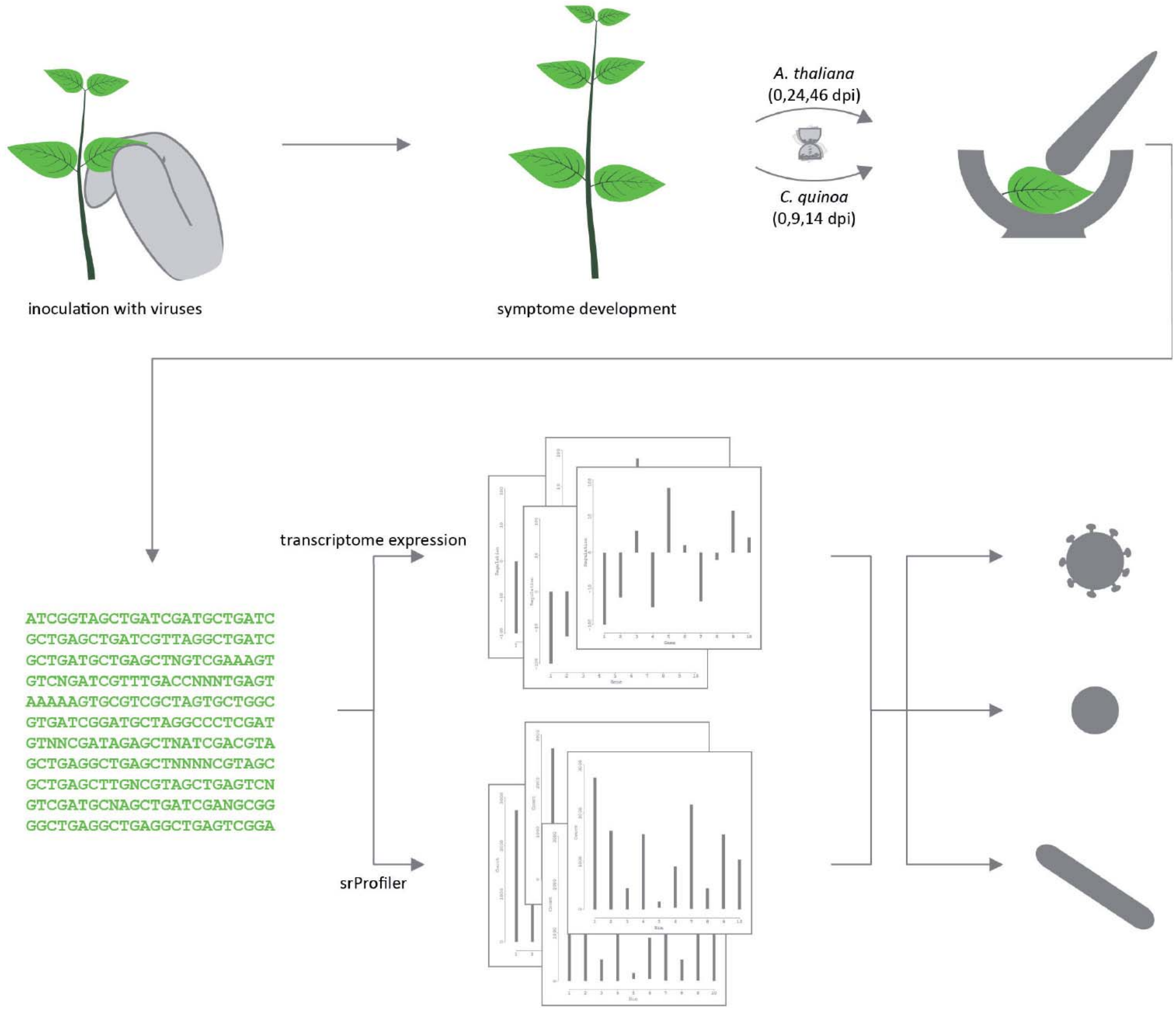
Berliner ökophysiologische und phytomedizinische Schriften

Steffen Pallarz

Data driven classification of host-plant response (virus-plant)

Band 44



Cuvillier Verlag Göttingen
Internationaler wissenschaftlicher Fachverlag

Berliner ökophysiologische
und phytomedizinische Schriften

Hrsg. von Christian Ulrichs und Carmen Büttner

Lebenswissenschaftliche Fakultät,
Humboldt-Universität zu Berlin

Band 44

Hrsg. von

Dr. Susanne von Bargen
Humboldt-Universität zu Berlin

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

Titel der Arbeit:
Data driven classification of host-plant response (virus-plant)

DISSERTATION
zur Erlangung des akademischen Grades

Doctor rerum naturalium
(Dr. rer. nat.)

eingereicht an der
Lebenswissenschaftlichen Fakultät der Humboldt-Universität zu Berlin

von
M.Sc., Steffen Pallarz

14.08.1984, Strausberg

Präsidentin/Präsident
der Humboldt-Universität zu Berlin

Prof. Dr.-Ing. Dr. Sabine Kunst

Dekanin/Dekan der Lebenswissenschaftlichen Fakultät
der Humboldt-Universität zu Berlin

Prof. Dr. Bernhard Grimm

Gutachter/innen

1. Prof. Dr. Carmen Büttner
2. Prof. Dr. Philipp Franken
3. Prof. Dr. Peter Beyerlein

Tag der mündlichen Prüfung: 30.01.2018

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der
Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet
über http://dnb.d-nb.de abrufbar.
1. Aufl. - Göttingen: Cuvillier, 2018
Zugl.: Berlin, Humboldt-Universität, Diss., 2018

© CUVILLIER VERLAG, Göttingen 2018
Nonnenstieg 8, 37075 Göttingen
Telefon: 0551-54724-0
Telefax: 0551-54724-21
www.cuvillier.de

Alle Rechte vorbehalten. Ohne ausdrückliche Genehmigung des Verlages ist
es nicht gestattet, das Buch oder Teile daraus auf fotomechanischem Weg
(Fotokopie, Mikrokopie) zu vervielfältigen.
1. Auflage, 2018
Gedruckt auf umweltfreundlichem, säurefreiem Papier aus nachhaltiger Forstwirtschaft.

ISBN 978-3-7369-9731-8
eISBN 978-3-7369-8731-9

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

Abstract

Background The amount of information that can be drawn from one NGS run is much greater than the binary answer provided by one ELISA test. However, the comparatively complex analyses required, especially when a reference is not applicable, result in limited usability in a routine diagnostic environment. With ever decreasing costs sequencing is becoming an ever more important technique in diagnostic medicine, especially when faced with unknown pathogens or diseases, thus increasing the demand for a robust and easy-to-use diagnostic tool incorporating the huge amount of data NGS provides.

Methods 48 plant specimen of two distinct species were each infected with one of three different viruses, sequenced and subsequently analyzed to discover whether the sample had been infected successfully, what the underlying pathogen had been and what species the plant belonged to. The analysis was performed using the standard approach, consisting of alignment to a reference genome, transcriptome expression analysis and classifying the samples using the most differentially expressed transcipts. A second analysis was performed, using a profiler approach, which is entirely independent of reference information. Each sample was transformed into a profile and the most different bins among the profiles were used for classification.

Results Comparing the two sets of results provided by the two distinct approaches it could be shown that the novel approach performed equally well and for certain classifications even better than the standard approach. It is a generic method independent of the host or pathogen in question and does not require assembly, alignment and subsequent expression analysis. Moreover it is substantially faster and requires less storage space than the standard approach. It is a fast and versatile tool which is capable of providing the same information as NGS, while being applicable as quickly and easily as ELISA in a routine environment without the need for substantial bioin-

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

formatic experience. It thus combines the robustness, reproducibility and generics of NGS with the usability and ease-of-operation of a routine diagnostic tool.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

Zusammenfassung

Hintergrund Die Menge an Informationen, die mit einem NGS Lauf erzeugt werden kann, ist sehr viel größer als die von ELISA Tests ausgegebene binäre Antwort. Allerdings schränken die im Vergleich sehr viel komplexeren Auswertungen der Daten und die um bis zu 50 Mal höheren Kosten pro Lauf, insbesondere wenn eine Referenz nicht anwendbar ist, die Benutzbarkeit der Technologie in einem diagnostischen Routinebetrieb stark ein. Aufgrund der stetig fallenden Kosten erhält das Verfahren der NGS Analyse einen immer wichtigeren Stellenwert im Bereich der medizinischen Diagnostik, insbesondere bei der Konfrontation mit unbekannten Pathogenen und Krankheiten. Diese steigende Nachfrage erhöht die Notwendigkeit eines einfach anzuwendenden und robusten diagnostischen Werkzeugs, das die große Datenmengen, die durch NGS erzeugt werden, verarbeitet.

Methoden Es wurden 48 Pflanzen aus zwei verschiedenen Spezies mit jeweils einem von drei unterschiedlichen Viren infiziert, sequenziert und anschließend analysiert um jeweils zu erkennen, ob die Infizierung erfolgreich war, welcher Erreger verwendet wurde und welcher Spezies die einzelnen Pflanzen angehörten. Die Analyse wurde nach dem Standardablauf durchgeführt, d. h. die Daten wurden gegen eine Referenz aligned, die Transkriptomexpression berechnet und die Proben unter Verwendung der am unterschiedlichsten exprimierten Transkripte klassifiziert. Ein zweites Analyseverfahren basierend auf der Erzeugung von Profilen direkt aus den NGS Daten ohne die Verwendung von Referenzen wurde ebenfalls benutzt. Aus diesen Profilen wurden die sich am stärksten unterscheidenden Bins innerhalb der Proben verglichen, um eine Klassifizierung durchzuführen.

Ergebnisse Ein Vergleich der Ergebnisse, die durch die unterschiedlichen Verfahren gewonnen wurden, zeigte, dass, verglichen mit dem Standardverfahren, das neue Verfahren ebenso gute und in einigen Fällen sogar bessere Klassifizierungen erzeugte. Dieser Profilansatz ist eine generische Methode, die unabhängig von dem

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

Wirt und dem Erreger funktioniert und keine Assemblierung benötigt. Dieser Ansatz ist sehr viel schneller und hat geringere Anforderungen an die Infrastruktur als der Standardansatz. Diese Eigenschaften machen den neuen Ansatz zu einem schnellen und vielseitigen Werkzeug, das die gleichen Informationen wie NGS produzieren kann und dabei im Aufwandsumfang mit ELISA vergleichbar ist. Es verbindet die Reproduzierbarkeit, die Informationsvielfalt und die generischen Eigenschaften von NGS mit der Verwendbarkeit und der einfachen Handhabung von ELISA in einem einzigen diagnostischen Werkzeug.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

Contents

Abstract I

Zusammenfassung III

1 Introduction 1

2 Materials and Methods 7
- 2.1 Plant-Viruses . . . 8
- 2.2 Model Plants . . . 11
- 2.3 Mechanical Inoculation . . . 12
- 2.4 ELISA . . . 14
- 2.5 Sequencing . . . 17
- 2.6 Transcriptome Analysis . . . 20
- 2.7 Data Analysis . . . 20
- 2.8 Pathogen Alignment . . . 25
- 2.9 n-Gram Based Sequence Profiling for Alignment Free Transcriptome Analysis . . . 25
- 2.10 Statistical Analysis . . . 26
  - 2.10.1 Welch’s unequal variances t-test . . . 27
  - 2.10.2 ANOVA-Test . . . 29
  - 2.10.3 Principal Component Analysis . . . 31

3 Results 33
- 3.1 ELISA-detection . . . 42
  - 3.1.1 DAS-ELISA for the detection of ArMV . . . 42
  - 3.1.2 DAS-ELISA for the detection of TSWV . . . 42
  - 3.1.3 DAS-ELISA for the detection of CLRV . . . 46
  - 3.1.4 TAS-ELISA for the detection of TSWV . . . 46
- 3.2 Sequencing . . . 50
- 3.3 Bioinformatical Analysis . . . 50
  - 3.3.1 Transcriptome Analysis . . . 50
  - 3.3.2 n-Gram Based Sequence Profiling . . . 60
- 3.4 Validation by means of Pathogen Alignment . . . 70

4 Discussion 72

A Appendix i
- A.1 Sequencing . . . ii
- A.2 Pathogen Alignment . . . iii

Eidesstattliche Erklärung

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

List of Tables

1 Viral isolates 9
2 Virus list 10
3 Hosts for propagation 10
4 Sample list *Chenopodium quinoa* 12
5 Sample list *Arabidopsis thaliana* 13
6 Classifiers 24
7 STRESS Interpretation 32

8 Classifier specificity, sensitivity and accuracy 79

9 Alignment results for *Arabidopsis thaliana* ii
10 Pathogen Alignment iii

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

List of Figures

1 Experimental design 7
2 Flowchart of the ELISA workflow 16
3 Flowchart of the data analysis in *Arabidopsis thaliana* 23
4 Profile transformation 26
5 Statistical Workflow 28

6 *Arabidopsis thaliana* control samples 34
7 *Chenopodium quinoa* control samples 35
8 *Arabidopsis thaliana* CLRV infected samples 36
9 *Chenopodium quinoa* CLRV infected samples 37
10 *Arabidopsis thaliana* ArMV infected samples 38
11 *Chenopodium quinoa* ArMV infected samples 39
12 *Arabidopsis thaliana* TSWV infected samples 40
13 *Chenopodium quinoa* TSWV infected samples 41
14 DAS ELISA results for ArMV propagation 43
15 DAS ELISA results for ArMV 44
16 DAS ELISA results for TSWV 45
17 DAS ELISA results for CLRV 47
18 TAS ELISA results for TSWV propagation 48
19 TAS ELISA results for TSWV 49
20 PCA expression A 53
21 PCA expression S1 54
22 PCA expression S2 55
23 PCA expression S3 56
24 PCA expression B1 57
25 PCA expression B2 58
26 PCA expression B4 59
27 PCA profile A 63
28 PCA profile S1 64
29 PCA profile S2 65
30 PCA profile S3 66
31 PCA profile B1 67
32 PCA profile B2 68
33 PCA profile B4 69
34 Viral load over time 71

35 Temporal distance of harvesting dates 76
36 Amount of entities covered by samples 80

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

1 Introduction

In the areas of phytomedicine and medicine at large the diagnosis of viral infections is extremely important. Over the last few years two methods emerged as gold standard for the diagnosis of viral infection, namely enzyme-linked-immunosorbent assay (ELISA) and quantitative Polymerase Chain Reaction (qPCR) (Boonham *et al.*, 2014). For most applications those methods are well suited, however, in some cases inherent shortcomings in both methods call for a different approach, to wit the highly versatile method of Next-Generation-Sequencing (NGS) (Boonham *et al.*, 2014).

