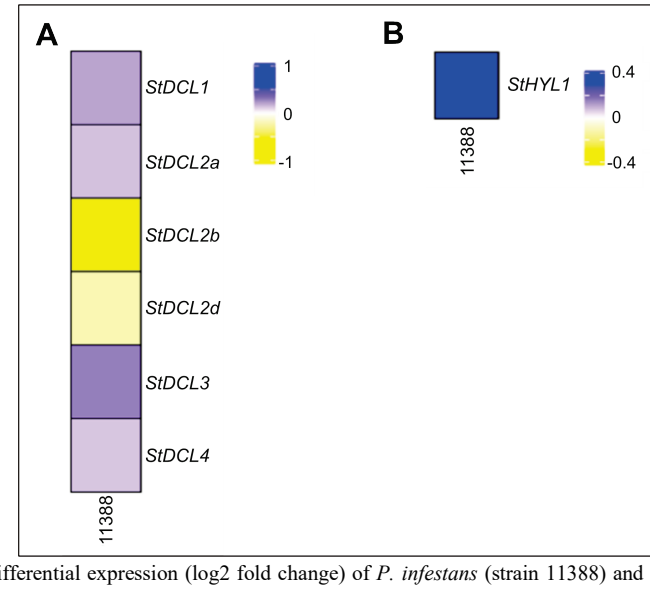


Figure 6. Differential expression (log2 fold change) of *P. infestans* (strain 11388) and H_2O inoculated potato (cv. Sarpo Mira), 5dpi. **A** ITAG DCL homologs in potato. StDCL1 (PGSC0003DMT400029301) and StDCL2a (PGSC0003DMT400042918) are also present in the PGSC annotation (v4.03) of the genome. **B** The only detected StHYL1 in potato. Colours of the heatmaps are related to the different log2 fold change ranges next to the heatmaps.

RDR homologs synthesize the complementary strand of a single-stranded RNA to generate double-stranded sRNA precursors (Polydore & Axtell, 2018). The six RDRs so far detected in plants are divided into two subclades: RDRα (RDR1, RDR2 and RDR6) and RDRγ (RDR3, RDR4 and RDR5) (Zong *et al.*, 2009). RDR1 is resistance related, induced upon virus infection in cucumber (Leibman *et al.*, 2018). RDR2 is involved in the heterochromatic siRNA pathway in *Arabidopsis* (Matzke et al., 2009). RDR6 is involved in production of phasiRNAs (Howell *et al.*, 2007), and involved in virus defence in *Arabidopsis* and *N. benthamiana* (Li *et al.*, 2014). The functions of *Arabidopsis* RDR3, RDR4 and RDR5 are so far unknown (Leuschen & Downing, 2020), although investigations have concluded that their catalytic domains share an atypical DFDGD amino acid motif (Willmann *et al.*, 2011).

A phylogenetic tree on RDRs in potato was reconstructed with *Arabidopsis* homologs (Kapoor *et al.*, 2008), tomato (*Solanum lycopersicum*, Sl) homologs (Bai *et al.*, 2012), already predicted potato RDR homologs in the ITAG potato genome version (Esposito *et al.*, 2018), and BLAST-mined homologs of the PGSC annotation (Fig. 7A). Based on homology in the phylogenetic tree and similarity percentage, the PGSC potato homologs were characterized. StRDR3b was more

similar (69%) to SlRDR3b than itagRDR3f (65%), why it was assigned the 3b extension. StRDR3g was 81% similar to itagRDR3d and itagRDR3e and hence assigned its own extension.

Potato RNAseq data surprisingly showed strong down-regulation (log2 fold change ~ -4) of both StRDR1 homologs upon *P. infestans* infection (Fig. 7B), indicating that although involved in virus defence RDR1 might not be involved in *P. infestans* defence. Instead, StRDR3b was most up-regulated upon infection, possibly involved in the defence against *P. infestans*. Redundancy among RNA interference-involved proteins has been summarised in *Arabidopsis* by Vazquez (2006), and might also explain potential similar functions between cucumber RDR1 and StRDR3b, just activated by different stresses. Potato has two RDR6 homologs, where StRDR6a was up-regulated and StRDR6b down-regulated upon infection, indicating that StRDR6a might be involved in the infection triggered biogenesis of phasiRNA.

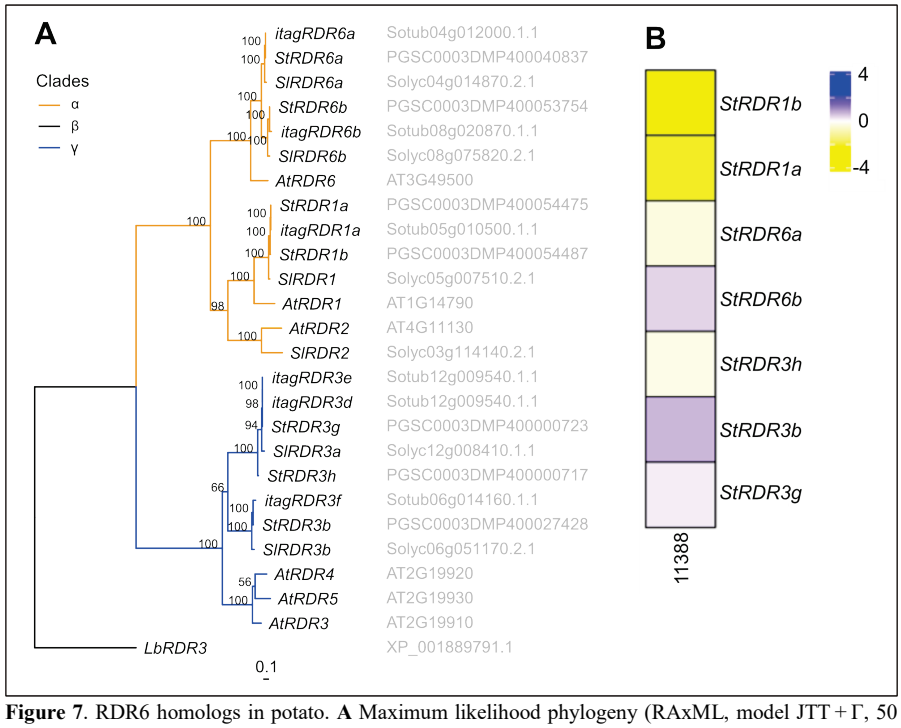


Figure 7. RDR6 homologs in potato. **A** Maximum likelihood phylogeny (RAxML, model JTT + Γ, 50 replicates) of the RDRs in *A. thaliana* (At), *S. lycopersicum* (Sl), *S. tuberosum* (ITAG denoted itagRDRx and PGSC denoted StRDRx). Outgroup = *Laccaria bicolor* RDR3 (clade β). Branches are coloured according to clade identity. Bootstrap values are indicated at the branch forks. Bar = number of substitutions per site. Accession numbers for each branch are annotated in column to the right of the tree. **B** Differential expression of *P. infestans* (strain 11388) and H_2O inoculated potato (cv. Sarpo Mira), 5dpi.

3.2 Analysis of *Phytophthora infestans* Ago1-associated sRNAs during infection (Paper II)

Co-immunoprecipitation and sRNAseq analysis of the Agos in *P. infestans* was performed by Åsman *et al.* (2016), resulting in the discovery that *P. infestans* miRNA and effector derived sRNAs were associated to PiAgo1. Hence, it was of interest to investigate in the association of sRNAs to PiAgo1 during infection. The material consisted of leaves infected with a *P. infestans* transformant expressing PiAgo1-GFP or a control expressing GFP and mycelia from the same transformants. sRNA co-immunoprecipitation was performed applying a GFP trap, after which libraries were prepared and sequenced. The data was quality controlled and separated into six datasets depending on what genome each sRNA read aligned to.

3.2.1 Increase in 5'U sRNAs upon infection

In the study performed by Åsman *et al.* (2016) a 5'C preference was detected among the PiAgo1-associating sRNA, which was also confirmed in the samples of this study. During infection, an increase was detected among the 5'C nt, however the proportional increase of 5'U was larger. Target predictions of the sRNA was performed applying psRNATarget (Dai *et al.*, 2018), revealing that the majority of all predicted targeting sRNAs had a 5'U preference. Further investigation revealed that only about 16% of the sRNA with 5'U in the infected samples were also present in the mycelia samples, indicating that there might be a mechanism altering the 5'U preference or production upon infection, potentially to invade potato with the 5'U sRNA. It has earlier been discovered that sRNA can act as effectors (Weiberg et al., 2013). The majority of the sRNAs with 5'U were derived from intergenic regions.

3.2.2 Resistance protein transcripts is one major sRNA target

Most of the *St*-sRNAs associated with PiAgo1 were derived from intergenic regions (64%). A similar pattern could be distinguished among the 33 *St*-miRNAs in the dataset. However, none were significantly enriched compared to the control sample (p-value < 0.05, log2 fold change > 2). Among the target predictions from the sRNAs derived from both genomes, groups of kinases, transferases, resistance proteins and transporters contained most predictions. Four predicted miRNAs discovered by homology-based detection had 21 potential target sites with most potential targets in resistance genes, transcription factors, kinases and synthases. Resistance proteins recognizes pathogen effector proteins and trigger one part of the innate plant immune system, reviewed by Han (2019). In total 638 mRNAs coding for resistance proteins were predicted as targets.

3.2.3 *P. infestans* miR8788 induces cleavage of *StABH1* mRNA

Among the *Pi*-sRNAs with predicted targets in potato, miR8788-3p was detected to target an alpha/beta hydrolase-type encoding gene (*StABH1*). The cleavage of *StABH1* was confirmed by 5'RACE upon infection, where no cleavage was detected in control samples of leaves inoculated with H_2O. The cleavage was also confirmed with a dual-luciferase reporter system, applied on Agro-infiltrated *Nicotiana benthamiana* leaf materials. PhasiRNAs were predicted applying PhaseTank (Guo *et al.*, 2015), and none of the discovered phasiRNAs were predicted to target *StABH1*.