ELISA is a very specific technique based on an antigen-antibody bond (Engvall *et al.*, 1971; Weemen *et al.*, 1971). The assays are very robust and require only some specific equipment. The sample preparation consists of little more than the homogenization of the sample in buffer in order to bring the antigen or antibody into solution. The reagents used are specific to the pathogen and are developed prior using an independent process (Boonham *et al.*, 2014). The signals produced by ELISA, as read by a plate reader, can be interpreted as a yes-no answer depending on the signal strength in comparison to an afore calculated cutoff (BIOREBA AG, 2014). As long as the number of possible viruses infecting a sample is very limited and the viruses are known, thus the corresponding specific reagents can be acquired, ELISA is the diagnostic method of choice (Büttner *et al.*, 2013). However, the method is not ideal if the number of possible viruses is great and therefore the amount of tests required to find the pathogen is very high. Also, if the infecting virus has not been discovered before, new specific reagents must be designed which is an expensive and time consuming process and requires a very specific laboratory (Boonham *et al.*, 2014; Büttner *et al.*, 2013). When dealing with viruses which have a high mutation rate, possibly resulting in quasi-species, the high specificity of ELISA can result in false negative results even for known viruses (Adams *et al.*, 2013).

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

qPCR requires nucleic acid (xNA) and uses pathogen specific oligo nucleotides (primers) to massively amplify very specific fragments of the input xNA, for instance a part of the pathogens genome (Khan *et al.*, 2001). The sample preparation is more complex compared to ELISA since purified deoxyribonucleic acid (DNA) or ribonucleic acid (RNA) is required. The amount of xNA is continually measured throughout the process of Polymerase Chain Reaction (PCR) (Boonham *et al.*, 2014). If the amount of xNA increases, the targeted material, which, confined by the primers, is amplified during the PCR cycles, must have been part of the input. Like ELISA the result of the qPCR is well interpretable and straight forward (either the material had been amplified or not). qPCR is a lot more sensitive than ELISA (Khan *et al.*, 2001). Since the method also requires pathogen specific reagents, qPCR is not ideal for pathogens which have not been discovered before. The primer design and construction is however much less expensive and less time consuming than the development of pathogen specific ELISA reagents (Boonham *et al.*, 2014). While the high specificity of the method is an advantage in most cases, like ELISA it can lead to false negative measurements for known viruses, if those are prone to mutations and the development of quasi-species (Adams *et al.*, 2013).

NGS encompasses a class of methods that is becoming ever more prominent as exploratory and diagnostic tool (Adams *et al.*, 2009; Boonham *et al.*, 2014; Hadidi *et al.*, 2016; Capobianchi *et al.*, 2013). These methods use different techniques to sequence the entire input xNA and provide the resulting sequences as a file to the analyst. The great advantage of NGS is that no prior knowledge about the pathogen is required (Selvarajan *et al.*, 2016). Since no pathogen-specific reagents are needed, NGS is a completely generic process. It can be used to discover viruses and even quasi-species of known viruses that ELISA and qPCR cannot discover due to their high specificity (Capobianchi *et al.*, 2013; Adams *et al.*, 2013). Moreover, NGS can be used to describe, assemble and annotate newly discovered pathogens (Prabha *et al.*, 2013). The sample preparation is comparable to qPCR, since purified DNA is required as input for most NGS based methods. While currently NGS

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

is expensive by comparison, the price per sequenced base is rapidly declining and will reach competitive prices in the near future (Boonham *et al.*, 2014). Since NGS was first used in phytomedicine in 2009 (Adams *et al.*, 2009; Al Rwahnih *et al.*, 2009; Kreuze *et al.*, 2009) its importance increased substantially, mainly in the area of pathogen discovery (Yanagisawa *et al.*, 2016; Barzon *et al.*, 2011). However, NGS is not yet used as a routine analysis tool like ELISA or qPCR. This is due to the complex data analysis required to interpret the results (Boonham *et al.*, 2014). While the analysis is independent of the pathogen, it is highly dependent on the host reference (complete genome) (Gogol-Döring *et al.*, 2012). Using a reference, the host specific information can be stripped from the data and the remaining fragments can be used to assemble and discover the pathogen without contaminations from the host (Barzon *et al.*, 2011; Studholme *et al.*, 2011). Conversely, the abundance of known pathogens within the host can be analyzed using references of the pathogens (Nagano *et al.*, 2015). It is also possible to measure the hosts response to a pathogen rather than the existence of said pathogen. This is useful when no information about the pathogen can be provided or if there is uncertainty of whether specific symptoms are actually caused by a pathogen, furthermore, this method enables the researcher to analyze the molecular mechanisms at work within the host (Chen *et al.*, 2016). Analyzing the host response to infection or disease can be done very well using the transcriptome expression, providing information about the expression of genes and, using a time series, the up and down regulation of specific genes which allows the analysis of pathway modifications (Wang *et al.*, 2009). This expression analysis also requires the use of a host specific reference and transcriptome annotation (host genome annotated with start and end positions of genes, exons, introns). If those reference informations are not accessible or do not exist, which is the case for almost every plant species (Yates *et al.*, 2016), those kinds of analyses, while still being possible, become much more time consuming. The required references must first be created. This process uses the information of the sequenc-

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

ing results and assembles a probable reference by constructing ever longer fragments into contigs (larger fragments constructed from overlapping fragments) and super contigs (larger contigs constructed from overlapping contigs) (Baker, 2012). A very high coverage (fragments covering a specific location) is needed to produce a good and trustworthy reference, which increases the cost of sequencing significantly (Sims *et al.*, 2014). The results of an analysis, being performed upon a newly assembled reference, are not reproducible by another researcher in a straight forward manner because any newly constructed reference is unique and in part dependent on the parameters used for the assembling algorithm (Baker, 2012). The eventual stability of any reference is the result of the collaboration of multiple groups and the thorough scrutiny by the scientific community.
Using a good reference and annotation, transcriptome analyses are based on multiple steps offering many possibilities to produce differing results. The alignment (mapping the sequencing results to the reference) can be run with different parameters resulting in fewer but qualitatively better results (Langmead *et al.*, 2009; 2012; Cox, 2007; Li *et al.*, 2009). The transcriptome analyses can be performed using only fragments aligned to a single location, or, in order to increase the pool of fragments, adding those aligned to multiple locations. During the expression analyses, the analyst has to decide whether a fragment is counted twice or only in part if it is located in two genes. Those examples show the complexity of the analyses and why it should be run and interpreted by an experienced bioinformatician (Boonham *et al.*, 2014).

This work proposes a novel approach, which reduces the complexity of NGS data analysis by removing multiple, otherwise necessary, steps from the analysis workflow. It is based on host response rather than the existence of pathogen RNA. The novel approach utilizes pattern classification in order to reduce the complexity within the data and answer multiple independent questions simultaneously. It does not require a reference for the host or the pathogen. The use or assembly of a transcriptome is not necessary. This has been accomplished by utilizing an alignment free

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

method (Song *et al.*, 2013; Bonham-Carter *et al.*, 2013) based on feature-frequency-profiles (Sims *et al.*, 2009), an n-gram (subsequence of size n) based approach, resulting in representative profiles which are used for classification. If the pathogen cannot be classified directly, on account of it being undiscovered of yet, the data can be used to assemble the new pathogen directly without the need for further wet-lab-work. A strong automation reduces the need for significant bioinformatic expertise and allows a competent lab technician to use the software in a routine environment. This offers a robustness and ease-of-operation comparable to ELISA or qPCR while offering the advantages of NGS in terms of amount and diversity of information and generic character.

This novel methods performance and accuracy is compared to a transcriptome analysis following a common workflow (Gogol-Döring *et al.*, 2012). An experiment has been performed, whereby 36 plants where mechanically inoculated with one of three distinct viruses and 12 plants served as control samples. All samples were sequenced resulting in the input files for the NGS based analyses. ELISA tests were performed to discover the infection state of each sample and alignments using the pathogen references were run to measure the viral load in each sample. The results of the sequencing runs were classified using the novel method and independently a transcriptome approach. The resulting classifications are compared in regards to similarity and accuracy given the results of the validation tests (ELISA and pathogen alignment).

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

2 Materials and Methods

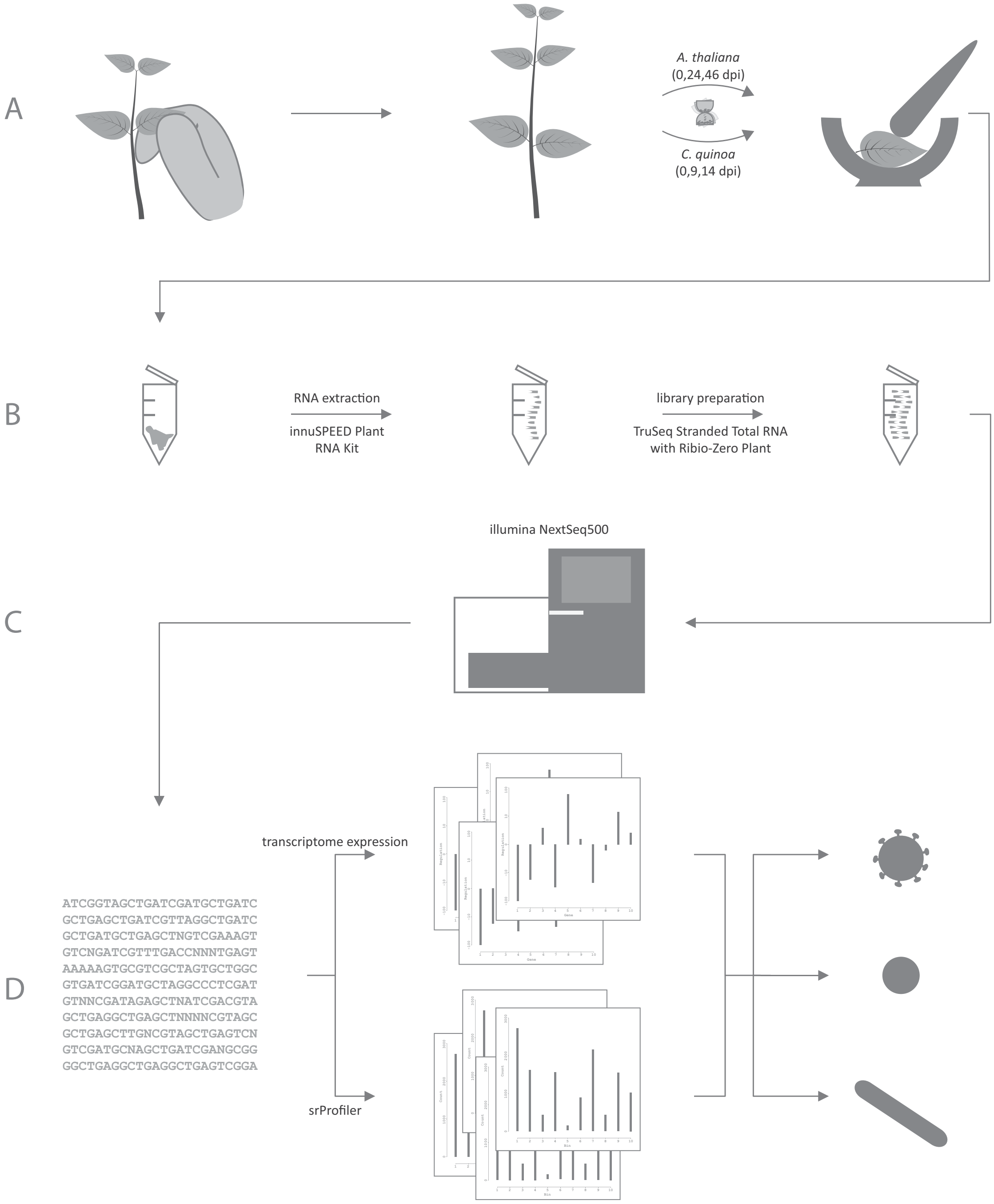


Figure 1: The experimental design is shown from inoculation and harvesting (A) over sample preparation for sequencing (B), sequencing (C) and subsequent data analysis resulting in profiles which classify the pathogens (D).

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

After an initial growth phase the sample plants were inoculated, cultivated for three different time spans and finally harvested (figure 1.A). The plant material was prepared to arrive at complementary DNA (cDNA) libraries (figure 1.B). Those were sequenced (figure 1.C). Using the generated reads, two independent approaches were used to answer multiple questions, for instance which the infecting virus had been (figure 1.D).

2.1 Plant-Viruses

Arabis mosaic virus (ArMV), *Tomato spotted wilt virus* (TSWV) and *Cherry leaf roll virus* (CLRV) are the pathogens used in the scope of this work (tables 1 and 2).

ArMV is a positive single stranded (+ss) RNA virus and belongs to the genus *nepovirus*. It was first described in 1944 (Smith *et al.*, 1944). Schmelzer (1962) reported that 93 different plant species could be successfully infected with this virus. Its hosts include important crops, such as hemp, raspberry, strawberry, cucumber, lettuce and more. The genome organization is comprised of two +ss RNAs (3820 base pair (bp) and 7334 bp in size). The complete sequence was published in its current version by Wetzel *et al.* (2001; 2004).

CLRV also belongs to the positive single stranded RNA *nepoviruses*. Its impact was first described in 1933, however, it was first designated CLRV in 1955 (Posnette *et al.*, 1955). While its host range, spanning 36 different plant families (EFSA, 2014; Hadidi *et al.*, 2011), is more limited than that of ArMV, new hosts are discovered frequently. In 2007 symptoms typical for a CLRV infection have been observed in two birch species in Finland, Sweden and Norway, while the virus could be detected in Finland (Jalkanen *et al.*, 2007). Genetically CLRV is described to have a high variability on interhost as well as intrahost level (Hadidi *et al.*, 2011), in some cases leading to different strains of the virus within the same host (Rumbou *et al.*, 2016).

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

In 2012 the two RNA sequences of CLRV (isolate E395), being 6360 bp and 7918 bp long, were published by von Bargen *et al.* (2012).

TSWV, a negative single stranded (-ss) RNA *tospovirus*, was first described in 1930. It was the pathogen that caused a disease first described in 1915 as tomato spotted wilt, which in the years from 1915 to 1930 spread over all southern states of Australia, causing great economic losses. The virus has an enormous host range of over 900 different plant species, amongst which are important agricultural crops such as tomato, peanut, watermelon, zucchini, tobacco and more (Rupert, 1968; Sherwood *et al.*, 2000). Its genome organization consists of three RNA strands. The sequences of RNA L (large 8897 bp), RNA M (middle 4821 bp) and RNA S (small 2916 bp) were published in 1991 (De Haan *et al.*, 1991), 1992 (Kormelink *et al.*, 1992) and 1990 (De Haan *et al.*, 1990) respectively.

Table 1: The table shows the viral isolates and their respective origins (Menzel, 2016).

		Virus	
	ArMV	**TSWV**	**CLRV**
Isolate	E53152	PC-0182 (L3)	E395
Host	*Sambucus nigra*	*Nicotiana rustica*	*Rheum rhabarbarum*
Origin	Sweden	Bulgaria	Germany
Year of isolation	2012	1988	1987
Supplier	division Phytomedicine	Deutsche Sammlung von Mikroorganismen und Zellkulturen (DSMZ)	division Phytomedicine

Two of the three required virus species, ArMV and TSWV, needed to be propagated personally, CLRV was provided. Different host species were chosen recommended for virus propagation according to description of the respective virus as can be seen in table 3. These hosts were mechanically inoculated (section 2.3) with the virus species in question and then left to be infected with the virus. After 14 days, the virus was “harvested” by choosing leaves of the host that showed strong signs of infection.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

Table 2: The table shows a list of the three virus species used in this work (Adams *et al.*, 2006; Büttner *et al.*, 2013).

	Cherry leaf roll virus	***Arabis mosaic virus***	***Tomato spotted wilt virus***
Genus	Nepovirus	Nepovirus	Tospovirus
Abbreviation	CLRV	ArMV	TSWV
Symptoms	leaf patterns, blackline disease, chlorotic mosaic, ring patterns, leaf rolling, chlorotic ringspot, yellow vein netting, dieback, plant death	yellow dwarf, mosaic, yellow crinkle, stunt mottle, chlorotic stunt, stunting, necrosis, yellow net	stunting, chlorotic rings, necrotic rings, necrosis, seed discoloration
first described	1955 (Posnette *et al.*, 1955)	1944 (Smith *et al.*, 1944)	1930 (Rupert, 1968)
Genome organisation	two (+)ss RNAs (6360 bp and 7918bp)	two (+)ss RNAs (3820 bp and 7334 bp)	three (-)ss RNAs (8897 bp, 4821 bp and 2916 bp)

Table 3: The table lists the respective host plants which were used for the propagation of ArMV and TSWV respectively.

	Virus	
	ArMV	**TSWV**
Host	*N. benthamiana*, *C. amaranticolor*	*N. clevelandii*, *N. benthamiana*, *N. tabacum*, *C. amaranticolor*

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

2.2 Model Plants

The samples used in this scope are made up of two plant species, one being *Arabidopsis thaliana cv. Columbia* (36 samples) and the other being *Chenopodium quinoa* (12 samples). Those two plant species were chosen based upon their importance to the fields of molecular plant pathology and bioinformatics respectively. While *Chenopodium quinoa* is a model organism for plant pathology and virology, since it is a host to numerous viruses and shows successful infection by clearly visible lesions on its leaves (Hollings, 1959; 1972), its genome is poorly explored and thus no reference genome does exist as of yet. On the other hand, the complete genome of *Arabidopsis thaliana* is well documented (Rhee *et al.*, 2003), so much that it is considered to be the model organism for molecular genetic experiments in plants (Meinke *et al.*, 1998). However, *Arabidopsis thaliana* shows minimal to no signs of successful inoculation.

The *Chenopodium quinoa* plants were grown in a greenhouse subjected to long-day conditions (16 hours of light), 22°C and between 20% to 40% humidity. The six-leaf stage was chosen for inoculation. The *Arabidopsis thaliana* specimen were kept under short-day conditions (eight hours of light), 22°C during day time and 16°C during night time and a humidity between 20% to 40%. The inoculation was performed five weeks after sowing. The difference in cultivation between *Arabidopsis thaliana* and *Chenopodium quinoa* is due to the different growth and infection characteristics.

All 48 plants were divided into four groups, each comprised of nine *Arabidopsis thaliana* and three *Chenopodium quinoa* plants, and inoculated with ArMV, TSWV and CLRV respectively. The final fourth group is comprised of the control specimen, which were mock inoculated. After three different periods of time, stated as days post inoculation (dpi), three *Arabidopsis thaliana* (biological replicates) and one *Chenopodium quinoa* individual of each group was harvested, frozen in liquid nitrogen and stored at -80°C (Yockteng *et al.*, 2013). For *Arabidopsis thaliana* those

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

periods were 0, 24 and 46 dpi, for *Chenopodium quinoa* it was 0, 9 and 14 dpi. The samples and their respective infection agents are listed in tables 4 and 5. The bioinformatical methods used were developed on *Arabidopsis thaliana*, since it is the model organism for bioinformatics and has a well described reference genome. Therefore all dpi have three *Arabidopsis thaliana* samples to ensure a higher degree of accuracy during development.

2.3 Mechanical Inoculation

The process of mechanical inoculation is used to introduce a viral pathogen into a plant by spreading the inoculum, a solution of viral particles and buffer, over its leaves. The virus enters the damaged epidermal cells within which the replication cycle of the virus starts (Dijkstra *et al.*, 1998).

In this work the respective virus species were provided by means of infected leaves. These leaves were mechanically homogenized in 10 ml of inoculation buffer (0.1 M $NaPO_4$, 0.2% Na_2SO_3, 2% Polyvinylpyrrolidon, pH 7.0) mixed with 0.1 g celite as an abrasive. The solution was finally rubbed on to the sample plants (tables 4 and 5).

Table 4: A table listing the *Chenopodium quinoa* samples, showing the sample name, the virus used during inoculation and the isolate of the virus.

ID	Species	Virus	Isolate
X01	*Chenopodium quinoa*	mock	none
X02	*Chenopodium quinoa*	mock	none
X03	*Chenopodium quinoa*	mock	none
A01	*Chenopodium quinoa*	ArMV	E53152
A02	*Chenopodium quinoa*	ArMV	E53152
A03	*Chenopodium quinoa*	ArMV	E53152
T01	*Chenopodium quinoa*	TSWV	E53460
T02	*Chenopodium quinoa*	TSWV	E53460
T03	*Chenopodium quinoa*	TSWV	E53460
C01	*Chenopodium quinoa*	CLRV	E395
C02	*Chenopodium quinoa*	CLRV	E395
C03	*Chenopodium quinoa*	CLRV	E395

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

Table 5: The list of *Arabidopsis thaliana* samples is displayed in this table, showing the sample name, the virus used during inoculation and the isolate of the virus.

ID	Species	Virus	Isolate
X01.1	*Arabidopsis thaliana*	mock	none
X01.2	*Arabidopsis thaliana*	mock	none
X01.3	*Arabidopsis thaliana*	mock	none
X02.1	*Arabidopsis thaliana*	mock	none
X02.2	*Arabidopsis thaliana*	mock	none
X02.3	*Arabidopsis thaliana*	mock	none
X03.1	*Arabidopsis thaliana*	mock	none
X03.2	*Arabidopsis thaliana*	mock	none
X03.3	*Arabidopsis thaliana*	mock	none
A01.1	*Arabidopsis thaliana*	ArMV	E53152
A01.2	*Arabidopsis thaliana*	ArMV	E53152
A01.3	*Arabidopsis thaliana*	ArMV	E53152
A02.1	*Arabidopsis thaliana*	ArMV	E53152
A02.2	*Arabidopsis thaliana*	ArMV	E53152
A02.3	*Arabidopsis thaliana*	ArMV	E53152
A03.1	*Arabidopsis thaliana*	ArMV	E53152
A03.2	*Arabidopsis thaliana*	ArMV	E53152
A03.3	*Arabidopsis thaliana*	ArMV	E53152
T01.1	*Arabidopsis thaliana*	TSWV	E53460
T01.2	*Arabidopsis thaliana*	TSWV	E53460
T01.3	*Arabidopsis thaliana*	TSWV	E53460
T02.1	*Arabidopsis thaliana*	TSWV	E53460
T02.2	*Arabidopsis thaliana*	TSWV	E53460
T02.3	*Arabidopsis thaliana*	TSWV	E53460
T03.1	*Arabidopsis thaliana*	TSWV	E53460
T03.2	*Arabidopsis thaliana*	TSWV	E53460
T03.3	*Arabidopsis thaliana*	TSWV	E53460
C01.1	*Arabidopsis thaliana*	CLRV	E395
C01.2	*Arabidopsis thaliana*	CLRV	E395
C01.3	*Arabidopsis thaliana*	CLRV	E395
C02.1	*Arabidopsis thaliana*	CLRV	E395
C02.2	*Arabidopsis thaliana*	CLRV	E395
C02.3	*Arabidopsis thaliana*	CLRV	E395
C03.1	*Arabidopsis thaliana*	CLRV	E395
C03.2	*Arabidopsis thaliana*	CLRV	E395
C03.3	*Arabidopsis thaliana*	CLRV	E395

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

2.4 ELISA

Proposed in 1971 (Engvall *et al.*, 1971; Weemen *et al.*, 1971), the ELISA utilizes the highly specific antigen-antibody bond. A surface is covered with an antibody, which is specific to the antigen in question, here a part of the virus. Upon this layer the antigen solution, which in case of this work was a leave mechanically brought into solution, is given. Any parts of the antigen-solution that did not bind to the antibody covered surface is washed off. A second antibody, specific to either the same or another part of the antigen, is added. The antibody is marked with an enzyme. The antibodies that did not bind to an antigen are washed off. Finally the substrate for the marker-enzyme is introduced. This procedure results in a measurable reaction in those areas where the antigen bond to the surface antibody, creating an antibody-antigen-antibody-enzyme complex. Due to the utilization of an antibody at the surface, and the resulting antibody-antigen-antibody complex, this procedure is called a double antibody sandwich (DAS) ELISA (Voller *et al.*, 1976; Clark *et al.*, 1977). A different approach relies on introducing another antibody into the reaction. The marker-enzyme is linked to this third antibody which is specific to the second antibody, which in turn is specific to the antigen. This procedure, resulting in an antibody-antigen-antibody-antibody-enzyme complex, is known as a triple antibody sandwich (TAS) ELISA (Wilson *et al.*, 2010).

As a quality control and as a method to determine viral infection even without clear symptoms, ELISA was used to detect the viral particles within the infected material produced in section 2.1. For ArMV a DAS ELISA was chosen using provided antibodies (ArMV-IgG 12/09 1.6 mg/ml and ArMV-IgG-AP 12/09 0.8 mg/ml). A TAS ELISA was used to analyze the TSWV samples using AS-0105-0106/0116 from DSMZ. After introducing the substrate for the marker enzymes, the complex was left to be incubated overnight, upon which the extinction at 405 nm was measured.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

In order to produce a baseline and as validation all samples were analyzed with ELISA tests after being stored at -80°C. The workflow for the DAS ELISAs to test for ArMV and CLRV as well as the workflow for the TAS ELISA is shown in figure 2. The TAS ELISA was performed using the AS0205 IgG (BN: 5089) from DSMZ and the corresponding mouse and rabbit-anti-mouse antibodies. The measurement at 405 nm was performed after an overnight incubation at 10°C. The DAS ELISA for the analysis of CLRV relied on the use of the in house antibodies CLRV36 and CLRV38. The ArMV detection with DAS ELISA used the AS-0008 IgG (BN: 5676) from DSMZ. As with the TAS ELISA the detection at 405 nm of both DAS ELISAs was performed after an over night incubation at 10°C.

In order to differentiate between negative and positive reactions a cutoff co was calculated (BIOREBA AG, 2014) using

$$co = (\mu + (3 \cdot \sigma)) \cdot 1.1\,. \tag{2.1}$$

The mean μ and the standard deviation σ in this formula are calculated by sorting the measured extinction values and using only those values following a linear regression.