3.2.4 PITG_10391 is presumably a pseudo-gene

miR8788 is located in *PITG_10391,* a predicted gene of unknown function. The gene could not be detected at cDNA levels. cDNA surrounding miR8788 was identified. The discovered transcript overlapped with an intron so it could not belong to the proposed *PITG_10391* gene. The transcript was hence concluded to origin from either pri- or pre-miR8788. Metazoan pre-miRNAs are ~70 nt long, however plant miRNAs can measure up to 900 nt (Bologna & Voinnet, 2014). In mirBASE the longest pre-miRNA measure over 2000 nt (Fig. 8). The cDNA transcript surrounding miR8788 measured 403 nt with no poly adenosine tail detected at 3' end of the transcript. A polyadenylation tail would have indicated an origin from a pri-miRNA so the detected transcript presumably is the pre-miR8788.

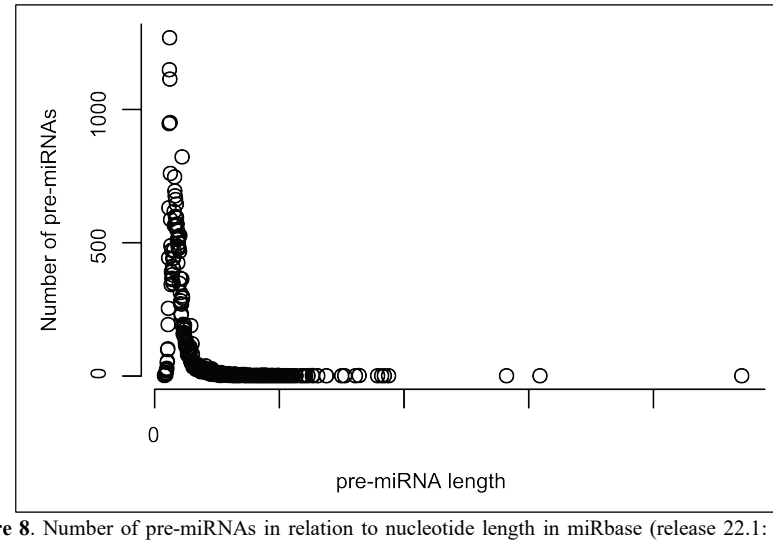


Figure 8. Number of pre-miRNAs in relation to nucleotide length in miRbase (release 22.1: October 2018, Kozomara *et al.*, 2019).

3.2.5 Tonoplast localization of *StABH1*

An *StABH1-GFP* construct was generated and Agro-infiltrated in *N. benthamiana*. Confocal microscopy located StABH1-GFP to the tonoplast, a membrane segregating the vacuole and the cytoplasm. In line with this discovery, domain prediction of StABH1 revealed a transmembrane domain upstream the ABH domain.

StABH1 is conserved in other potato cultivars and orthologs of the gene are observed in other plants. When analysing the orthologs of *StABH1*, a paralog was discovered in potato. Through phylogenetic analysis, *StABH1* and the paralog were located to two different clades among the *Solanaceae* species, indicating that the gene was derived from the same sequence in an early *Solanaceae* ancestor.

3.2.6 *StABH1* is vital for potato defence

Transgenic potato lines were produced, over expressing the earlier mentioned construct StABH1-GFP. StABH1-GFP transformants infected with *P. infestans* displayed smaller lesions 5 days past inoculation than control plants treated correspondingly. Transcript levels of *StABH1* were 10-fold lower than StABH1-GFP transformants inoculated with H_2O. In addition, transgenic potato lines were produced, expressing an artificial miRNA (*StamiRNA*) silencing the *StABH1* transcript. Three days post inoculation, *P. infestans* lesions covered the whole leaves of the *StamiRNA* lines. DNA content was significantly higher in the *StamiRNA* lines than in control plants. Although *StABH1* transcript level already was reduced due to the silencing by the artificial miRNA, *StABH1* was further reduced by infection. The further reduction in *StABH1*-levels was probably caused by miR8788-3p. To inhibit miR8788 in *P. infestans*, six miRNA target mimic candidate strains were constructed. Two strains showed reduced expression of miR8788-3p, enhancing *StABH1* levels upon infection.

3.3 Degradome analysis reveals infection induced targets (Paper III)

The extensive number of sRNA target predictions in Paper II motivated more in-depth methods to confirm the targets. Degradome sequencing is in simplicity a high-throughput modification of the 5′-rapid amplification of cDNA ends (5'RACE) (German *et al.*, 2008); a standard method for sRNA cleavage confirmation in plants (Llave *et al.*, 2002; Wang *et al.*, 2016; Huen *et al.*, 2018). We performed the degradome sequencing on material from potato inoculated with H_2O or spore solutions from the *P. infestans* transformant harbouring PiAgo1-GFP or the wild type strain 88069 (wt). Also, mycelia degradomes from both the earlier mentioned strains were sequenced. Three sRNA datasets constructed from the following material were compared: i) Potato leaves infected with the *P. infestans* transformant PiAgo1-GFP and mycelia from the same transformant. ii) Potato leaves from a transformant harboring StAGO1a-GFP inoculated with H_2O and *P. infestans* (wt). iii) Background set with potato leaves inoculated with H_2O and *P. infestans* (wt) and *P. infestans* (wt) mycelia.

3.3.1 Degradome analysis improvements

Analysis of the sRNA and degradome data with PAREsnip2 predicted 32,886 cleavages. Manual plotting of BAM raw data from the degradome in cleavage windows revealed that most images lacked the characteristic degradome cleavage appearance (Fig. 9). Instead, the images displayed background noise. In earlier comparison of PAREsnip2, only evaluating miRNA targets, predictions were 90% true (Thody *et al.*, 2018). To separate the true cleavages from the false in this study, the R package smartPARE was constructed, based on a deep learning CNN. The CNN was implemented based on the R interface to Keras (Chollet, 2015), and was designed to comprise cyclical learning rate (CLR) and Bayesian optimization to enhance the classification accuracy of the trained model (Snoek *et al.*, 2012; Smith, 2017). Cross-validation of the final model revealed an accuracy of 100% and a loss of 0.10. Evaluation of 65,772 cleavage window images (two replicates of all 32,886 predicted cleavages) identified 4,073 true cleavages in potato and 702 in *P. infestans*.

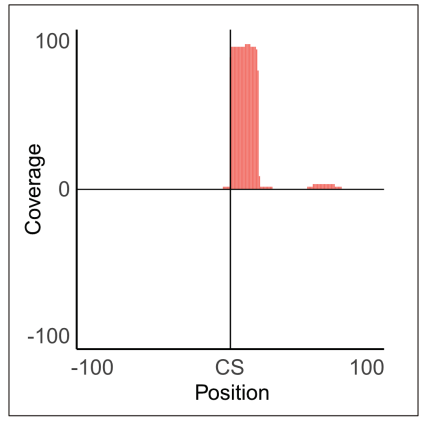


Figure 9. Cleavage plot displaying a characteristic appearance of a true cleavage in the centre of the plot at the 5' Watson strand. CS = cleavage site.

3.3.2 Evaluation of smartPARE

Apart from the outstanding cross-validation mentioned, seven miRNA targets were detected that matched pairs of miRNA/target-gene combinations already confirmed in potato or in other species. To further evaluate smartPARE, *Arabidopsis* miRNA and degradome data was applied to PAREsnip2 to generate miRNA cleavage predictions. smartPARE was utilized to test the cleavages, confirming that all predicted miRNA cleavages in the *Arabidopsis* dataset were true.

3.3.3 Infection affected cleavages revealed cascades and dual effector-functionality

To clarify which specific cleavage sites were affected by infection, comparison datasets were generated comparing normalized fragment abundance (NFA) between infection-based datasets and control datasets. With this approach, resistance genes and transcription factors protruded as the largest groups with both increased and decreased NFA in potato. As mentioned in Paper I, 638 mRNA coding for resistance genes were predicted targets by sRNA associated with PiAgo1. Analysis of the same dataset with smartPARE only mRNA from seven resistance genes could be confirmed. Expansion of the analysis to also include the background and StAGO1 datasets, confirmed cleavage of totally 105 resistance involved mRNAs.

In *P. infestans* the greatest groups of targets were genes producing ribosomal RNA and enzymes. Translocating sRNAs were discovered from both *P. infestans* and potato, where sRNA translocating from *P. infestans* effectors was a major group. This indicates that effectors have evolved to assist infection, both at the protein level and at the post-transcriptional level. Potential sRNA cascades were also detected in both *P. infestans* and potato. In *P. infestans* several potential cascades were located

to ribosomal RNAs, whereas in potato, numerous cascades were located to *PHAS* and *TAS* loci, often related to resistance genes. Furthermore, several cis-regulatory sRNAs were detected that were generated from the same loci they targeted.

3.3.4 Other sRNA might trigger phasiRNA biogenesis

From the extensive sRNA pool of all the sRNA datasets of the study phasiRNA and corresponding *PHAS* loci were predicted applying PhaseTank (Guo *et al.*, 2015), revealing phasiRNA from 114 *PHAS* loci. Although all potato miRNAs to date uploaded to mirBASE were included in the analysis only 17 of the previously mentioned *PHAS* loci were predicted triggered by miRNAs. We found 22 transcripts containing *PHAS* loci targeted by other sources of sRNA (three sRNA from *R* genes, five from other types of protein coding genes, 17 from intergenic sequences and one from an mRNA in *P. infestans*). The diversity of *PHAS* loci targeting sRNAs together with the scarcity of targeting miRNAs is raising the hypothesis that phasiRNA biogenesis could be triggered by multiple sources of sRNA. Furthermore, only one miRNA among the miRNAs predicted to trigger phasiRNA biogenesis were detected in our sRNA datasets from StAGO1a and PiAgo1, indicating that also other AGOs might be involved in the triggering of the phasiRNA biogenesis.