The material showing the most significant distance from the cut off was used to inoculate the model plants.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

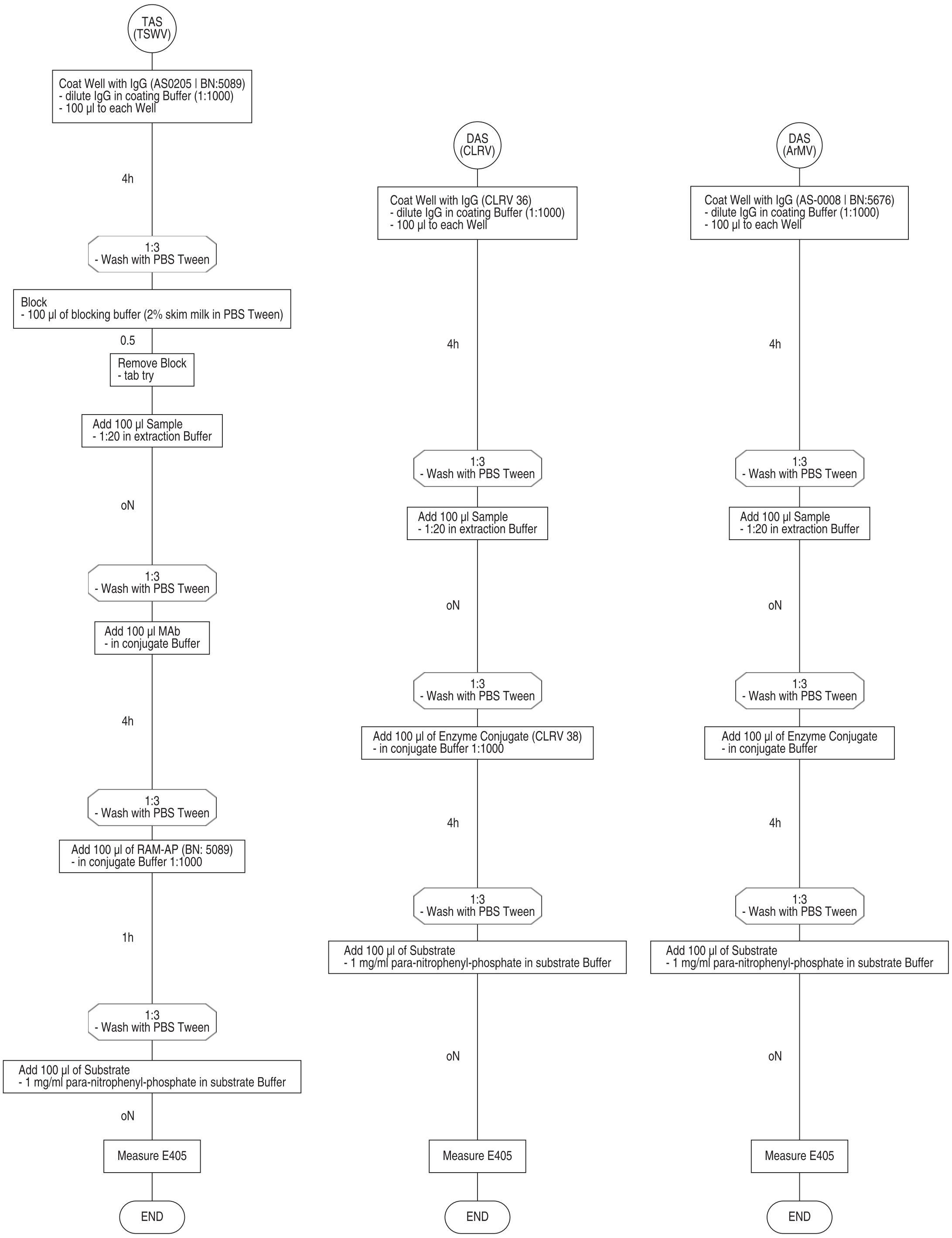

TAS
(TSWV)
Coat Well with IgG (AS0205 | BN:5089)
- dilute IgG in coating Buffer (1:1000)
- 100 µl to each Well
4h
1:3
- Wash with PBS Tween
Block
- 100 µl of blocking buffer (2% skim milk in PBS Tween)
0.5
Remove Block
- tab try
Add 100 µl Sample
- 1:20 in extraction Buffer
oN
1:3
- Wash with PBS Tween
Add 100 µl MAb
- in conjugate Buffer
4h
1:3
- Wash with PBS Tween
Add 100 µl of RAM-AP (BN: 5089)
- in conjugate Buffer 1:1000
1h
1:3
- Wash with PBS Tween
Add 100 µl of Substrate
- 1 mg/ml para-nitrophenyl-phosphate in substrate Buffer
oN
Measure E405
END
DAS
(CLRV)
Coat Well with IgG (CLRV 36)
- dilute IgG in coating Buffer (1:1000)
- 100 µl to each Well
4h
1:3
- Wash with PBS Tween
Add 100 µl Sample
- 1:20 in extraction Buffer
oN
1:3
- Wash with PBS Tween
Add 100 µl of Enzyme Conjugate (CLRV 38)
- in conjugate Buffer 1:1000
4h
1:3
- Wash with PBS Tween
Add 100 µl of Substrate
- 1 mg/ml para-nitrophenyl-phosphate in substrate Buffer
oN
Measure E405
END
DAS
(ArMV)
Coat Well with IgG (AS-0008 | BN:5676)
- dilute IgG in coating Buffer (1:1000)
- 100 µl to each Well
4h
1:3
- Wash with PBS Tween
Add 100 µl Sample
- 1:20 in extraction Buffer
oN
1:3
- Wash with PBS Tween
Add 100 µl of Enzyme Conjugate
- in conjugate Buffer
4h
1:3
- Wash with PBS Tween
Add 100 µl of Substrate
- 1 mg/ml para-nitrophenyl-phosphate in substrate Buffer
oN
Measure E405
END

Figure 2: The ELISA workflows for the TAS ELISA testing for TSWV (left), the DAS ELISA testing for CLRV (middle) and the DAS ELISA testing for ArMV (right) is shown.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

2.5 Sequencing

Ever since Sanger *et al.* (1977) published the first feasible method of sequencing DNA molecules using chain-terminating inhibitors many new discoveries resulted in an ever faster and affordable technology to uncover a DNA sequence. One class of these was named Next-Generation-Sequencing (NGS) technologies, as those are considered to be the next generation after Sanger sequencing. Nowadays these technologies are also called second generation sequencing, since new technologies are being developed, capable of single molecule sequencing which themselves are called next generation or third generation sequencing. These High-Throughput-Sequencing (HTS) technologies are developed mainly by companies and are incorporated into commercial sequencers.

In this study a HTS-technology developed by illumina Inc. called sequencing by synthesis (SBS) was used. This method is comprised of several steps and requires DNA. In order to arrive at the cDNA the total RNA was extracted from plant tissue (randomly chosen leaves) using innuSPEED Plant RNA Kit from Analytik Jena AG (Analytik Jena AG, 2010). The extracted RNA concentration was measured using Qubit RNA HS assay as well as independently with a NanoDrop RNA measurement. Ribosomal RNA was removed from the total RNA and the remaining RNA was transcribed into cDNA using 'TruSeq Stranded Total RNA with Ribo-Zero Plant kit' from illumina Inc. (illumina Inc., 2015). During this step two adapter sequences are added to the sample DNA, one adapter for the 3' and one for the 5' end. The complementary strands to those adapter sequences are covalently bound to a surface, called flowcell. The concentration of cDNA was measured using Qubit as well as LabChip. Finally the prepared cDNA was sequenced using illumina's NextSeq500 sequencer and 'NextSeq reagent kit V2 (300 cycles)' (illumina Inc., 2016) also from illumina Inc. This process begins by introducing the prepared DNA to the flowcell, where it will bind to the complementary adapter sequences. Polymerases form the com-

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

plementary strand to the sample DNA, using the surface adapter oligo as a starting point. These double stranded sequences are denatured and the original strand is washed off. The strand is amplified using a process called bridge amplification. The DNA sequence, now bound to the surface, bends to hybridize its second adapter sequence to its complementary sequence on the flowcell forming a bridge. Polymerases form the complementary strand, using the second adapter sequence as a starting point. The hybridized strands are denatured, leaving the original strand and its complementary counterpart. Both strands can now be used for amplification. These steps are repeated several times forming clusters on the surface of the flowcell. The 3' ends of these sequences are protected to suppress unwanted hybridizations. During the next step all reverse sequences are removed from the flowcell. The sequencing itself is performed by adding primers which bind to parts of the adapter sequence. This step is followed by injecting fluorescently tagged nucleotides onto the flowcell, one of which hybridizes to the strand. The fluorescent tag prohibits the hybridization of any further nucleotides to the sequence. It is excited and the emitted wavelength is recorded. To allow for the next nucleotide to hybridize the fluorescent tag is cleaved off. The process of hybridizing one nucleotide, exciting it and measuring its emitted wavelength is repeated until a predefined readlength is reached. The newly formed strand is washed off. After deprotection of the 3' end of the sequence, it folds over forming once again a bridge. Polymerases form the complementary strand. After denaturing, the original strand is washed off leaving only the newly formed strand behind. The sequencing steps performed on the original strand are now repeated for its reverse complement (illumina Inc., 2014; 2010). This procedure results in two sequences per original DNA molecule, one forward and one reverse. Depending on the length of the original DNA molecules and the readlength, these paired-end reads may either overlap (in part or completely) resulting a double determination of either part or all of the original DNA, or be a number of bases apart resulting in the start- and end-sequences of larger regions.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

The intensity of the emitted wavelength is recorded in multiple `bcl` basecall files and converted into a `fastQ` file for later analysis. The `fastQ` format is a widely used American Standard Code for Information Interchange (ASCII) file, holding both the sequence and per base quality. One sequence, also known as a read, is represented by four lines, a header line, starting with '@', the read itself, a '+' and a quality per base string (Cock *et al.*, 2010), see example listing 1 for clarification.

Listing 1: The listing shows an example excerpt of a `fastQ` file.

```
1 @example Read
2 ATCGGAAGAGCACACGTCTGAACTCCAGTCACTTCGGAGAATCTCGTATGCCGTCTTCTG
3 +
4 CCCCCGEEGFAFCFGGCFGGGGGGGGCFEFFF@,C6,@@:CCE9FFGGGGG<@8@FF5@C
```

The per base quality is the ASCII encoded PHRED score and is defined as

$$Q_{PHRED} = -10 \cdot log_{10}(P_e)\,, \tag{2.2}$$

where P_e is the probability of the base being called erroneously (Ewing *et al.*, 1998). This score is added to 33 (Illumina 1.8+ Phred+33) and the resulting number is represented by its ASCII value (Cock *et al.*, 2010). Thus a base being called with a probability of being wrong of $P_e = 0.01$, so being 99% accurate, has a PHRED score of 20, adding this to 33 results in 53 which encodes the '5' in ASCII. If a base is called with an error probability of 0.0004, its PHRED score would be $-10 \cdot log_{10}(0.0004) = 33.9$ which is added to 33 and transformed into a character using ASCII encoding $67 \xrightarrow{ASCII}$ 'C'. This is done for all bases and the resulting characters are the quality string as shown in line four of listing 1.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

2.6 Transcriptome Analysis

The transcriptome of a given individual is a collection of all RNA molecules transcribed from those genes being active at a particular point in time either in a single cell or a conglomerate of cells (Brown, 2006). Unlike the exome, which is a static representation of all genes on a DNA, the transcriptome can be used to analyse expression changes over time (Wang *et al.*, 2009; Chen *et al.*, 2016), i.e. the up-/down-regulation of specific genes.

In RNA-Sequencing, there are multiple ways to clean up the RNA in order to only retain those kinds of RNAs which are of particular interest. Capturing only RNA molecules which have a polyadenylated (poly(A)) tail ensures that only RNA transcribed from protein coding genes is later sequenced. This method, however, is not ideal when working with degraded RNA or when the molecules of interest do not show the characteristic of a poly(A) tail (sRNA, microRNA, non-eukaryotic RNA, lincRNA, etc). Another method is based on removing only the highly prevalent ribosomal RNA (80% of a cell's RNA (Lodish *et al.*, 2000)), leaving all other RNA molecules to be sequenced. This procedure retains a lot of information crucial for further analysis and was thus chosen.

After the total RNA has been depleted of either ribosomal RNA or all non-poly(A)-tail-RNA, it is reverse transcribed into cDNA and further handled like normal DNA during sequencing (see section 2.5) (Wilhelm *et al.*, 2008; Nagalakshmi *et al.*, 2008).

2.7 Data Analysis

With the sequencing finalized and the `fastQ` files generated using `bcl2fastq` v2.15, the analysis of the large dataset commences. While there are numerous approaches of how to deal with the data, a common workflow is to first filter the reads by quality

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

using the PHRED scores. Following this filtering the reads are mapped to a reference genome (Gogol-Döring *et al.*, 2012). This alignment can be done using different programs (i.e. bowtie (Langmead *et al.*, 2009; 2012), eland (Cox, 2007), bwa (Li *et al.*, 2009)). By mapping a read onto a reference genome, the alignment further classifies said read into one of three categories: a read that could be mapped to the reference in exactly one position (uniquely matching read (UMR)), multiple positions (multiply matching read (MMR)) or no position at all (nonmatching read (NMR)). Utilizing the information of where a given read stems from, its position on the reference, missmatches and alignment quality, many different analysis steps can be performed. The UMRs and MMRs can be used to discover differences between the reads and the reference, pointing to single nucleotide polymorphisms (SNPs) or other variations within the sample DNA. Using an annotated reference, which holds the information about start and end positions of all known genes, exons and transcripts, the UMRs and MMRs can be utilized to uncover copy number variations or, in case of transcriptome sequencing (section 2.6), the amount of RNA molecules belonging to a specific transcript, gene or exon can be counted, effectively measuring the expression of said transcript, gene or exon. Doing this measurement over all known transcripts, exons and genes can thus be used to create a transcription profile of an individual at a given point in time. The NMRs can be matched to other references, pointing for example to infections (section 2.8).

Without the alignment, the reads can be used to assemble a new genome or use the n-gram based sequence profiler (`srProfiler`) approach to gain further insights into the data.

All reads were aligned using `bowtie2` v2.2.6 (Langmead *et al.*, 2009; 2012). The alignment was first performed against the reference of *Arabidopsis thaliana* tair v10.29 (Berardini *et al.*, 2015), followed by the alignment against the references of the three viruses (section 2.1). UMRs were used for further analysis of viral load as well as transcriptome expressions in the samples.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

Moreover, the reads were transformed according to section 2.9 and the resulting profiles were used to run multiple statistical comparisons. Figure 3 shows the major steps being performed starting with the separation of the sample pool into a test and a training set. The sets were constructed using only the odd reads of each sample for the training set and only the even reads for the test set. The training set was used to create classifiers for different questions. The newly formed classifiers are run against the independent test set. In addition one sample was chosen at random and deliberately left out of any pathogen related calculations in order to simulate an entirely independent sample. A3.2, an *Arabidopsis thaliana* individual successfully infected with ArMV, so unanimously diagnosed by ELISA and pathogen alignment, was chosen to be said sample. The left side shows the steps required to gather the necessary information using the reference alignment, whereas the right side depicts the steps performed using the `srProfiler` approach. The feature vectors both approaches result in are the basis for independent classifiers, which are used to classify the test set.