3.3.5 Sequencing depth and noise limits degradome coverage

Sequencing coverage is dependent on the sequencing depth (Wang *et al.*, 2009a). If the depth in RNAseq is too low, detection of rarely occurring transcripts is not possible. Theoretically, this should also apply for degradome sequencing because depth also varies at the cleavage sites, making the probability to sequence a read from a cleavage of high depth higher than a cleavage of low depth. In course of this study, several established cleavage sites in potato could not be detected in the raw cleavage data.

The level of noise is a significant factor for cleavage detection difficulties in two ways. First, the background noise makes low read cleavages blend in with the surrounding noise, making the cleavages indistinguishable from the noise. Second, the level of noise minimizes chances of detecting the rare cleavages. As a simple example, if half the reads were inferred by noise during sequencing, only half the number of reads would represent true data. The probability of detecting the rare cleavage would hence only be half the probability of detecting the same event in the same sample given there were no noise at all. This example neglects the decreased probability to detect rare events among the noise. Degradome sequencing would hence benefit from improvements in the protocol that would decrease the level of noise in the sample.

4. Conclusions

The main conclusions presented in this thesis are the following:

- The potato genome contains 14 AGOs
- Potato AGO15 diverged early in evolution from the AGO4 clade
- PiAgo1 induces 5'U sRNA preference upon infection
- *P. infestans* miR8788 induces cleavage of potato mRNA from *StABH1*
- *StABH1* is a vital gene for potato defence against *P. infestans*
- 4,073 cleavages are identified in potato and 702 in *P. infestans* by degradome analysis
- *P. infestans* effectors might possess dual functionality

Due to the covid-19 pandemic, a number of laboratory analyses originally planned are significantly delayed.

5. Future perspectives

This thesis presents a series of results opening up for further research in the area of sRNA interactions between potato and *P. infestans*.

The number of potato AGO proteins were clarified and their evolution was reconstructed. It would be interesting to further investigate in their structure and functions. Protein crystallization can be performed to deduce differences in the AGO structures. However, this process is often expensive and time consuming. Protein structure prediction tools like AlphaFold are currently developing the field and might become an alternative to making crystals (Senior et al., 2020).

sRNA association of one potato AGO and one Ago from *P. infestans* were examined in this thesis. It would be extensive but optimal to pull down sRNA from all sRNA-associating AGOs of these interacting organisms during infection. This would provide the community with an expanding picture of the potato and *P. infestans* sRNA landscape.

RXLR effectors have been shown to enter host vesicles (Petre *et al.*, 2021). However, the mechanism of sRNA transport between *P. infestans* and potato is still unknown. For example, investigation of miR8788 transport to potato would be of interest. Furthermore, knowledge in the transport of the *P. infestans* effector derived sRNAs targeting in potato could be incorporated in resistance work. If understood how these sRNAs enter the potato, methods could be developed to prevent the sRNAs from entering.

The degradome analysis revealed extensive numbers of sRNAs potentially involved in mRNA silencing during infection. These targets might be individually studied to identify the impact and implications of each diverse targeting event. It would be for example be interesting to study the cis-regulatory sRNAs, identified to target the same loci they were derived from. Moreover, only a minority of all detected *PHAS* loci also had a predicted triggering miRNA. Further work is required to deduce the trigger related to the other *PHAS* loci.

Further optimization of the degradome sequencing protocol would be of great value for the sRNA community. Noise reduction through laboratory work would increase the percentage of true cleavage reads achieved per sequencing which in turn would result in a higher true cleavage site coverage, exposing more rare cleavages.

References

Aceto S, Sica M, Paolo SD, *et al.* 2014. The analysis of the inflorescence miRNome of the orchid *Orchis italica* reveals a DEF-like MADS-box gene as a new miRNA target. *PLOS ONE* **9**: e97839.

Allen E, Xie Z, Gustafson AM, *et al.* 2005. microRNA-directed phasing during trans-acting siRNA biogenesis in plants. *Cell* **121**: 207–221.

Altschul SF, Boguski MS, Gish W, *et al.* 1994. Issues in searching molecular sequence databases. *Nature Genetics* **6**: 119–129.

Amar D, Frades I, Danek A, *et al.* 2014. Evaluation and integration of functional annotation pipelines for newly sequenced organisms: the potato genome as a test case. *BMC Plant Biology* **14**: 329.

Anders S, McCarthy DJ, Chen Y, *et al.* 2013. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nature Protocols* **8**: 1765–1786.

Andrews S. 2010. FastQC: A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Arribas-Hernández L, Marchais A, Poulsen C, *et al.* 2016. The slicer activity of ARGONAUTE1 is required specifically for the phasing, not production, of *trans*-acting short interfering RNAs in Arabidopsis. *The Plant Cell* **28**: 1563–1580.

Artzi S, Kiezun A, Shomron N. 2008. miRNAminer: A tool for homologous microRNA gene search. *BMC Bioinformatics* **9**: 39.

Ashkenazy H, Sela I, Levy Karin E, *et al.* 2019. Multiple sequence alignment averaging improves phylogeny reconstruction. *Systematic Biology* **68**: 117–130.

Åsman AKM, Fogelqvist J, Vetukuri RR, *et al.* 2016. *Phytophthora infestans* Argonaute 1 binds microRNA and small RNAs from effector genes and transposable elements. *The New Phytologist* **211**: 993–1007.

Axtell MJ. 2013. ShortStack: Comprehensive annotation and quantification of small RNA genes. *RNA* **19**: 740.

Axtell M. 2020. MikeAxtell/ShortStack. https://github.com/MikeAxtell/ShortStack .

Bai M, Yang G-S, Chen W-T, *et al.* 2012. Genome-wide identification of Dicer-like, Argonaute and RNA-dependent RNA polymerase gene families and their expression analyses in response to viral infection and abiotic stresses in *Solanum lycopersicum*. *Gene* **501**: 52–62.

Bartel DP. 2009. MicroRNAs: Target recognition and regulatory functions. *Cell* **136**: 215–233.

Batzoglou S, Jaffe DB, Stanley K, *et al.* 2002. ARACHNE: A whole-genome shotgun assembler. *Genome Research* **12**: 177–189.

Bauer FL, Wössner H. 1972. The 'Plankalkül' of Konrad Zuse: a forerunner of today's programming languages. *Communications of the ACM* **15**: 678–685.

Baumberger N, Baulcombe DC. 2005. Arabidopsis ARGONAUTE1 is an RNA Slicer that selectively recruits microRNAs and short interfering RNAs. *Proceedings of the National Academy of Sciences USA* **102**: 11928–11933.

Bayón GF, Fernández AF, Fraga MF. 2016. Chapter 4 - Bioinformatics tools in epigenomics studies. In: Fraga MF, Fernández AF, eds. Epigenomics in health and disease. Boston: Academic Press, 73–107.

Beckers M, Mohorianu I, Stocks M, *et al.* 2017. Comprehensive processing of high-throughput small RNA sequencing data including quality checking, normalization, and differential expression analysis using the UEA sRNA Workbench. *RNA* **23**: 823–835.

Beerli P. 2006. Comparison of Bayesian and maximum-likelihood inference of population genetic parameters. *Bioinformatics* **22**: 341–345.

Blackwell AF. 2002. What is programming? In: In Proceedings of PPIG 2002. MIT Press, 204–218.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114–2120.

Bologna NG, Voinnet O. 2014. The diversity, biogenesis, and activities of endogenous silencing small RNAs in Arabidopsis. *Annual Review of Plant Biology* **65**: 473–503.

Bonnet E, He Y, Billiau K, *et al.* 2010. TAPIR, a web server for the prediction of plant microRNA targets, including target mimics. *Bioinformatics* **26**: 1566–1568.

Borges F, Martienssen RA. 2015. The expanding world of small RNAs in plants. *Nature Reviews Molecular Cell Biology* **16**: 727–741.

Branscheid A, Marchais A, Schott G, *et al.* 2015. SKI2 mediates degradation of RISC 5′-cleavage fragments and prevents secondary siRNA production from miRNA targets in Arabidopsis. *Nucleic Acids Research* **43**: 10975–10988.

Brant EJ, Budak H. **2018**. Plant small non-coding RNAs and their roles in biotic stresses. *Frontiers in Plant Science* **9**: 1038.

Bray NL, Pimentel H, Melsted P, *et al.* 2016. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology* **34**: 525–527.

Brousse C, Liu Q, Beauclair L, *et al.* 2014. A non-canonical plant microRNA target site. *Nucleic Acids Research* **42**: 5270–5279.

Brunette ES, Flemmer RC, Flemmer CL. **2009**. A review of artificial intelligence. In: 2009 4th International Conference on Autonomous Robots and Agents. Wellington: IEEE, 385–392.

Buermans HPJ, den Dunnen JT. **2014**. Next generation sequencing technology: Advances and applications. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* **1842**: 1932–1941.

Camacho C, Coulouris G, Avagyan V, *et al.* 2009. BLAST+: architecture and applications. *BMC Bioinformatics* **10**: 421.

Cano L, Kamoun S, Win J. **2019**. The effector secretome of the Irish potato famine pathogen *Phytophthora infestans*. https://zenodo.org/record/3574589#.YFRIXhZ7mXI.

Chang J-M, Di Tommaso P, Notredame C. **2014**. TCS: A new multiple sequence alignment reliability measure to estimate alignment accuracy and improve phylogenetic tree reconstruction. *Molecular Biology and Evolution* **31**: 1625–1637.

Chen H-M, Chen L-T, Patel K, *et al.* 2010. 22-nucleotide RNAs trigger secondary siRNA biogenesis in plants. *Proceedings of the National Academy of Sciences USA* **107**: 15269–15274.

Chen L, Heikkinen L, Wang C, *et al.* 2019. Trends in the development of miRNA bioinformatics tools. *Briefings in Bioinformatics* **20**: 1836–1852.

Chen C, Zeng Z, Liu Z, *et al.* 2018a. Small RNAs, emerging regulators critical for the development of horticultural traits. *Horticulture Research* **5**.

Chen S, Zhou Y, Chen Y, *et al.* 2018b. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**: i884–i890.

Chi SW, Zang JB, Mele A, *et al.* 2009. Ago HITS-CLIP decodes miRNA-mRNA interaction maps. *Nature* **460**: 479–486.

Chollet F. **2015**. Keras. https://github.com/fchollet/keras.

Chorostecki U, Palatnik JF. **2014**. comTAR: a web tool for the prediction and characterization of conserved microRNA targets in plants. *Bioinformatics* **30**: 2066–2067.

Collins K, Warnow T. **2018**. PASTA for proteins. *Bioinformatics* **34**: 3939–3941.

Compeau PEC, Pevzner PA, Tesler G. **2011**. Why are de Bruijn graphs useful for genome assembly? *Nature Biotechnology* **29**: 987–991.

Conesa A, Madrigal P, Tarazona S, *et al.* 2016. A survey of best practices for RNA-seq data analysis. *Genome Biology* **17**: 13.

Copeland BJ. **2020**. Artificial intelligence. *Encyclopædia Britannica*. https://www.britannica.com/technology/artificial-intelligence.

Corchete LA, Rojas EA, Alonso-López D, *et al.* 2020. Systematic comparison and assessment of RNA-seq procedures for gene expression quantitative analysis. *Scientific Reports* **10**: 19737.

Cuperus JT, Carbonell A, Fahlgren N, *et al.* 2010. Unique functionality of 22 nt miRNAs in triggering RDR6-dependent siRNA biogenesis from target transcripts in Arabidopsis. *Nature Structural & Molecular Biology* **17**: 997–1003.

Dai X, Zhao PX. **2011**. psRNATarget: a plant small RNA target analysis server. *Nucleic Acids Research* **39**: W155–W159.

Dai X, Zhuang Z, Zhao PX. **2018**. psRNATarget: a plant small RNA target analysis server (2017 release). *Nucleic Acids Research* **46**: W49–W54.

Desvignes T, Batzel P, Berezikov E, *et al.* 2015. microRNA nomenclature: A view incorporating genetic origins, biosynthetic pathways, and sequence variants. *Trends in Genetics* **31**: 613–626.

Devers EA, Branscheid A, May P, *et al.* 2011. Stars and symbiosis: microRNA- and microRNA*-mediated transcript cleavage involved in arbuscular mycorrhizal symbiosis. *Plant Physiology* **156**: 1990–2010.

Di Franco A, Poujol R, Baurain D, *et al.* 2019. Evaluating the usefulness of alignment filtering methods to reduce the impact of errors on evolutionary inferences. *BMC Evolutionary Biology* **19**: 21.

van Dijk EL, Jaszczyszyn Y, Naquin D, *et al.* 2018. The third revolution in sequencing technology. *Trends in Genetics* **34**: 666–681.

Doench JG, Petersen CP, Sharp PA. **2003**. siRNAs can function as miRNAs. *Genes & Development* **17**: 438–442.

Dong Z, Han M-H, Fedoroff N. **2008**. The RNA-binding proteins HYL1 and SE promote accurate in vitro processing of pri-miRNA by DCL1. *Proceedings of the National Academy of Sciences USA* **105**: 9970–9975.

Drenth A, Tas I, Govers F. **1994**. DNA fingerprinting of the potato late blight fungus, *Phytophthora infestans*. *European Journal of Plant Pathology* **100**: 97–107.

Du AX, Emam S, Gniadecki R. 2020. Review of machine learning in predicting dermatological outcomes. *Frontiers in Medicine* **7**: 266.

Dumschott K, Schmidt MH-W, Chawla HS, *et al.* 2020. Oxford Nanopore sequencing: new opportunities for plant genomics? *Journal of Experimental Botany* **71**: 5313–5322.

Eamens AL, Smith NA, Curtin SJ, *et al.* 2009. The Arabidopsis thaliana double-stranded RNA binding protein DRB1 directs guide strand selection from microRNA duplexes. *RNA* **15**: 2219–2235.

Eisenhaber F. 2013. Prediction of Protein Function Two Basic Concepts and One Practical Recipe. https://www.ncbi.nlm.nih.gov/books/NBK6301/. Landes Bioscience.

Eisenhaber B, Kuchibhatla D, Sherman W, *et al.* 2016. The recipe for protein sequence-based function prediction and its implementation in the ANNOTATOR software environment. In: Methods in Molecular Biology (Clifton, N.J.). 477–506.

El-Gebali S, Mistry J, Bateman A, *et al.* 2019. The Pfam protein families database in 2019. *Nucleic Acids Research* **47**: D427–D432.

Emrich SJ, Barbazuk WB, Li L, *et al.* 2007. Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Research* **17**: 69–73.

Esposito S, Aversano R, D'Amelia V, *et al.* 2018. Dicer-like and RNA-dependent RNA polymerase gene family identification and annotation in the cultivated *Solanum tuberosum* and its wild relative *S. commersonii*. *Planta* **248**: 729–743.

Ewels P, Magnusson M, Lundin S, *et al.* 2016. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**: 3047–3048.

Faehnle CR, Elkayam E, Haase AD, *et al.* 2013. The making of a slicer: activation of human Argonaute-1. *Cell Reports* **3**: 1901–1909.

Fahlgren N, Carrington JC. 2010. miRNA target prediction in plants. In: Methods in Molecular Biology. Plant microRNAs: methods and protocols. Totowa, NJ: Humana Press, 51–57.

Fahlgren N, Howell MD, Kasschau KD, *et al.* 2007. High-throughput sequencing of Arabidopsis microRNAs: Evidence for frequent birth and death of MIRNA genes. *PLOS ONE* **2**: e219.

Fang X, Qi Y. 2016. RNAi in plants: An Argonaute-centered view. *The Plant Cell* **28**: 272–285.

Fei Q, Li P, Teng C, *et al.* 2015. Secondary siRNAs from *Medicago NB-LRRs* modulated via miRNA–target interactions and their abundances. *The Plant Journal* **83**: 451–465.

Fei Q, Xia R, Meyers BC. 2013. Phased, secondary, small interfering RNAs in posttranscriptional regulatory networks. *The Plant Cell* **25**: 2400–2415.

Fire A, Xu S, Montgomery MK, *et al.* 1998. Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* **391**: 806–811.

Folkes L, Moxon S, Woolfenden HC, *et al.* 2012. PAREsnip: a tool for rapid genome-wide discovery of small RNA/target interactions evidenced through degradome sequencing. *Nucleic Acids Research* **40**: e103.

Franco-Zorrilla JM, Valli A, Todesco M, *et al.* 2007. Target mimicry provides a new mechanism for regulation of microRNA activity. *Nature Genetics* **39**: 1033–1037.

Furuichi Y. 2014. Caps on Eukaryotic mRNAs. In: eLS. American Cancer Society.

Furuichi Y. 2015. Discovery of m(7)G-cap in eukaryotic mRNAs. Proceedings of the Japan Academy. Series B, Physical and Biological Sciences **91**: 394–409.

Gao Z, Nie J, Wang H. 2020. MicroRNA biogenesis in plant. *Plant Growth Regulation*.

Gebert D, Hewel C, Rosenkranz D. 2017. unitas: the universal tool for annotation of small RNAs. *BMC Genomics* **18**: 644.

Gebhardt C. 2016. The historical role of species from the Solanaceae plant family in genetic research. *Theoretical and Applied Genetics* **129**: 2281–2294.

Gentleman RC, Carey VJ, Bates DM, *et al.* 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology* **5**: R80.

German MA, Pillay M, Jeong D-H, *et al.* 2008. Global identification of microRNA–target RNA pairs by parallel analysis of RNA ends. *Nature Biotechnology* **26**: 941–946.

Grant GR, Farkas MH, Pizarro AD, *et al.* 2011. Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). *Bioinformatics* **27**: 2518–2528.

Guindon S, Dufayard J-F, Lefort V, *et al.* 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Systematic Biology* **59**: 307–321.

Guo G, Liu X, Sun F, *et al.* 2018. Wheat miR9678 affects seed germination by generating phased siRNAs and modulating abscisic acid/gibberellin signaling. *The Plant Cell* **30**: 796–814.

Guo Q, Qu X, Jin W. 2015. PhaseTank: genome-wide computational identification of phasiRNAs and their regulatory cascades. *Bioinformatics* **31**: 284–286.

Gusev Y, Brackett DJ. 2007. MicroRNA expression profiling in cancer from a bioinformatics prospective. *Expert Review of Molecular Diagnostics* **7**: 787–792.

Haas BJ, Kamoun S, Zody MC, *et al.* 2009. Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*. *Nature* **461**: 393–398.

Hamilton AJ, Baulcombe DC. 1999. A species of small antisense RNA in posttranscriptional gene silencing in plants. *Science* **286**: 950–952.

Han G-Z. 2019. Origin and evolution of the plant immune system. *New Phytologist* **222**: 70–83.

Hardcastle TJ, Kelly KA. 2010. baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics* **11**: 422.

Hardigan MA, Crisovan E, Hamilton JP, *et al.* 2016. Genome reduction uncovers a large dispensable genome and adaptive role for copy number variation in asexually propagated *Solanum tuberosum*. *The Plant Cell* **28**: 388–405.

Haverkort AJ, Boonekamp PM, Hutten R, *et al.* 2008. Societal costs of late blight in potato and prospects of durable resistance through cisgenic modification. *Potato Research* **51**: 47–57.

Heather JM, Chain B. 2016. The sequence of sequencers: The history of sequencing DNA. *Genomics* **107**: 1–8.

Höck J, Meister G. 2008. The Argonaute protein family. *Genome Biology* **9**: 210.