The classification was performed using the groupings shown in table 6. The grouping for the single classifiers was determined by the experimental design, however, based on the results of the validating steps (ELISA and pathogen alignment) the groupings for the infectant classifiers had to be adapted. Since a definitive infection state (infected, not infected) could not be declared for all samples, multiple plants were removed from the groupings to ensure a bias free classifier.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

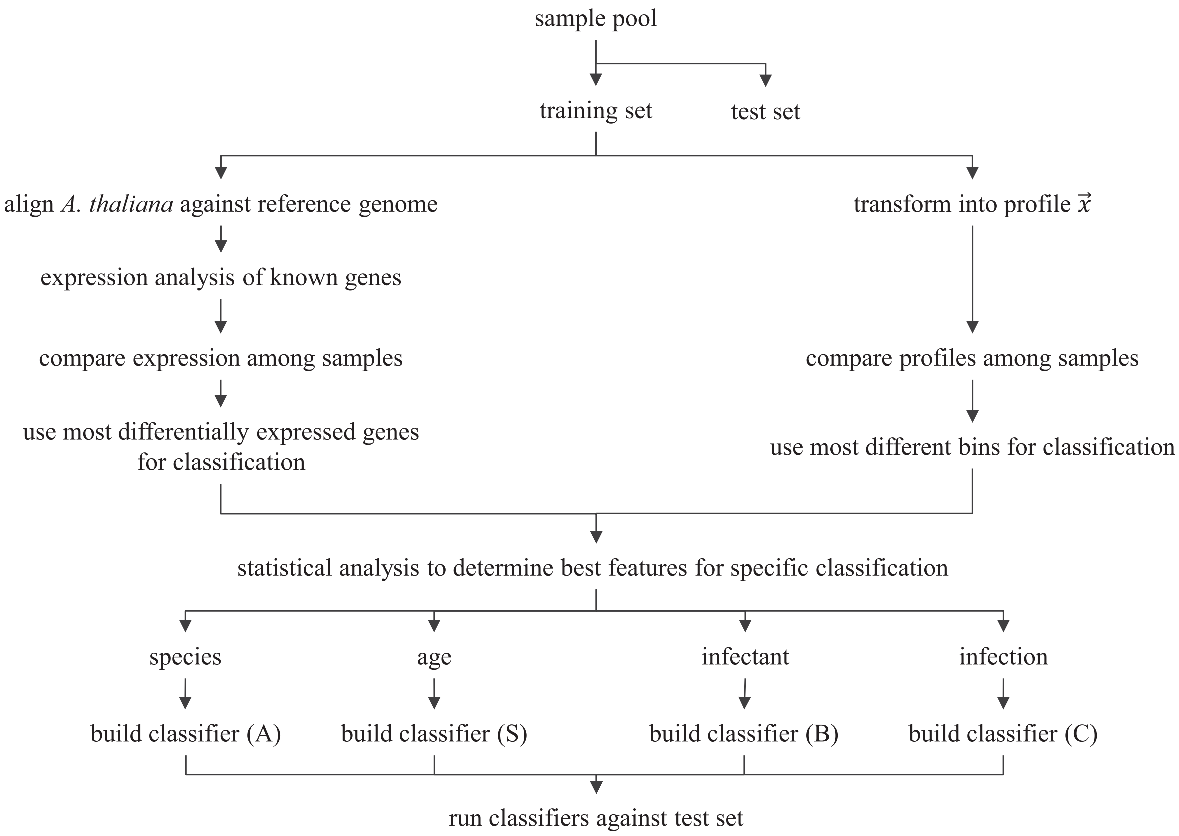

sample pool
training set
test set
align A. thaliana against reference genome
transform into profile $\vec{x}$
expression analysis of known genes
compare expression among samples
compare profiles among samples
use most differentially expressed genes for classification
use most different bins for classification
statistical analysis to determine best features for specific classification
species
age
infectant
infection
build classifier (A)
build classifier (S)
build classifier (B)
build classifier (C)
run classifiers against test set

Figure 3: The major data analysis steps in *Arabidopsis thaliana* are shown. Before the analysis starts the data set is devided into a test and a training set. The training set is analysed using a reference genome (left side) and using `srProfiler` (right side). Both approaches arrive at specific feature vectors upon which designated classifiers are build. Those classifiers are used to classify the test set.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

Table 6: The table shows the groupings according to the chosen classifiers. The symbol 'NA' represents samples that were not used during the respective classification. The symbol '!' means 'NOT'. The abbreviation TP stands for time point and represents the respective harvest dpi. The infectant classifiers are shown in two versions ('in Theory' and 'in Praxis'). The first version, 'in Theory', represents the grouping as it should have been based on experimental design. The second version, 'in Praxis', lists the groupings as they were used based on the results of ELISA and pathogen alignment.

Sample	**Species**	**Age**	**Infectant (in Theory)**					**Infectant (in Praxis)**				
	A	**S**	**B**	**B1**	**B2**	**B3**	**B4**	**B**	**B1**	**B2**	**B3**	**B4**
A01	*C. quinoa*	TP1	ArMV	ArMV	!CLRV	!TSWV	!Control	ArMV	ArMV	!CLRV	NA	!Control
A02	*C. quinoa*	TP2	ArMV	ArMV	!CLRV	!TSWV	!Control	ArMV	ArMV	!CLRV	NA	!Control
A03	*C. quinoa*	TP3	ArMV	ArMV	!CLRV	!TSWV	!Control	ArMV	ArMV	!CLRV	NA	!Control
A1.1	*A. thaliana*	TP1	ArMV	ArMV	!CLRV	!TSWV	!Control	NA	NA	!CLRV	NA	NA
A1.2	*A. thaliana*	TP1	ArMV	ArMV	!CLRV	!TSWV	!Control	NA	NA	!CLRV	NA	NA
A1.3	*A. thaliana*	TP1	ArMV	ArMV	!CLRV	!TSWV	!Control	NA	NA	!CLRV	NA	NA
A2.1	*A. thaliana*	TP2	ArMV	ArMV	!CLRV	!TSWV	!Control	ArMV	ArMV	!CLRV	NA	!Control
A2.2	*A. thaliana*	TP2	ArMV	ArMV	!CLRV	!TSWV	!Control	ArMV	ArMV	!CLRV	NA	!Control
A2.3	*A. thaliana*	TP2	ArMV	ArMV	!CLRV	!TSWV	!Control	NA	NA	!CLRV	NA	NA
A3.1	*A. thaliana*	TP3	ArMV	ArMV	!CLRV	!TSWV	!Control	ArMV	ArMV	!CLRV	NA	!Control
A3.2	*A. thaliana*	TP3	ArMV	ArMV	!CLRV	!TSWV	!Control	NA	NA	!CLRV	NA	NA
A3.3	*A. thaliana*	TP3	ArMV	ArMV	!CLRV	!TSWV	!Control	NA	NA	!CLRV	NA	NA
C01	*C. quinoa*	TP1	CLRV	!ArMV	CLRV	!TSWV	!Control	NA	!ArMV	NA	NA	NA
C02	*C. quinoa*	TP2	CLRV	!ArMV	CLRV	!TSWV	!Control	CLRV	!ArMV	CLRV	NA	!Control
C03	*C. quinoa*	TP3	CLRV	!ArMV	CLRV	!TSWV	!Control	CLRV	!ArMV	CLRV	NA	!Control
C1.1	*A. thaliana*	TP1	CLRV	!ArMV	CLRV	!TSWV	!Control	NA	!ArMV	NA	NA	NA
C1.2	*A. thaliana*	TP1	CLRV	!ArMV	CLRV	!TSWV	!Control	NA	!ArMV	NA	NA	NA
C1.3	*A. thaliana*	TP1	CLRV	!ArMV	CLRV	!TSWV	!Control	NA	!ArMV	NA	NA	NA
C2.1	*A. thaliana*	TP2	CLRV	!ArMV	CLRV	!TSWV	!Control	NA	!ArMV	NA	NA	NA
C2.2	*A. thaliana*	TP2	CLRV	!ArMV	CLRV	!TSWV	!Control	NA	!ArMV	NA	NA	NA
C2.3	*A. thaliana*	TP2	CLRV	!ArMV	CLRV	!TSWV	!Control	NA	!ArMV	NA	NA	NA
C3.1	*A. thaliana*	TP3	CLRV	!ArMV	CLRV	!TSWV	!Control	NA	!ArMV	NA	NA	NA
C3.2	*A. thaliana*	TP3	CLRV	!ArMV	CLRV	!TSWV	!Control	NA	!ArMV	NA	NA	NA
C3.3	*A. thaliana*	TP3	CLRV	!ArMV	CLRV	!TSWV	!Control	NA	!ArMV	NA	NA	NA
T01	*C. quinoa*	TP1	TSWV	!ArMV	!CLRV	TSWV	!Control	NA	!ArMV	!CLRV	NA	NA
T02	*C. quinoa*	TP2	TSWV	!ArMV	!CLRV	TSWV	!Control	NA	!ArMV	!CLRV	NA	NA
T03	*C. quinoa*	TP3	TSWV	!ArMV	!CLRV	TSWV	!Control	NA	!ArMV	!CLRV	NA	NA
T1.1	*A. thaliana*	TP1	TSWV	!ArMV	!CLRV	TSWV	!Control	NA	!ArMV	!CLRV	NA	NA
T1.2	*A. thaliana*	TP1	TSWV	!ArMV	!CLRV	TSWV	!Control	NA	!ArMV	!CLRV	NA	NA
T1.3	*A. thaliana*	TP1	TSWV	!ArMV	!CLRV	TSWV	!Control	NA	!ArMV	!CLRV	NA	NA
T2.1	*A. thaliana*	TP2	TSWV	!ArMV	!CLRV	TSWV	!Control	NA	!ArMV	!CLRV	NA	NA
T2.2	*A. thaliana*	TP2	TSWV	!ArMV	!CLRV	TSWV	!Control	NA	!ArMV	!CLRV	NA	NA
T2.3	*A. thaliana*	TP2	TSWV	!ArMV	!CLRV	TSWV	!Control	NA	!ArMV	!CLRV	NA	NA
T3.1	*A. thaliana*	TP3	TSWV	!ArMV	!CLRV	TSWV	!Control	NA	!ArMV	!CLRV	NA	NA
T3.2	*A. thaliana*	TP3	TSWV	!ArMV	!CLRV	TSWV	!Control	NA	!ArMV	!CLRV	NA	NA
T3.3	*A. thaliana*	TP3	TSWV	!ArMV	!CLRV	TSWV	!Control	NA	!ArMV	!CLRV	NA	NA
X01	*C. quinoa*	TP1	Control	!ArMV	!CLRV	!TSWV	Control	Control	!ArMV	!CLRV	NA	Control
X02	*C. quinoa*	TP2	Control	!ArMV	!CLRV	!TSWV	Control	Control	!ArMV	!CLRV	NA	Control
X03	*C. quinoa*	TP3	Control	!ArMV	!CLRV	!TSWV	Control	Control	!ArMV	!CLRV	NA	Control
X1.1	*A. thaliana*	TP1	Control	!ArMV	!CLRV	!TSWV	Control	Control	!ArMV	!CLRV	NA	Control
X1.2	*A. thaliana*	TP1	Control	!ArMV	!CLRV	!TSWV	Control	NA	!ArMV	NA	NA	NA
X1.3	*A. thaliana*	TP1	Control	!ArMV	!CLRV	!TSWV	Control	Control	!ArMV	!CLRV	NA	Control
X2.1	*A. thaliana*	TP2	Control	!ArMV	!CLRV	!TSWV	Control	Control	!ArMV	!CLRV	NA	Control
X2.2	*A. thaliana*	TP2	Control	!ArMV	!CLRV	!TSWV	Control	Control	!ArMV	!CLRV	NA	Control
X2.3	*A. thaliana*	TP2	Control	!ArMV	!CLRV	!TSWV	Control	NA	!ArMV	NA	NA	NA
X3.1	*A. thaliana*	TP3	Control	!ArMV	!CLRV	!TSWV	Control	Control	!ArMV	!CLRV	NA	Control
X3.2	*A. thaliana*	TP3	Control	!ArMV	!CLRV	!TSWV	Control	NA	!ArMV	NA	NA	NA
X3.3	*A. thaliana*	TP3	Control	!ArMV	!CLRV	!TSWV	Control	Control	!ArMV	!CLRV	NA	Control

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

2.8 Pathogen Alignment

Section 2.6 explained that in order to purify the sample RNA only the rRNA had been removed. Hence, if an infection did in fact occur in a given sample, the viruses are still discoverable after purification, since, as described in section 2.1, all three viruses are RNA-viruses and are not removed with the rRNA. Aligning the sample data from section 2.5 against the reference sequences of the three viruses respectively (section 2.1), is thus a method to describe the presence of the pathogen as well as the amount of viral RNA present within a given sample (Nagano *et al.*, 2015).

2.9 n-Gram Based Sequence Profiling for Alignment Free Transcriptome Analysis

In many cases the transcriptome analysis of organisms for which a reference genome has not been described is a time consuming, multi step approach that includes the alignment to the reference of a relatively similar species and a subsequent den-ovo assembly (Collins *et al.*, 2008; Ward *et al.*, 2012). In order to classify a set of plants based on different criteria, a complete transcriptome analysis is not required if the data can be harmoniously transformed into a unique, representative profile. Thus neither the alignment nor the denovo assembly is needed (Song *et al.*, 2013; Bonham-Carter *et al.*, 2013). A unique and representative profile can be constructed using a window of length n moving over a given sequence. The occurrence of each possible n-sized subsequence (n-gram) within a sequence is counted, resulting in a profile $\vec{x}$ of a^n counts (bins), where a is the size of the alphabet within the underlying sequence. Here a is 4 since the alphabet is $A := \{A, T, G, C\}$ (Sims *et al.*, 2009). The `srProfiler` constructs those n-gram frequency profiles and, additionally, transforms the single n-grams into binary strings, each consisting of $n \cdot 2$ bits (Amberg, 2014) as shown in figure 4. This not only reduces the size of the resulting profile ($2n$ bits

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

rather than $8n$ bits), it also markably increases the speed with which the profiles can be handled.

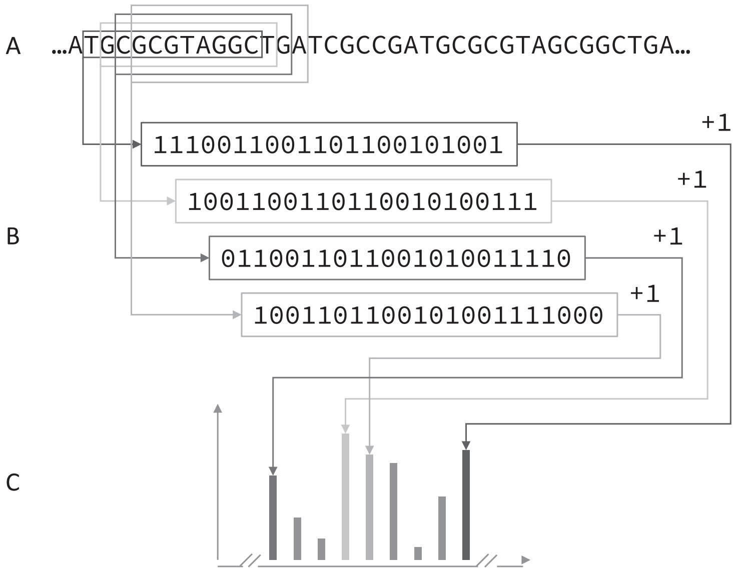


Figure 4: Using a window size n of 11 the sequence (A) is transformed into multiple binary n-grams (B), each consisting of $n \cdot 2$ bits with an offset of 2 bits to the previous n-gram. Each time a specific n-gram is encountered its count is increased by 1 resulting in the profile $\vec{x}$ depicted as a histogram (C).