Hogarty D, Mackey D, Hewitt A. 2018. Current state and future prospects of artificial intelligence in ophthalmology: a review: Artificial intelligence in ophthalmology. *Clinical & Experimental Ophthalmology* **47**.

Hogarty DT, Su JC, Phan K, *et al.* 2020. Artificial intelligence in dermatology—where we are and the way to the future: A review. *American Journal of Clinical Dermatology* **21**: 41–47.

Hogeweg P. 2011. The roots of bioinformatics in theoretical biology. *PLOS Computational Biology* **7**: e1002021.

Holley RW, Apgar J, Everett GA, *et al.* 1965. Structure of a ribonucleic acid. *Science* **147**: 1462–1465.

Howell MD, Fahlgren N, Chapman EJ, *et al.* 2007. Genome-wide analysis of the RNA-DEPENDENT RNA POLYMERASE6/DICER-LIKE4 pathway in *Arabidopsis* reveals dependency on miRNA- and tasiRNA-directed targeting. *The Plant Cell* **19**: 926–942.

Huen A, Bally J, Smith P. 2018. Identification and characterisation of microRNAs and their target genes in phosphate-starved *Nicotiana benthamiana* by small RNA deep sequencing and 5'RACE analysis. *BMC Genomics* **19**: 940.

Jain M, Olsen HE, Turner DJ, *et al.* 2018. Linear assembly of a human centromere on the Y chromosome. *Nature Biotechnology* **36**: 321–323.

Jeske T, Huypens P, Stirm L, *et al.* 2019. DEUS: an R package for accurate small RNA profiling based on differential expression of unique sequences. *Bioinformatics* **35**: 4834–4836.

Jin H, Wan Y-W, Liu Z. 2017. Comprehensive evaluation of RNA-seq quantification methods for linearity. *BMC Bioinformatics* **18**: 117.

Johnson C, Kasprzewska A, Tennessen K, *et al.* 2009. Clusters and superclusters of phased small RNAs in the developing inflorescence of rice. *Genome Research* **19**: 1429–1440.

Kahanwal B. 2013. Abstraction level taxonomy of programming language frameworks. *International Journal of Programming Languages and Applications* **3**: 1–12.

Kakrana A, Hammond R, Patel P, *et al.* 2014. sPARTA: a parallelized pipeline for integrated analysis of plant miRNA and cleaved mRNA data sets, including new miRNA target-identification software. *Nucleic Acids Research* **42**: e139–e139.

Kalyanaraman A. 2011. Genome assembly. In: Padua D, ed. Encyclopedia of Parallel Computing. Boston, MA: Springer US, 755–768.

Kapoor M, Arora R, Lama T, *et al.* 2008. Genome-wide identification, organization and phylogenetic analysis of Dicer-like, Argonaute and RNA-dependent RNA Polymerase gene families and their expression analysis during reproductive development and stress in rice. *BMC Genomics* **9**: 451.

Kemena C, Notredame C. 2009. Upcoming challenges for multiple sequence alignment methods in the high-throughput era. *Bioinformatics* **25**: 2455–2465.

Kim D, Paggi JM, Park C, *et al.* 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology* **37**: 907–915.

Kim D, Pertea G, Trapnell C, *et al.* 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology* **14**: R36.

Knaus BJ, Tabima JF, Davis CE, *et al.* 2016. Genomic analyses of dominant U.S. clonal lineages of *Phytophthora infestans* reveals a shared common ancestry for clonal lineages US11 and US18 and a lack of recently shared ancestry among all other U.S. lineages. *Phytopathology* **106**: 1393–1403.

Kozlov AM, Aberer AJ, Stamatakis A. 2015. ExaML version 3: a tool for phylogenomic analyses on supercomputers. *Bioinformatics* **31**: 2577–2579.

Kozomara A, Birgaoanu M, Griffiths-Jones S. 2019. miRBase: from microRNA sequences to function. *Nucleic Acids Research* **47**: D155–D162.

Krichevsky AM, King KS, Donahue CP, *et al.* 2003. A microRNA array reveals extensive regulation of microRNAs during brain development. *RNA* **9**: 1274–1281.

Kück P, Meusemann K, Dambach J, *et al.* 2010. Parametric and non-parametric masking of randomness in sequence alignments can be improved and leads to better resolved trees. *Frontiers in Zoology* **7**: 10.

Kurihara Y, Watanabe Y. **2004**. Arabidopsis micro-RNA biogenesis through Dicer-like 1 protein functions. *Proceedings of the National Academy of Sciences USA* **101**: 12753–12758.
Kwak PB, Tomari Y. **2012**. The N domain of Argonaute drives duplex unwinding during RISC assembly. *Nature Structural & Molecular Biology* **19**: 145–151.
Langmead B, Salzberg SL. **2012**. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**: 357–359.
Law CW, Chen Y, Shi W, *et al.* 2014. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology* **15**: R29.
LeCun Y, Bengio Y, Hinton G. **2015**. Deep learning. *Nature* **521**: 436–444.
Lee Y, Cho K-S, Seo J-H, *et al.* 2020. Improved genome sequence and gene annotation resource for the potato late blight pathogen *Phytophthora infestans*. *Molecular Plant-Microbe Interactions* **33**: 1025–1028.
Lee RC, Feinbaum RL, Ambros V. **1993**. The C. elegans heterochronic gene *lin-14* encodes small RNAs with antisense complementarity to *lin-14*. *Cell* **75**: 843–854.
Lee T van der, Robold A, Testa A, *et al.* 2001. Mapping of avirulence genes in *Phytophthora infestans* with amplified fragment length polymorphism markers selected by bulked segregant analysis. *Genetics* **157**: 949–956.
Lees JA, Kendall M, Parkhill J, *et al.* 2018. Evaluation of phylogenetic reconstruction methods using bacterial whole genomes: a simulation based study. *Wellcome Open Research* **3**.
Leesutthiphonchai W, Vu AL, Ah-Fong AMV, *et al.* 2018. How does *Phytophthora infestans* evade control efforts? Modern insight into the late blight disease. *Phytopathology* **108**: 916–924.
Leibman D, Kravchik M, Wolf D, *et al.* 2018. Differential expression of cucumber RNA-dependent RNA polymerase 1 genes during antiviral defence and resistance. *Molecular Plant Pathology* **19**: 300–312.
Leuschen RL, Downing BL. **2020**. Gene editing of *Arabidopsis thaliana* RNA dependent RNA polymerases using CRISPR/Cas9. *The FASEB Journal* **34**: 1–1.
Li B, Dewey CN. **2011**. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**: 323.
Li F, Huang C, Li Z, *et al.* 2014. Suppression of RNA silencing by a plant DNA virus satellite requires a host Calmodulin-like protein to repress RDR6 expression. *PLOS Pathogens* **10**: e1003921.
Li J, Kho AT, Chase RP, *et al.* 2020. COMPSRA: a COMprehensive Platform for Small RNA-Seq data Analysis. *Scientific Reports* **10**: 4552.
Li Y, Shen H, Zhou Q, *et al.* 2017. Changing ploidy as a strategy: the Irish potato famine pathogen shifts ploidy in relation to its sexuality. *Molecular Plant-Microbe Interactions:* **30**: 45–52.
Ling H, Fabbri M, Calin GA. **2013**. MicroRNAs and other non-coding RNAs as targets for anticancer drug development. *Nature Reviews. Drug discovery* **12**: 847–865.
Liu Y, Teng C, Xia R, *et al.* 2020. PhasiRNAs in plants: Their biogenesis, genic sources, and roles in stress responses, development, and reproduction. *The Plant Cell* **32**: 3059–3080.
Liu Q, Wang F, Axtell MJ. **2014**. Analysis of complementarity requirements for plant microRNA targeting using a *Nicotiana benthamiana* quantitative transient assay. *The Plant Cell* **26**: 741–753.
Llave C, Xie Z, Kasschau KD, *et al.* 2002. Cleavage of *Scarecrow-like* mRNA targets directed by a class of *Arabidopsis* miRNA. *Science* **297**: 2053–2056.
Lu W, Tong Y, Yu Y, *et al.* 2018. Applications of artificial intelligence in ophthalmology: General overview. *Journal of Ophthalmology* **2018**: e5278196.
Lucas SJ, Budak H. **2012**. Sorting the wheat from the chaff: Identifying miRNAs in genomic survey sequences of *Triticum aestivum* chromosome 1AL. *PLOS ONE* **7**: e40859.
Lutzmayer S, Enugutti B, Nodine MD. **2017**. Novel small RNA spike-in oligonucleotides enable absolute normalization of small RNA-Seq data. *Scientific Reports* **7**: 5913.
Ma J-B, Ye K, Patel DJ. **2004**. Structural basis for overhang-specific small interfering RNA recognition by the PAZ domain. *Nature* **429**: 318–322.
Ma J-B, Yuan Y-R, Meister G, *et al.* 2005. Structural basis for 5′-end-specific recognition of guide RNA by the *A. fulgidus* Piwi protein. *Nature* **434**: 666–670.
Machida-Hirano R. **2015**. Diversity of potato genetic resources. *Breeding Science* **65**: 26.
Manavella PA, Hagmann J, Ott F, *et al.* 2012a. Fast-forward genetics identifies plant CPL phosphatases as regulators of miRNA processing factor HYL1. *Cell* **151**: 859–870.
Manavella PA, Koenig D, Weigel D. **2012b**. Plant secondary siRNA production determined by microRNA-duplex structure. *Proceedings of the National Academy of Sciences of the USA* **109**: 2461–2466.
Mangalam H. **2002**. The Bio* toolkits — a brief overview. *Briefings in Bioinformatics* **3**: 296–302.
Manrique-Carpintero NC, Coombs JJ, Pham GM, *et al.* 2018. Genome reduction in tetraploid potato reveals genetic load, haplotype variation, and loci associated with agronomic traits. *Frontiers in Plant Science* **9**.
Margis R, Fusaro AF, Smith NA, *et al.* 2006. The evolution and diversification of Dicers in plants. *FEBS Letters* **580**: 2442–2450.

Marin E, Jouannet V, Herz A, *et al.* 2010. miR390, *Arabidopsis TAS3* tasiRNAs, and their *AUXIN RESPONSE FACTOR* targets define an autoregulatory network quantitatively regulating lateral root growth. *The Plant Cell* **22**: 1104–1117.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**: 10–12.

Martinez J, Patkaniowska A, Urlaub H, *et al.* 2002. Single-stranded antisense siRNAs guide target RNA cleavage in RNAi. *Cell* **110**: 563–574.

Mattick JS, Makunin IV. 2006. Non-coding RNA. *Human Molecular Genetics* **15**: R17–R29.

McDermaid A, Monier B, Zhao J, *et al.* 2019. Interpretation of differential gene expression results of RNA-seq data: review and integration. *Briefings in Bioinformatics* **20**: 2044–2054.

Mendes ND, Freitas AT, Sagot M-F. 2009. Current tools for the identification of miRNA genes and their targets. *Nucleic Acids Research* **37**: 2419–2433.

Merico D, Isserlin R, Stueker O, *et al.* 2010. Enrichment map: A network-based method for gene-set enrichment visualization and interpretation. *PLOS ONE* **5**: e13984.

Meyer SU, Pfaffl MW, Ulbrich SE. 2010. Normalization strategies for microRNA profiling experiments: a 'normal' way to a hidden layer of complexity? *Biotechnology Letters* **32**: 1777–1788.

Mi S, Cai T, Hu Y, *et al.* 2008. Sorting of small RNAs into *Arabidopsis* Argonaute complexes is directed by the 5′ terminal nucleotide. *Cell* **133**: 116–127.

Mishra AK, Duraisamy GS, Matoušek J. 2015. Discovering microRNAs and their targets in plants. *Critical Reviews in Plant Sciences* **34**: 553–571.

Mitchell AL, Attwood TK, Babbitt PC, *et al.* 2019. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Research* **47**: D351–D360.

Miyoshi T, Ito K, Murakami R, *et al.* 2016. Structural basis for the recognition of guide RNA and target DNA heteroduplex by Argonaute. *Nature Communications* **7**: 11846.

Mo W, Luo X, Zhong Y, *et al.* 2019. Image recognition using convolutional neural network combined with ensemble learning algorithm. *Journal of Physics: Conference Series* **1237**: 022026.

Montgomery SB, Sammeth M, Gutierrez-Arcelus M, *et al.* 2010. Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**: 773–777.

Moore MM, Slonimsky E, Long AD, *et al.* 2019. Machine learning concepts, concerns and opportunities for a pediatric radiologist. *Pediatric Radiology* **49**: 509–516.

Morel J-B, Godon C, Mourrain P, *et al.* 2002. Fertile hypomorphic ARGONAUTE (ago1) mutants impaired in post-transcriptional gene silencing and virus resistance. *The Plant Cell* **14**: 629–639.

Morgado L, Johannes F. 2019. Computational tools for plant small RNA detection and categorization. *Briefings in Bioinformatics* **20**: 1181–1192.

Motameny S, Wolters S, Nürnberg P, *et al.* 2010. Next generation sequencing of miRNAs – strategies, resources and methods. *Genes* **1**: 70–84.

Mukkulainen R. 1990. Script recognition with hierarchical feature maps. *Connection Science* **2**: 83–101.

Nazarie FW, Shih B, Angus T, *et al.* 2019. Visualization and analysis of RNA-Seq assembly graphs. *Nucleic Acids Research* **47**: 7262–7275.

Nguyen L-T, Schmidt HA, von Haeseler A, *et al.* 2015. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution* **32**: 268–274.

Nikhitha NK, Afzal AL, Asharaf S. 2020. Deep Kernel machines: a survey. *Pattern Analysis and Applications*.

Nyrén P. 1987. Enzymatic method for continuous monitoring of DNA polymerase activity. *Analytical Biochemistry* **167**: 235–238.

Ogden TH, Rosenberg MS. 2006. Multiple sequence alignment accuracy and phylogenetic inference. *Systematic Biology* **55**: 314–328.

Oliphant TE. 2007. Python for scientific computing. *Computing in Science Engineering* **9**: 10–20.

Paeglis A, Strumfs B, Mezale D, *et al.* 2018. A review on machine learning and deep learning techniques applied to liquid biopsy. *Liquid Biopsy*.

Pascanu R, Gulcehre C, Cho K, *et al.* 2013. How to construct deep recurrent neural networks.

Patro R, Duggal G, Love MI, *et al.* 2017. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods* **14**: 417–419.

Patro R, Mount SM, Kingsford C. 2014. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nature Biotechnology* **32**: 462–464.

Peng T, Lv Q, Zhang J, *et al.* 2011. Differential expression of the microRNAs in superior and inferior spikelets in rice (*Oryza sativa*). *Journal of Experimental Botany* **62**: 4943–4954.

Peragine A, Yoshikawa M, Wu G, *et al.* 2004. *SGS3* and *SGS2/SDE1/RDR6* are required for juvenile development and the production of *trans*-acting siRNAs in *Arabidopsis*. *Genes & Development* **18**: 2368–2379.

Petre B, Contreras MP, Bozkurt TO, *et al.* 2021. Host-interactor screens of *Phytophthora infestans* RXLR proteins reveal vesicle trafficking as a major effector-targeted process. *The Plant Cell*. koab069.
Pham GM, Hamilton JP, Wood JC, *et al.* 2020. Construction of a chromosome-scale long-read reference genome assembly for potato. *GigaScience* **9**.
Pogorelcnik R, Vaury C, Pouchin P, *et al.* 2018. sRNAPipe: a Galaxy-based pipeline for bioinformatic in-depth exploration of small RNAseq data. *Mobile DNA* **9**: 25.
Polydore S, Axtell MJ. 2018. Analysis of *RDR1/RDR2/RDR6*-independent small RNAs in *Arabidopsis thaliana* improves *MIRNA* annotations and reveals unexplained types of short interfering RNA loci. *The Plant Journal* **94**: 1051–1063.
Price MN, Dehal PS, Arkin AP. 2010. FastTree 2 – Approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**.
Qi Y, Denli AM, Hannon GJ. 2005. Biochemical specialization within *Arabidopsis* RNA silencing pathways. *Molecular Cell* **19**: 421–428.
Qin L-X, Zou J, Shi J, *et al.* 2020. Statistical Assessment of depth normalization for small RNA sequencing. *JCO Ciinical Cancer Informatics* **4**: 567–582.
R Core Team. 2017. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.
Raplee ID, Evsikov AV, Marín de Evsikova C. 2019. Aligning the aligners: comparison of RNA sequencing data alignment and gene expression quantification tools for clinical breast cancer research. *Journal of Personalized Medicine* **9**: 18.
Reimand J, Arak T, Adler P, *et al.* 2016. g:Profiler—a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Research* **44**: W83–W89.
Reimand J, Isser R, Voisin V, *et al.* 2019. Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nature Protocols* **14**: 482–517.
Risca VI, Greenleaf WJ. 2015. Beyond the linear genome: Paired-end sequencing as a biophysical tool. *Trends in Cell Biology* **25**: 716–719.
Ritchie ME, Phipson B, Wu D, *et al.* 2015. *limma* powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* **43**: e47–e47.
Roberts A, Pachter L. 2013. Streaming fragment assignment for real-time analysis of sequencing experiments. *Nature Methods* **10**: 71–73.
Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**: 139–140.
Rojas R, Göktekin C, Friedland G, *et al.* 2000. Plankalkül: The first high-level programming language and its implementation. *Berlin: Feinarbeit.*
Rueda A, Barturen G, Lebrón R, *et al.* 2015. sRNAtoolbox: an integrated collection of small RNA research tools. *Nucleic Acids Research* **43**: W467-473.
Rutter L, Moran Lauter AN, Graham MA, *et al.* 2019. Visualization methods for differential expression analysis. *BMC Bioinformatics* **20**: 458.
Salakhutdinov R, Hinton G. 2012. An efficient learning procedure for deep Boltzmann machines. *Neural Computation* **24**: 1967–2006.
Salimi H, Bahar M, Mirlohi A, *et al.* 2016. Assessment of the genetic diversity among potato cultivars from different geographical areas using the genomic and EST microsatellites. *Iranian Journal of Biotechnology* **14**: 270–277.
Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the USA* **74**: 5463–5467.
Sanz-Carbonell A, Marques MC, Martinez G, *et al.* 2020. Dynamic architecture and regulatory implications of the miRNA network underlying the response to stress in melon. *RNA Biology* **17**: 292–308.
Sato S, Tabata S, Hirakawa H, *et al.* 2012. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **485**: 635–641.
Scalzitti N, Jeannin-Girardon A, Collet P, *et al.* 2020. A benchmark study of ab initio gene prediction methods in diverse eukaryotic organisms. *BMC Genomics* **21**: 293.
Schäfer M, Ciaudo C. 2020. Prediction of the miRNA interactome – Established methods and upcoming perspectives. *Computational and Structural Biotechnology Journal* **18**: 548–557.
Schwab R, Palatnik JF, Riester M, *et al.* 2005. Specific effects of microRNAs on the plant transcriptome. *Developmental Cell* **8**: 517–527.
Searle JR. 1980. Minds, brains, and programs. *Behavioral and Brain Sciences* **3**: 417–424.
Senior AW, Evans R, Jumper J, *et al.* 2020. Improved protein structure prediction using potentials from deep learning. *Nature* **577**: 706–710.
Shahid S, Kim G, Johnson NR, *et al.* 2018. MicroRNAs from the parasitic plant *Cuscuta campestris* target host messenger RNAs. *Nature* **553**: 82–85.