2.10 Statistical Analysis

The statistical analyses are used to first reduce the dimensionality of the sample vectors (33.605 dimensions using expression vectors and 4.194.304 dimensions using `srProfiler` vectors) and then build classifiers to differentiate between different groupings of sample properties (i.e. species, pathogen, age). The workflow is shown in figure 5. The first step is to find those candidates amongst all dimensions which can best distinguish among the given groups using ANOVA (section 2.10.2) and Student's t test (section 2.10.1) and thus building a candidate vector. A classification is possible using those candidate vectors, however, due to the complex nature of the underlying processes, a combination of multiple dimensions produces clearer results. Therefore the best 25 candidates are used to calculate the principal

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

components (section 2.10.3). These components can be utilized to project the high dimensional candidate vector (25 dimensions) into an x-dimensional space, where $x << 25$. Reducing the dimensionality of the original sample vectors removes information that has no relevance for the particular classification. It is not a useful information that a sample is infected with a pathogen if the classification is supposed to differentiate the species of the samples. Reducing this background noise from the sample vectors allows for a better classification. Moreover, using only a few dimensions for the classification results in a faster response and less computational requirements (i.e. storage or CPU). Once a classifier is constructed the only input it requires for differentiation are the lower dimensional candidate vectors of the samples in question and not the high dimensional sample vectors.

For each classifier (table 6) the procedure of reducing dimensionality was repeated until each dimension was weighted with a pValue unique to the classifier being used. The calculations were performed on dimensions that were greater than zero in all training samples involved in the construction of a given classifier. Therefore only genes with an expression value greater than zero and bins which were counted at least once in all training samples making up a classifier group were used. On the dimensions weighted with the lowest pValue, thus having a high probability of differentiating between the classifier's groups, Principal Component Analysis (PCA) was performed.

2.10.1 Welch's unequal variances t-test

Student's t-test was first published in 1908 (Student, 1908) and can be used to determine whether the sample means $\bar{x}_g$ of two samples $g \varepsilon \{1,2\}$ with sample sizes ranging from $i = 1...N_g$ and elements x_{ig} are significantly equal to one another, which is defined as the null hypothesis $H_0 :: \bar{x}_1 = \bar{x}_2$. This is done by calculating the t-value by using the mean of each sample $\bar{x}_g$

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

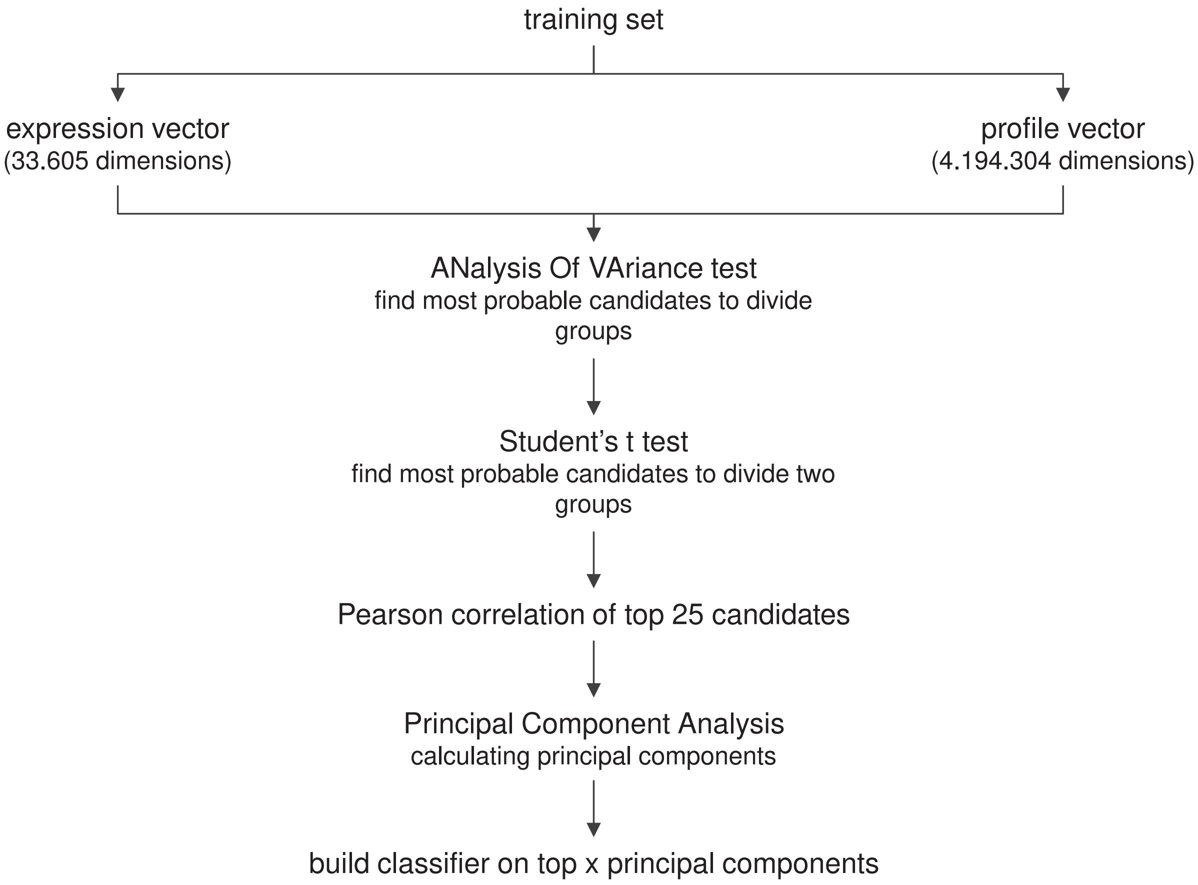

training set
expression vector
(33.605 dimensions)
profile vector
(4.194.304 dimensions)
ANalysis Of VAriance test
find most probable candidates to divide groups
Student's t test
find most probable candidates to divide two groups
Pearson correlation of top 25 candidates
Principal Component Analysis
calculating principal components
build classifier on top x principal components

Figure 5: The statistical workflow is shown. The goal of these analyses is to reduce the high dimensionality inherent to each sample vector, independent of the method this vector was produced by. The sample vectors derived from either transcriptome analysis or `srProfiler` approach are grouped corresponding to their properties (i.e. species). ANOVA testing is used to find probable candidates among all dimensions which differentiate the groups. The probable candidates are run through Student's t test to discover which two groups those candidates effectively differentiate. The best 25 dimensions differentiating the groups in question are Pearson correlated and a Principal Component Analysis is run upon the resulting matrix. Using the first x principal components the classifiers are designed.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

$$\bar{x}_g = \frac{1}{N_g} \sum_{i=1}^{N_g} x_{ig} \tag{2.3}$$

and the sample standard deviation s_g of each sample g using Bessels's correction

$$s_g = \sqrt{\frac{1}{N_g - 1} \sum_{i=1}^{N_g} (x_{ig} - \bar{x}_g)} \,. \tag{2.4}$$

In 1947 Bernard L. Welch proposed an adaptation for a more robust t-Test with samples of unequal variances and sample sizes (Welch, 1947). The value t according to Welch is calculated using

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}} \,, \tag{2.5}$$

with the corresponding degrees of freedom d being

$$d = \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)^2}{\frac{\left(\frac{s_1^2}{N_1}\right)^2}{N_1 - 1} + \frac{\left(\frac{s_2^2}{N_2}\right)^2}{N_2 - 1}} \,. \tag{2.6}$$

2.10.2 ANOVA-Test

Introduced in 1925 by Ronald Fisher, the analysis of variance (ANOVA) test is used to determine whether the means $\bar{x}$ of multiple samples $g \; \varepsilon \; 1...G$ comprised of the elements $x \; \varepsilon \; 1...X$ are significantly equal (Fisher, 1925). It does so by comparing the sample variation, or sum of squares, ss_g within sample g of sample size N_g with the variation between samples ss_b. Calculating the mean $\bar{x}$ of a given sample g as in equation 2.3 and the corresponding variance ss_g with

$$ss_g = \sum_{i=1}^{N_g} (x_i - \bar{x}_g)^2 \,, \tag{2.7}$$

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

the variation within all samples ss can be calculated as

$$ss_w = \sum_{g=1}^{G} ss_g \,. \tag{2.8}$$

Further, the variation between samples ss_b is calculated by first joining all samples $g_1..g_G$ into one sample g_{total} with group size

$$N_{total} = \sum_{g=1}^{G} N_g \tag{2.9}$$

and calculating the mean for said sample $\bar{x}_{total}$ analogous to equation 2.3 and its variance ss_{total} as in equation 2.7 using N_{total} and all elements x from all samples with

$$\bar{x}_{total} = \frac{1}{N_{total}} \sum_{i=1}^{N_{total}} x_i \tag{2.10}$$

and

$$ss_{total} = \sum_{i=1}^{N_{total}} (x_i - \bar{x}_{total})^2 \,. \tag{2.11}$$

Now ss_b can be calculated by subtracting the total variation ss_{total} from the variation within all samples ss_w arriving at

$$ss_b = ss_{total} - ss_w \,. \tag{2.12}$$

With these values an f-test can be performed to evaluate whether the null-hypothesis $H_0 :: \bar{x}_i = \bar{x}_j$ (the means of all samples are significantly equal) can be rejected. The degrees of freedom for ss_b are defined as the number of samples G subtracted by one and the degrees of freedom for ss_w are given by the amount of elements X subtracted by the amount of samples G, thus

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

$$F\left((G-1),(X-G)\right)=\frac{\frac{ss_b}{G-1}}{\frac{ss_w}{X-G}}\,. \tag{2.13}$$

2.10.3 Principal Component Analysis

PCA, proposed by Pearson (1901), is a method that converts a vector of observations into a vector of linear combinations, so called principal components. The number of principal components is smaller or equal to the number of original observations. Amongst all principal components the first one describes the largest possible variability of the given observations. The second principal component describes the second largest variability and so forth. It is thus possible to describe a given data set using principal components instead of all dimensions. Depending on the required correlation between original data set and resulting principal component set fewer components can be used allowing for a significant reduction of dimensionality.

The observation vectors are joined into a matrix $M(n,d)$ with n rows (samples) and d columns (dimensions). By means of double centering M is transformed into the Z matrix. Each element $Z_{i,j}$ is calculated using $M_{i,j}$, being part of dimension $i\ \varepsilon\ 1...d$ and sample $j\ \varepsilon\ 1...n$. $M_{i,j}$ is subtracted from the mean of its dimension $\bar{x_i}$ and then divided by its dimensions' standard deviation s_i

$$Z_{i,j}=\frac{M_{i,j}-\bar{x_i}}{s_i}\,. \tag{2.14}$$

The pearson correlation matrix of Z is calculated followed by the calculation of the eigenvectors of said correlation matrix. The double centered sample vector $Z_{,j}$ is multiplied with the first eigenvector to arrive at the first dimension of the scaled sample vector. Multiplying $Z_{,j}$ with the second eigenvector yields the second dimension of the scaled sample vector. This step is repeated depending on the intended di-

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

mensionality of the final scaled sample vector. The matrix of scaled sample vectors $\hat{M}$ can be regarded as a map within which the scale and rotation can be ignored, solely the distance of the points relative to one another matters and represents the same distance relations as the original higher dimensional space D.

As a measure of the goodness of fit of the scaled space, Kruskal (1964) proposed the STRESS. This value uses the distance of the original points and the scaled points.

$$STRESS = \sqrt{\frac{\sum_{i<j} \left(d_{i,j} - \hat{d}_{i,j}\right)^2}{\sum_{i<j} d_{i,j}^2}} \tag{2.15}$$

Where d is the distance matrix of the double centered matrix Z and $\hat{d}$ is the distance matrix of the final scaled matrix $\hat{M}$. In order to interpret the STRESS value table 7 shows a reference as proposed by Kruskal (1964).

Table 7: The table shows the reference values to interpret the STRESS value as proposed by Kruskal (1964).

STRESS	**Fit**
0.20	Poor
0.10	Fair
0.05	Good
0.02	Excellent

As a further measure the squared Pearson Correlation r^2 between d and $\hat{d}$ can be used.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

3 Results

After inoculation the sample plants where divided into three groups. The *Arabidopsis thaliana* groups grew for 0, 24 and 46 days respectively after inoculation before being harvested. The *Chenopodium quinoa* groups were harvested after 0, 9 and 14 days respectively after inoculation. Figures 6 through 13 show specimen of each group at the day of harvesting.

Arabidopsis thaliana control samples, which were mock inoculated, are shown before or directly after harvesting in figure 6. Figure 7 shows the *Chenopodium quinoa* control samples directly before harvesting. The samples depicted show no obvious signs of viral infection.

The CLRV infected *Arabidopsis thaliana* samples can be seen in figure 8. The sample harvested 24 dpi shows necroses on some leaves. The sample harvested 46 dpi shows no obvious symptoms. The *Chenopodium quinoa* samples infected with CLRV are shown in figure 9. The samples harvested 9 and 14 dpi show strong signs of infection. The sample harvested 9 dpi is wilting and shows leaf rolling as well as chloroses. The chloroses in the sample harvested 14 dpi are less obvious, however the wilting and leaf rolling are more prominent, all but two leaves are completely withered.

The ArMV infected samples are shown in figures 10 and 11. The *Arabidopsis thaliana* samples harvested 24 and 46 dpi show necroses on several leaves. The *Chenopodium quinoa* samples harvested 9 and 14 dpi show strong chloroses on all leaves as well as strong signs of wilting.

The samples depicted in figures 12 and 13 were inoculated with TSWV. Obvious symptoms of viral infection cannot be observed in any of the samples.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

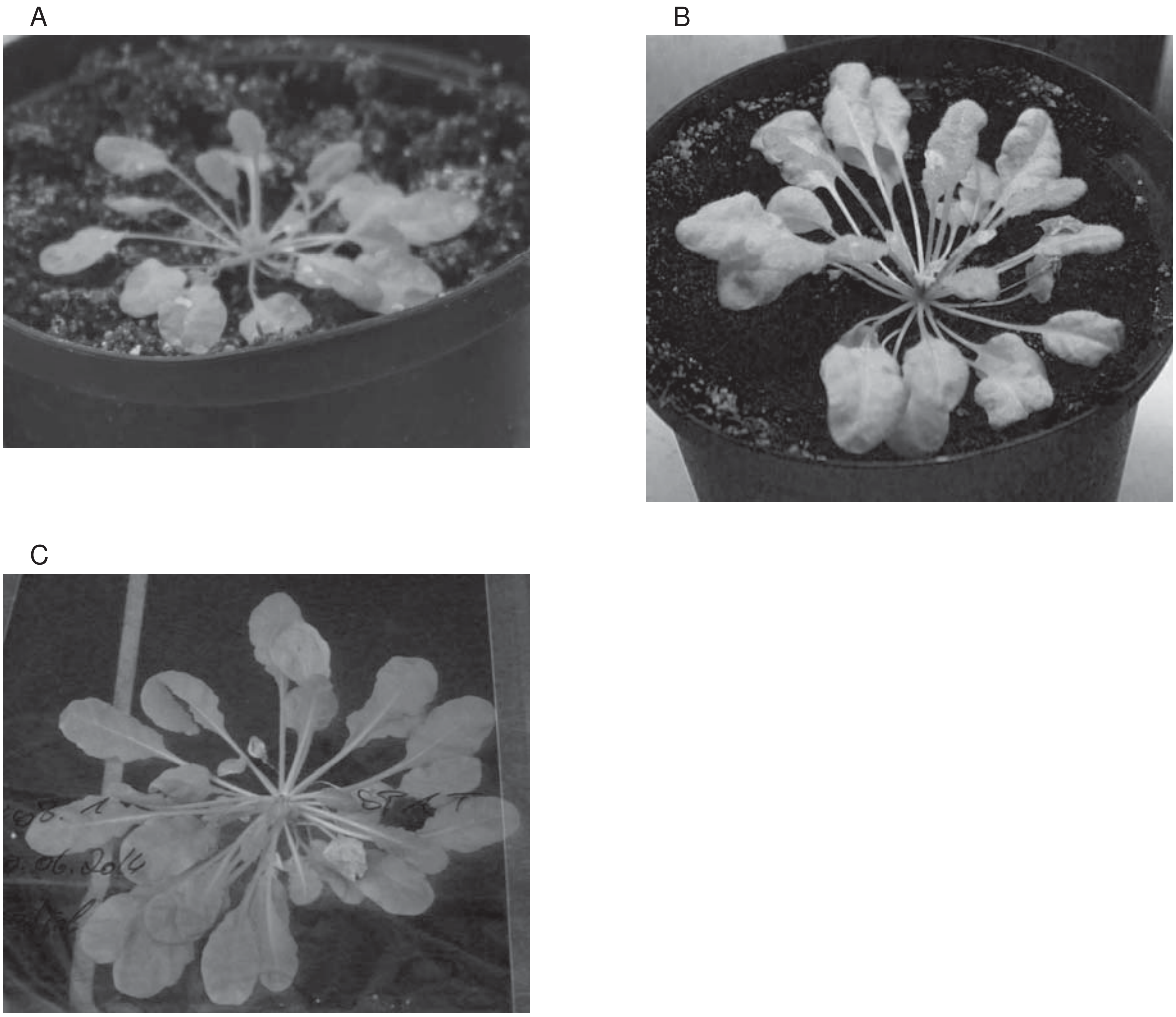


Figure 6: *Arabidopsis thaliana* control samples harvested 0 dpi (A), 24 dpi (B) and 46 dpi (C). The samples harvested 0 and 9 dpi are shown directly before, while the sample harvested 46 dpi is shown post harvesting.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

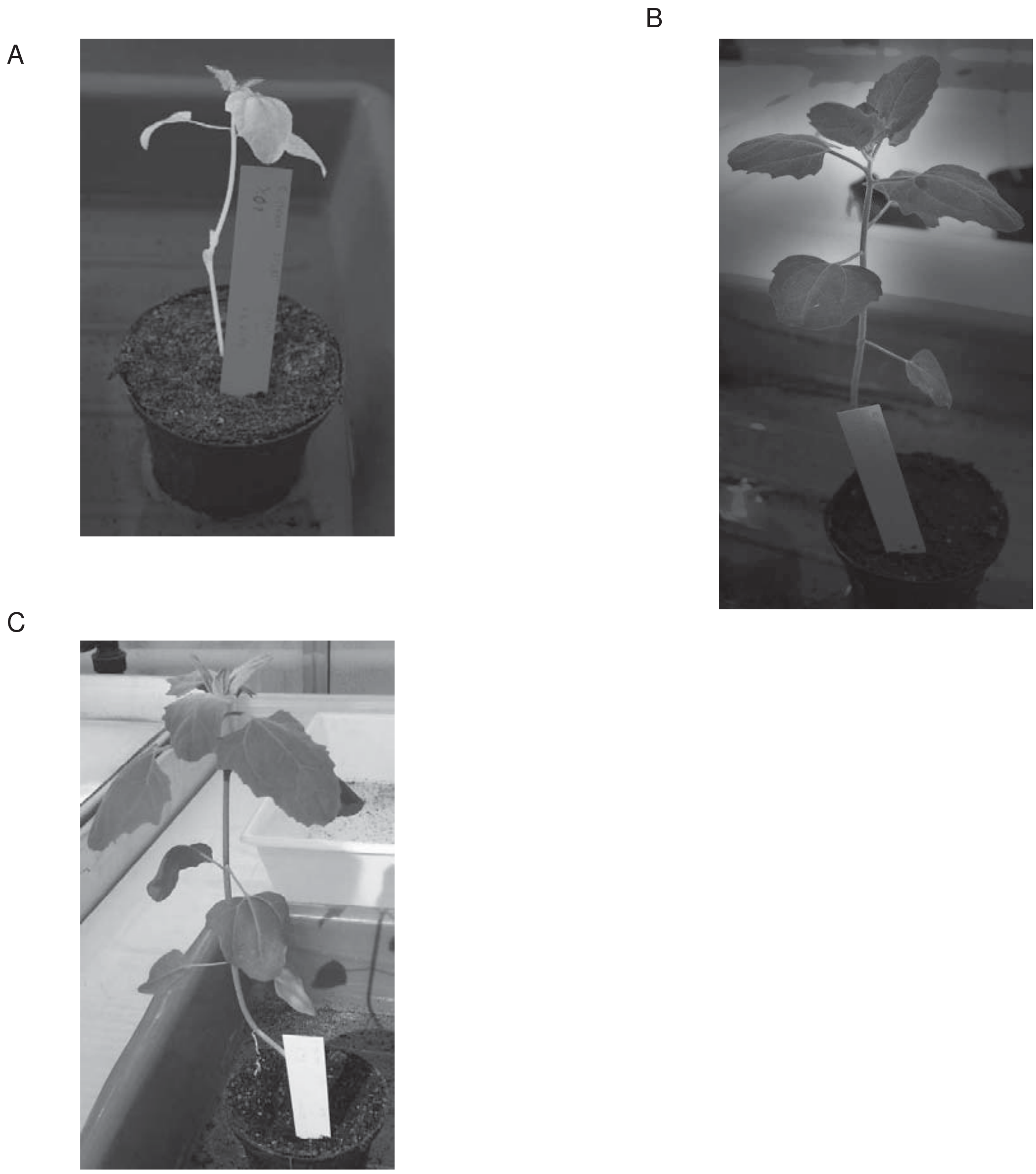


Figure 7: *Chenopodium quinoa* control samples harvested 0 dpi (A), 9 dpi (B) and 14 dpi (C) are shown.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

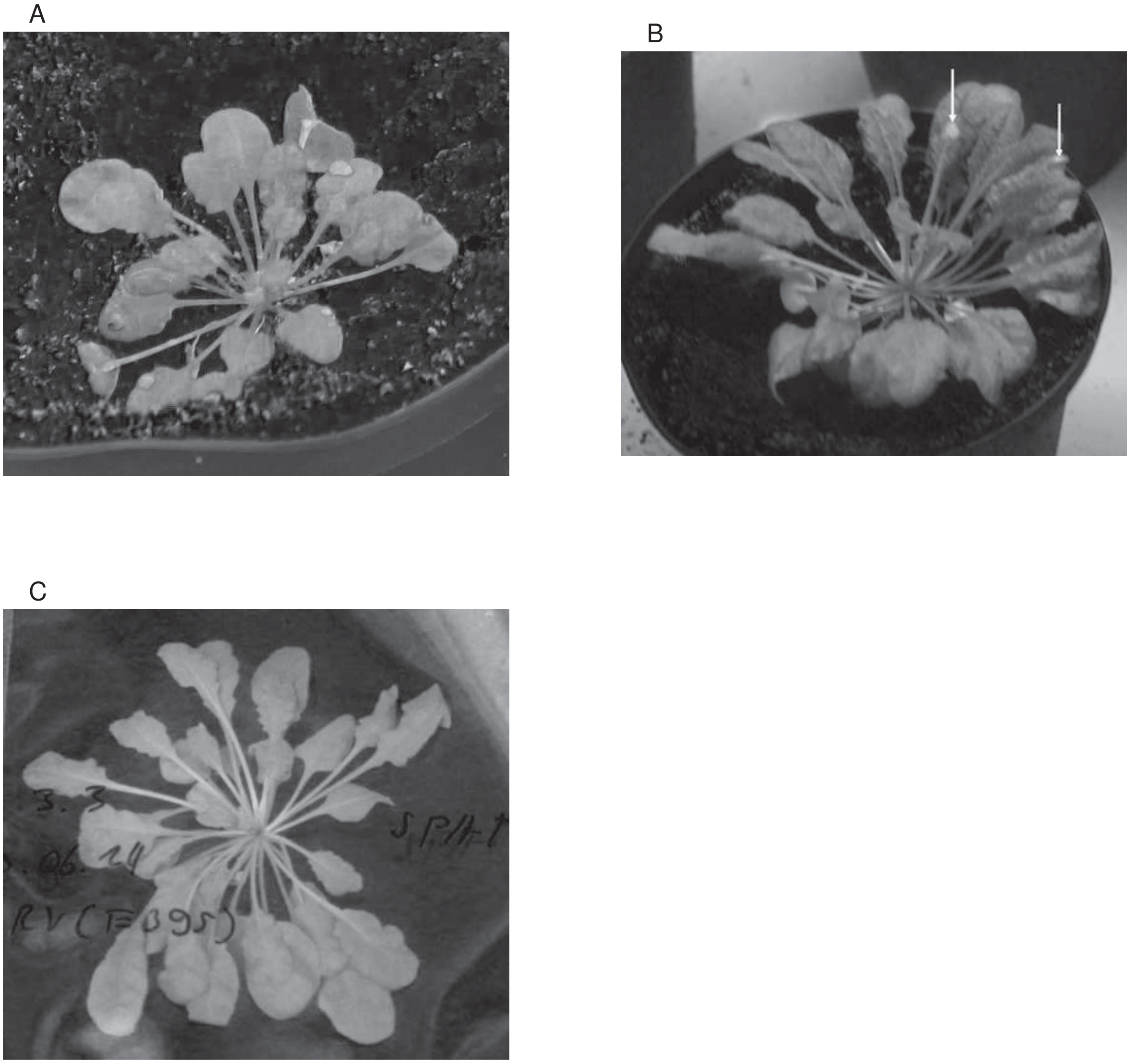


Figure 8: *Arabidopsis thaliana* samples infected with CLRV harvested 0 dpi (A), 24 dpi (B) and 46 dpi (C) are shown. The sample harvested 46 dpi is shown post harvesting while the samples harvested 0 and 9 dpi are depicted shortly before harvesting. The sample depicted in B shows necroses on two leaves (white arrows).

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

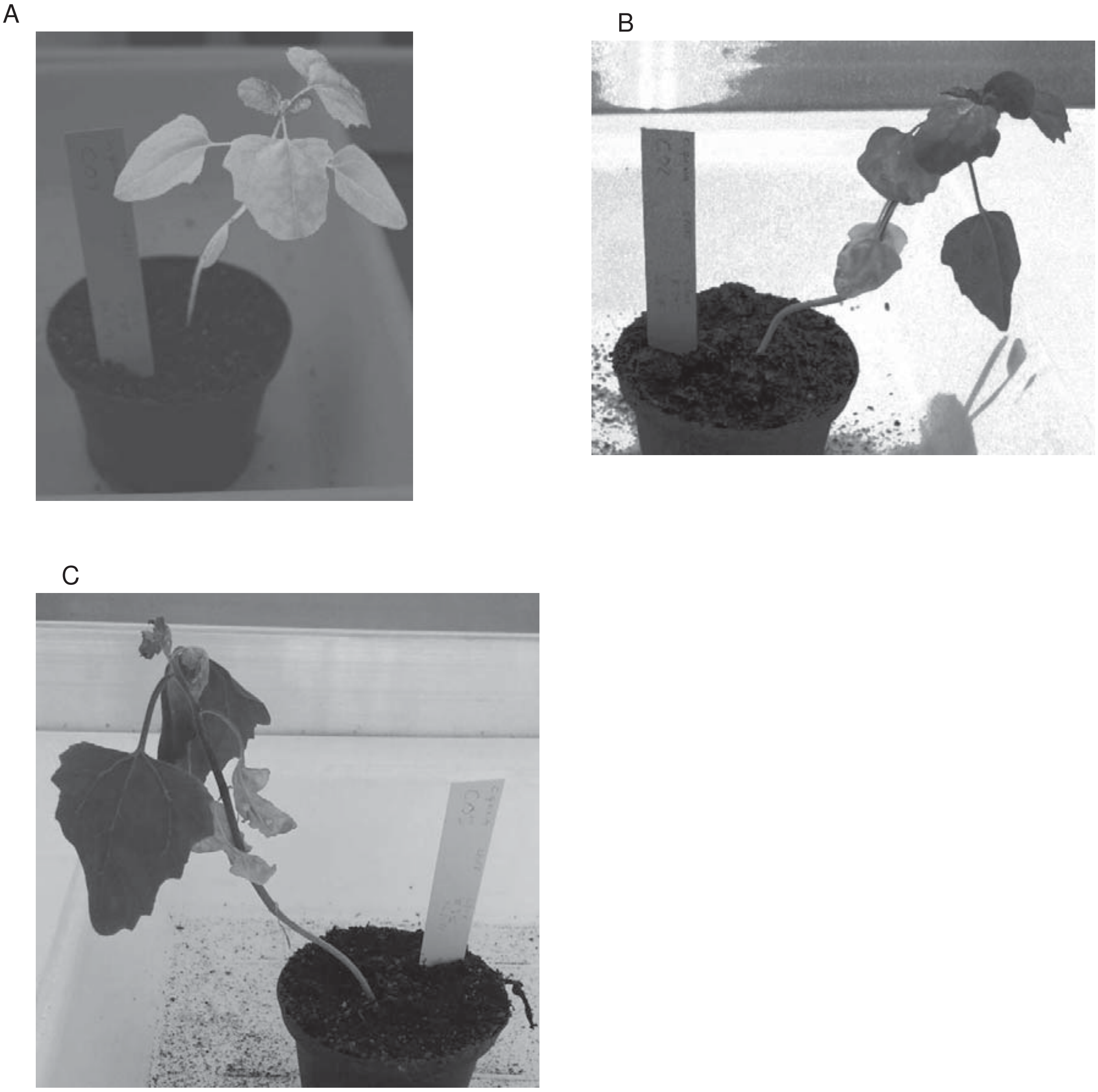


Figure 9: *Chenopodium quinoa* samples infected with CLRV harvested 0 dpi (A), 9 dpi (B) and 14 dpi (C) are shown. The sample shown in B shows chloroses on multiple leaves and presents signs of wilting. The plant in C shows strong signs of wilting in all but two leaves.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