Shannon P, Markiel A, Ozier O, *et al.* 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research* **13**: 2498–2504.
Sharma SK, Bolser D, Boer J de, *et al.* 2013. Construction of reference chromosome-scale pseudomolecules for potato: integrating the potato genome with genetic and physical maps. *G3: Genes, Genomes, Genetics* **3**: 2031–2047.
Sharpee WC, Dean RA. 2016. Form and function of fungal and oomycete effectors. *Fungal Biology Reviews* **30**: 62–73.
Shin C, Nam J-W, Farh KK-H, *et al.* 2010. Expanding the microRNA targeting code: Functional sites with centered pairing. *Molecular Cell* **38**: 789–802.
Smith LN. 2017. Cyclical learning rates for training neural networks. In: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV). 464–472.
Snoek J, Larochelle H, Adams RP. 2012. Practical Bayesian optimization of machine learning algorithms. *arXiv:1206.2944*.
Som A. 2015. Causes, consequences and solutions of phylogenetic incongruence. *Briefings in Bioinformatics* **16**: 536–548.
Song L, Han M-H, Lesicka J, *et al.* 2007. *Arabidopsis* primary microRNA processing proteins HYL1 and DCL1 define a nuclear body distinct from the Cajal body. *Proceedings of the National Academy of Sciences USA* **104**: 5437–5442.
Srivastava PK, Moturu TR, Pandey P, *et al.* 2014. A comparison of performance of plant miRNA target prediction tools and the characterization of features for genome-wide target prediction. *BMC Genomics* **15**: 348.
Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**: 1312–1313.
Stark R, Grzelak M, Hadfield J. 2019. RNA sequencing: the teenage years. *Nature Reviews Genetics* **20**: 631–656.
Subramanian A, Tamayo P, Mootha VK, *et al.* 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the USA* **102**: 15545–15550.
Tan G, Muffato M, Ledergerber C, *et al.* 2015. Current methods for automated filtering of multiple sequence alignments frequently worsen single-gene phylogenetic inference. *Systematic Biology* **64**: 778–791.
Tarazona S, García-Alcalde F, Dopazo J, *et al.* 2011. Differential expression in RNA-seq: A matter of depth. *Genome Research* **21**: 2213–2223.
Taylor RS, Tarver JE, Hiscock SJ, *et al.* 2014. Evolutionary history of plant microRNAs. *Trends in Plant Science* **19**: 175–182.
Teng M, Love MI, Davis CA, *et al.* 2016. A benchmark for RNA-seq quantification pipelines. *Genome Biology* **17**: 74.
The potato genome sequencing consortium. 2011. Genome sequence and analysis of the tuber crop potato. *Nature* **475**: 189–195.
Thody J, Folkes L, Medina-Calzada Z, *et al.* 2018. PAREsnip2: a tool for high-throughput prediction of small RNA targets from degradome sequencing data using configurable targeting rules. *Nucleic Acids Research* **46**: 8730–8739.
Thomas PD, Campbell MJ, Kejariwal A, *et al.* 2003. PANTHER: A library of protein families and subfamilies indexed by function. *Genome Research* **13**: 2129–2141.
Thorvaldsdóttir H, Robinson JT, Mesirov JP. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics* **14**: 178–192.
Tong L, Wu P-Y, Phan JH, *et al.* 2020. Impact of RNA-seq data analysis algorithms on gene expression estimation and downstream prediction. *Scientific Reports* **10**: 17925.
Trapnell C, Williams BA, Pertea G, *et al.* 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology* **28**: 511–515.
Tzelepis G, Hodén KP, Fogelqvist J, *et al.* 2020. Dominance of mating type A1 and indication of epigenetic effects during early stages of mating in *Phytophthora infestans*. *Frontiers in Microbiology* **11**.
The UniProt Consortium. 2017. UniProt: the universal protein knowledgebase. *Nucleic Acids Research* **45**: D158–D169.
Van Rossum G, Drake FL. 2009. Python 3 Reference Manual. Scotts Valley, CA: CreateSpace.
Vazquez F. 2006. *Arabidopsis* endogenous small RNAs: highways and byways. *Trends in Plant Science* **11**: 460–468.
Vetukuri RR, Avrova AO, Grenville-Briggs LJ, *et al.* 2011. Evidence for involvement of Dicer-like, Argonaute and histone deacetylase proteins in gene silencing in *Phytophthora infestans*. *Molecular Plant Pathology* **12**: 772–785.
Vetukuri RR, Tripathy S, Malar C M, *et al.* 2018. Draft genome sequence for the tree pathogen *Phytophthora plurivora*. *Genome Biology and Evolution* **10**: 2432–2442.
Voelkerding KV, Dames SA, Durtschi JD. 2009. Next-generation sequencing: From basic research to diagnostics. *Clinical Chemistry* **55**: 641–658.
Waldrop MM. 2019. News Feature: What are the limits of deep learning? *Proceedings of the National Academy of Sciences* **116**: 1074–1077.
Wang X, Cairns MJ. 2013. Gene set enrichment analysis of RNA-Seq data: integrating differential expression and splicing. *BMC Bioinformatics* **14**: S16.

Wang Z, Chen Y, Li Y. **2004**. A brief review of computational gene prediction methods. *Genomics, Proteomics & Bioinformatics* **2**: 216–221.
Wang Y, Ding Y, Liu J-Y. **2016**. Identification and profiling of microRNAs expressed in elongating cotton fibers using small RNA deep sequencing. *Frontiers in Plant Science* **7**.
Wang Z, Gerstein M, Snyder M. **2009a**. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics* **10**: 57–63.
Wang Y, Juranek S, Li H, *et al.* 2009b. Nucleation, propagation and cleavage of target RNAs in Ago silencing complexes. *Nature* **461**: 754–761.
Wang J, Mei J, Ren G. **2019a**. Plant microRNAs: Biogenesis, homeostasis, and degradation. *Frontiers in Plant Science* **10**.
Wang F, Polydore S, Axtell MJ. **2015**. More than meets the eye? Factors that affect target selection by plant miRNAs and heterochromatic siRNAs. *Current Opinion in Plant Biology* **27**: 118–124.
Wang Y, Rashid MAR, Li X, *et al.* 2019b. Collection and evaluation of genetic diversity and population structure of potato landraces and varieties in china. *Frontiers in Plant Science* **10**.
Wang L, Song X, Gu L, *et al.* 2013. NOT2 proteins promote polymerase II–dependent transcription and interact with multiple microRNA biogenesis factors in *Arabidopsis*. *The Plant Cell* **25**: 715–727.
Wang R, Wang Z, Wang J, *et al.* 2019c. SpliceFinder: ab initio prediction of splice sites using convolutional neural network. *BMC Bioinformatics* **20**: 652.
Watson JD, Crick FH. **1953**. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**: 737–738.
Watt D. **2004**. Programming language design concepts. Wiley.
Weiberg A, Wang M, Lin F-M, *et al.* 2013. Fungal small RNAs suppress plant immunity by hijacking host RNA interference pathways. *Science* **342**: 118–123.
Willmann MR, Endres MW, Cook RT, *et al.* 2011. The functions of RNA-dependent RNA polymerases in *Arabidopsis*. *The Arabidopsis Book* **2011**.
Wu X, Kim T-K, Baxter D, *et al.* 2017. sRNAnalyzer—a flexible and customizable small RNA sequencing data analysis pipeline. *Nucleic Acids Research* **45**: 12140–12151.
Xie Z, Allen E, Fahlgren N, *et al.* 2005. Expression of Arabidopsis MIRNA Genes. *Plant Physiology* **138**: 2145–2154.
Yang SW, Chen H-Y, Yang J, *et al.* 2010. Structure of *Arabidopsis* HYPONASTIC LEAVES1 and its molecular implications for miRNA processing. *Structure* **18**: 594–605.
Yang Z, Rannala B. **2012**. Molecular phylogenetics: principles and practice. *Nature Reviews Genetics* **13**: 303–314.
Young MD, Wakefield MJ, Smyth GK, *et al.* 2010. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biology* **11**: R14.
Zhai J, Arikit S, Simon SA, *et al.* 2014. Rapid construction of parallel analysis of RNA end (PARE) libraries for Illumina sequencing. *Methods* **67**: 84–90.
Zhai J, Jeong D-H, De Paoli E, *et al.* 2011. MicroRNAs as master regulators of the plant *NB-LRR* defense gene family via the production of phased, *trans*-acting siRNAs. *Genes & Development* **25**: 2540–2553.
Zhang X, Niu D, Carbonell A, *et al.* 2014. ARGONAUTE PIWI domain and microRNA duplex structure regulate small RNA sorting in *Arabidopsis*. *Nature Communications* **5**: 5468.
Zhang H, Xia R, Meyers BC, *et al.* 2015. Evolution, functions, and mysteries of plant ARGONAUTE proteins. *Current Opinion in Plant Biology* **27**: 84–90.
Zhao T, Tao X, Li M, *et al.* 2020. Role of phasiRNAs from two distinct phasing frames of *GhMYB2* loci in *cis*- gene regulation in the cotton genome. *BMC Plant Biology* **20**: 219.
Zhou X, Shen X-X, Hittinger CT, *et al.* 2018. Evaluating fast maximum likelihood-based phylogenetic programs using empirical phylogenomic data sets. *Molecular Biology and Evolution* **35**: 486–503.
Zong J, Yao X, Yin J, *et al.* 2009. Evolution of the RNA-dependent RNA polymerase (RdRP) genes: Duplications and possible losses before and after the divergence of major eukaryotic groups. *Gene* **447**: 29–39.

Popular science summary

Potato is, with a worldwide production of 368 million tons, the most produced non-grain food crop in the world. To feed a constantly growing world population it is important to get a rich harvest. Hence, breeders attempt to breed potato with higher yields and defence mechanisms against pathogens. Potato late blight is the world's worst potato disease causing the world losses estimated to 4.8 billion € yearly, which represents approximately 15% of the total yield. Late blight is caused by the water mould *P. infestans*, which effectively adapts to pesticides and host-based resistance. Hence, it is important to investigate further in interactions between potato and *P. infestans*. Small RNA (sRNA) are short RNA molecules that, instead of getting translated into proteins, bind to Argonaute protein (AGOs) in a complex. The sRNA guides the complex to an mRNA, which is silenced by cleavage or blocked to inhibit the production of protein. sRNAs are reported to spread between host and pathogen to contribute in the infection process. To explore the sRNA world we first needed to investigate in the AGO content of potato. Through phylogenetic analysis we characterized several AGOs in closely related species to potato and determined the number of AGOs in potato to 14. Potato AGO15 was one of the AGOs unique in potato and related species. Further, we investigated in the sRNA binding to *P. infestans* AGO1 during infection. We found that a *P. infestans* sRNA of type microRNA entered potato and silenced the expression of a potato gene. The gene was deduced very important for potato defence as over-expression of the gene resulted in a very resilient plant and down-regulation of the gene resulted in increased susceptibility. To further examine sRNA silencing-events in potato and *P. infestans*, degradome sequencing was performed. A degradome sequencing comprises sequencing of the pieces of sRNA-cleaved mRNAs. Through combination of sRNA- and degradome data with cleavage rules, it is possible to deduce which sRNA guides the cleavage of what mRNA. Unfortunately, the degradome data contained noise, which forced us to develop a deep learning-based computer tool with functionality to distinguish between noise and true cleavages. In total, the tool identified almost 5000 cleavages in potato and *P. infestans*. One of the greatest groups of targets was resistance genes, a group we continue to investigate in.

Populärvetenskaplig sammanfattning

Potatis är tillsammans med vete, ris och majs en av de fyra mest producerade grödorna, sett ur ett världsproduktionsperspektiv. Potatisodlingen har en lång tradition i Europa och Nordamerika. I dagsläget ökar den framförallt i Kina men även i Afrika ökar produktionsvolymen. Potatis är en gröda som kan användas inom många områden förutom livsmedel vilket är en faktor bakom det allt mer ökande intresset. Potatis som många andra grödor kan drabbas av många sjukdomsalstrande organismer och växtskadegörare. *Phytophthora infestas* är en algsvamp som orsakar sjukdomen bladmögel när blasten angrips och när den har spridit sig till knölarna går angreppen under namnet brunröta. En total skördeförlust kan uppkomma om knölarna angrips. Två alternativa kontrollmetoder används idag och ofta i kombination. Kemisk bekämpning på blasten för att undvika utveckling av brunröta och resistensförädling. Den här patogenen har en speciell organisation av sitt genom där det stora antalet gener som utnyttjas för infektion (ca 1000) är insprängda bland transposoner som har förmåga att skapa förändringar dvs mutationer. Då miljoner sporer kan produceras varje vecka i ett fält sprids mutationerna snabbt i en population. Detta är bakgrunden till att denna algsvamp är så framgångsrik sjukdomsalstrare. Anpassning till kemisk bekämpning samt nya resistenta potatissorter sker mycket snabbt. Vi försöker förstå de bakomliggande molekylära mekanismerna vid denna sjukdomsutveckling. En typ av molekyl involverad i bland annat växters immunförsvar kallas små RNA (sRNA). sRNA är korta RNA-molekyler som binder till så kallade Argonaut-proteiner (AGOs) i komplex istället för att själva översättas till protein. sRNAt leder komplexet till ett mRNA, vilket ”tystas” genom klyvning eller blockering och förhindrar bildandet av protein. sRNAn har rapporterats kunna spridas mellan växt och patogen för att bidra i infektionsprocessen. För att undersöka sRNA-världen behövde vi först undersöka vilka sorts AGOs som finns i potatis. Fylogenetisk släktskapsanalys påvisade att arter i potatisfamiljen har tappat och skapat nya *AGO*-gener under evolutionens gång samt att dagens potatis har 14 AGOs. En av dessa är AGO15 som är unik för potatis och dess nära släktingar. AGO15 är högt uttryckt under angrepp av *P. infestans*. Därefter undersökte vi vilka sRNAs som associeras till *P. infestans* AGO1 under infektion. Vi hittade i den studien ett *P. infestans* sRNA av typen mikroRNA, som

tystar en potatisgen vilket vi i uppföljande analyser kunde bekräfta. Den här genen visade sig vara mycket viktig för potatisens försvar då överuttryck av genen ledde till en mycket motståndskraftig växt och nedreglering av genen gav upphov till en mycket sjuk potatisplanta. För att vidare undersöka sRNA-nedtystningar i potatis och *P. infestans* genomfördes en degradomsekvensering, det vill säga, en sekvensering av de mRNA-fragment som klyvs av sRNA. Genom att kombinera sRNA- och degradomdata med statistiska klyvningsregler kan man härleda vilket sRNA som matchar klyvning av individuella mRNA sekvenser. Den första analysen innehöll en hög frekvens falska sRNA-mRNA par. Därför skapade vi ett nytt bioinformatiskt verktyg (smartPARE) som kunde särskilja störningar från riktiga klyvningar. Verktyget identifierade totalt nästan 5000 klyvningar i olika gensekvenser i potatis och *P. infestans*.

Acknowledgements

Like leaves, we fall from branches. Caught by the wind we travel through the air. Some cruise with a gust. Some are pushed to the ground. On the way we bump into each other. We hitch for some time to later disentangle.

When I started my bachelor studies, it was not my intention to pursue all the way to a graduate degree. I did not even intend to work with plants. However, along the road the horizon changed. Experiences and encounters made me alter my mind and suddenly I was mucking up my hands with soil and filling my nostrils with the *Phytophthora* scent.

I want to thank **everyone** that I have bumped into during my PhD education. You have all in your particular way enriched these years and just because your name might not be mentioned in this text, it does not mean your name is not sealed in my heart.

I would like to express my deepest appreciation to **Christina Dixelius** that took me in as PhD student and has supervised me throughout these years. Thanks to you, I have challenged myself and developed my scientific knowledge vigorously. Thanks also for your extremely hard work and for always being there for me!

I wish to thank my co-supervisors **Anna Åsman**, **Germán Martinez-Arias** and **Georgios Tzelepis** for all being helpful and positive whenever I needed advice**.** More specifically**,** I want to express gratitude to **Anna** for teaching me almost everything I know about wet-lab work. It must have been challenging to convey the bioinformatician with shaking hands understand all those methods. I have been very fortunate to have you in the building, also recent years whenever having *Phytophthora* or sRNA related questions. Thanks **Germán** for motivating us to pursue with the degradome sequencing and for making the related libraries. This project has been really cool! Thanks **Georgios** for teaching me the rest I know about wet-lab work and for being my office mate the longest time of anyone throughout

these years. Your spirit also helped me digest the exposure to extensive levels of science at conferences.

I am greatful to **Johan Fogelqvist** for being my co-supervisor during my first year and providing me with invaluable sRNA-bioinformatics knowledge.

Many thanks to my 50%, 75% and 18 week evaluation committee members: **Anders Hafrén, Andrea Hinas, Jens Sundström, Juan Santos-González, Johan Reimegård and Pär Ingvarsson.**

I wish to thank all, not already mentioned, past (**Anna T, Linnea, Louise, Ravi, Sara and Zhen**) and present (**Anushka, Fredrik, Shailja, Shridhar, Suzana, Tua and Xinyi**) "Dixan's group" members that I have encountered. The collaborations and friendships have been vital to reach this point.

Thanks **Björn** and **Christian** for the computer support! Thanks to **Fredric**, **Kathrin, Per** and **Urban** for helping with the plants and taking good care of the growth facilities**.** Thanks **Anna S**, **Lotta**, **Monica** and **Qing** for the help with administrative stuff.

Thanks a million to the rest of the **Plant Biology department** for providing such a great working environment. I have had such nice chats with so many of you in front of the coffee machine before the start of covid-19 and I deeply regret not knowing newcomers better because of the social distancing.

To all my **musician friends**: Thank you for the music and all the joy! You have been the candle when tough experiments have shaded my mind. A special thanks to my fellow band members in **Caolmhar** and **Pelle med två** and for the nice jams at the **Irish sessions** and at **Vdala**.

To **my other friends:** Thanks a million for all the interesting conversations, small chats and greetings! Sometimes, the simplest things mean the most.

Thanks **Annika** for all support and love! You are a true blessing to me! I wish to also send my appreciation to your **family** for all the enjoyable time spent together.

Thanks to **mamms** and **papps** for teaching me that practice makes perfect and that patience is a virtue. These words of wisdom have been vital for accomplishing my PhD education. Thanks also for carrying me through the tough times, as easily now as when I was a baby. Thanks **Seb** and **Sandro** for the nice celebrations and conversations! I am also grateful towards all my other **relatives** that have been inspiring and supportive throughout time.

With huge gratitude,

Acta Universitatis Agriculturae Sueciae

Doctoral Thesis No. 2021:14

During *Phytophthora infestans* infection of potato, small RNA fingerprints in both species are altered. small RNAs are loaded into Argonauts which in a larger protein complex (RISC) silence gene expression by sequence complementation. This thesis cover analysis of Argonauts in the potato family, and extensive small RNA sequence profiling. Precursor and target sites were predicted in both organisms. A deep learning-based R package, smartPARE, was developed to refine the data.

Kristian Persson Hodén received his graduate education at Umeå University, Umeå, Sweden, and obtained a M.Sc. in Engineering with specialisation in Engineering Biotechnology.

Acta Universitatis Agriculturae Sueciae presents doctoral theses from the Swedish University of Agricultural Sciences (SLU).

SLU generates knowledge for the sustainable use of biological natural resources. Research, education, extension, as well as environmental monitoring and assessment are used to achieve this goal.

Online publication of thesis summary: http://pub.epsilon.slu.se/

ISSN 1652-6880
ISBN (print version) 978-91-7760-706-9
ISBN (electronic version) 978-91-7760-707-6