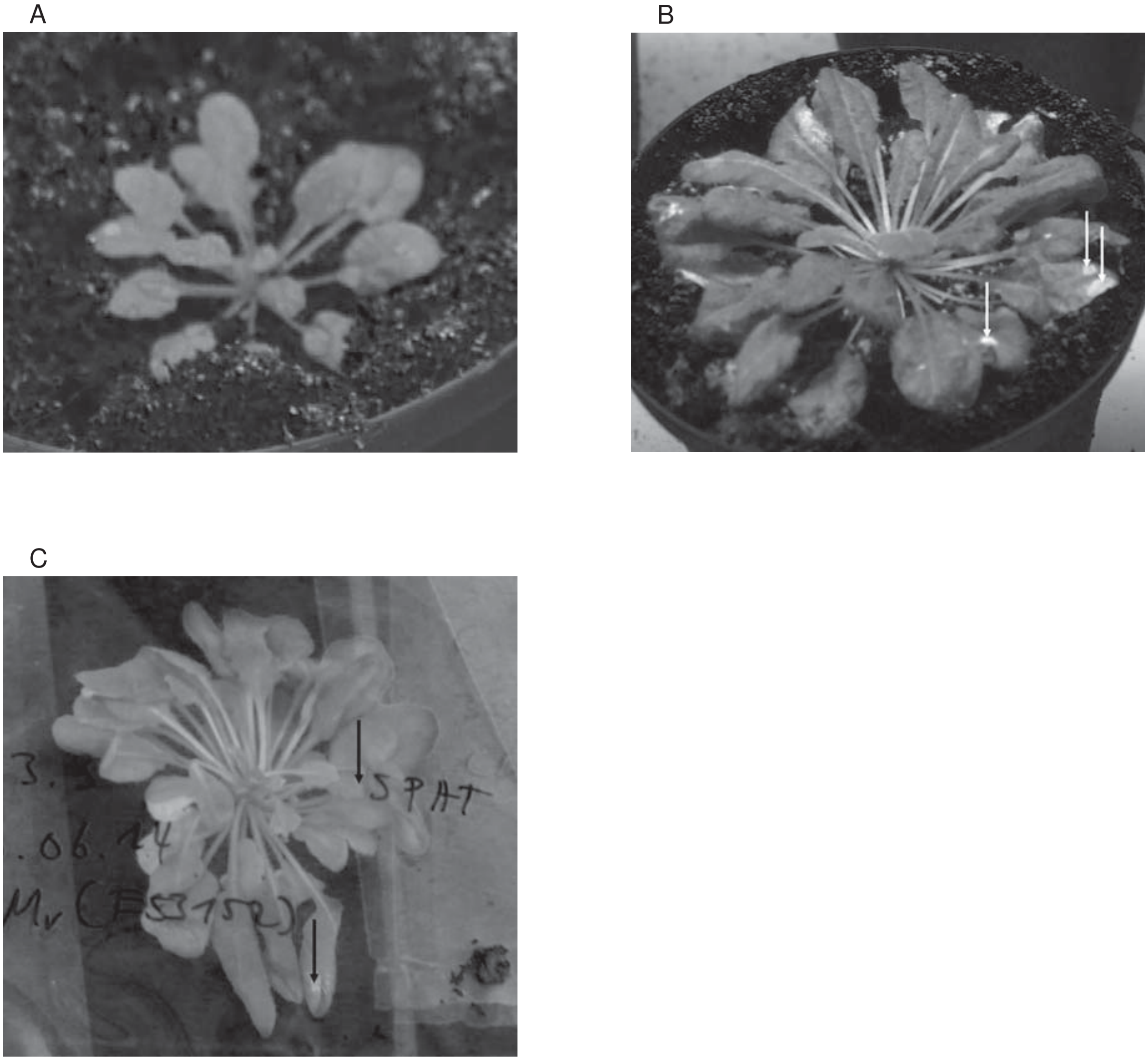


Figure 10: *Arabidopsis thaliana* samples infected with ArMV harvested 0 dpi (A), 24 dpi (B) and 46 dpi (C) are displayed. The plants harvested prior 46 dpi are shown before harvesting, the remaining sample is depicted after harvesting. The sample harvested 24 dpi displays necroses on multiple leaves (white arrows). The sample depicted in C also shows signs of necroses (black arrows).

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

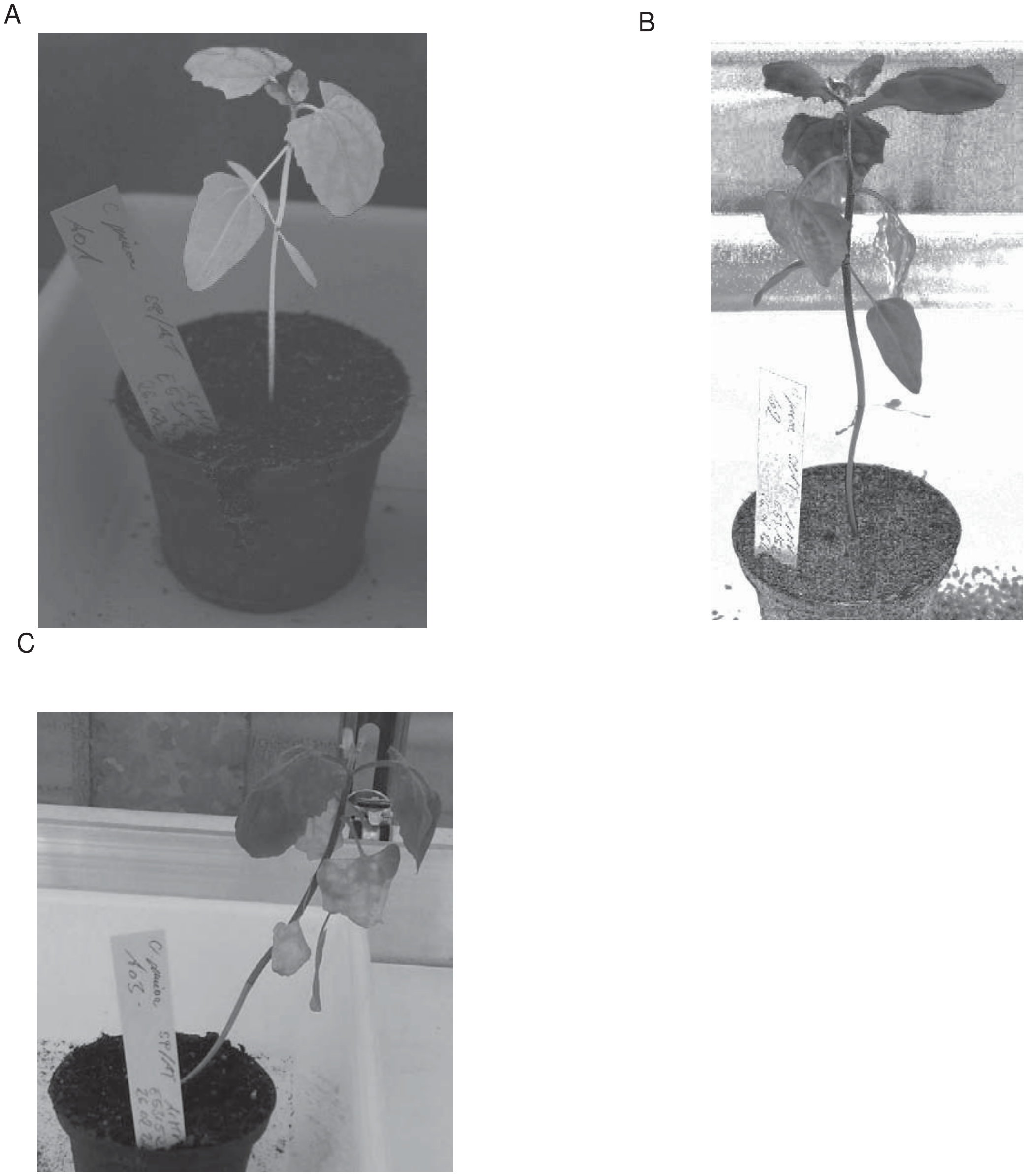


Figure 11: *Chenopodium quinoa* samples infected with ArMV harvested 0 dpi (A), 9 dpi (B) and 14 dpi (C) are shown. The specimen harvested 9 dpi presents chloroses on multiple leaves. The plant shown in C shows signs of wilting as well as chloroses.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

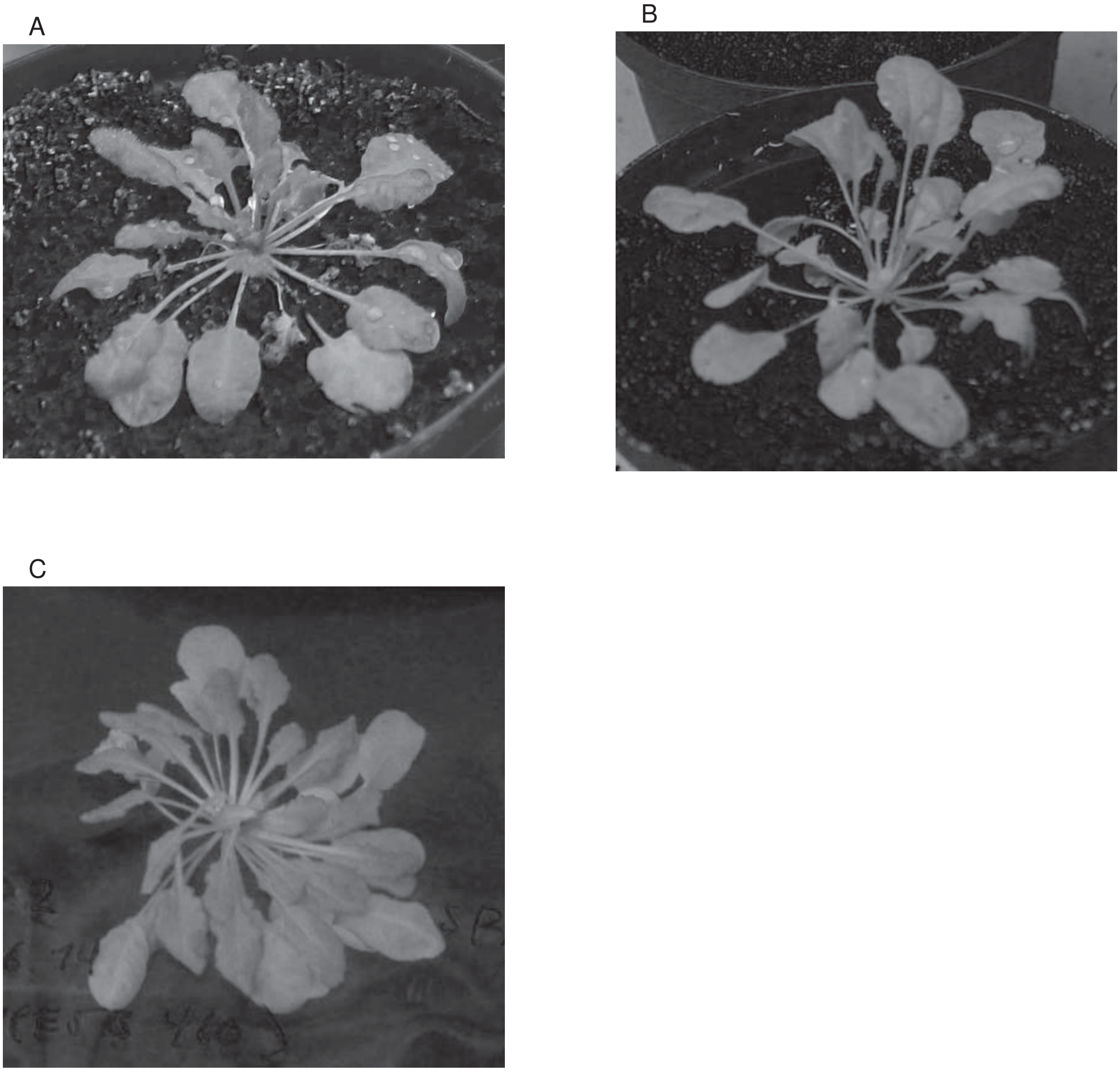


Figure 12: *Arabidopsis thaliana* samples infected with TSWV harvested 0 dpi (A), 24 dpi (B) and 46 dpi (C) are shown. The sample harvested 46 dpi is shown after harvesting. The samples collected 0 and 24 dpi are shown before harvesting. Neither one of the specimen shows obvious signs of viral infection.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

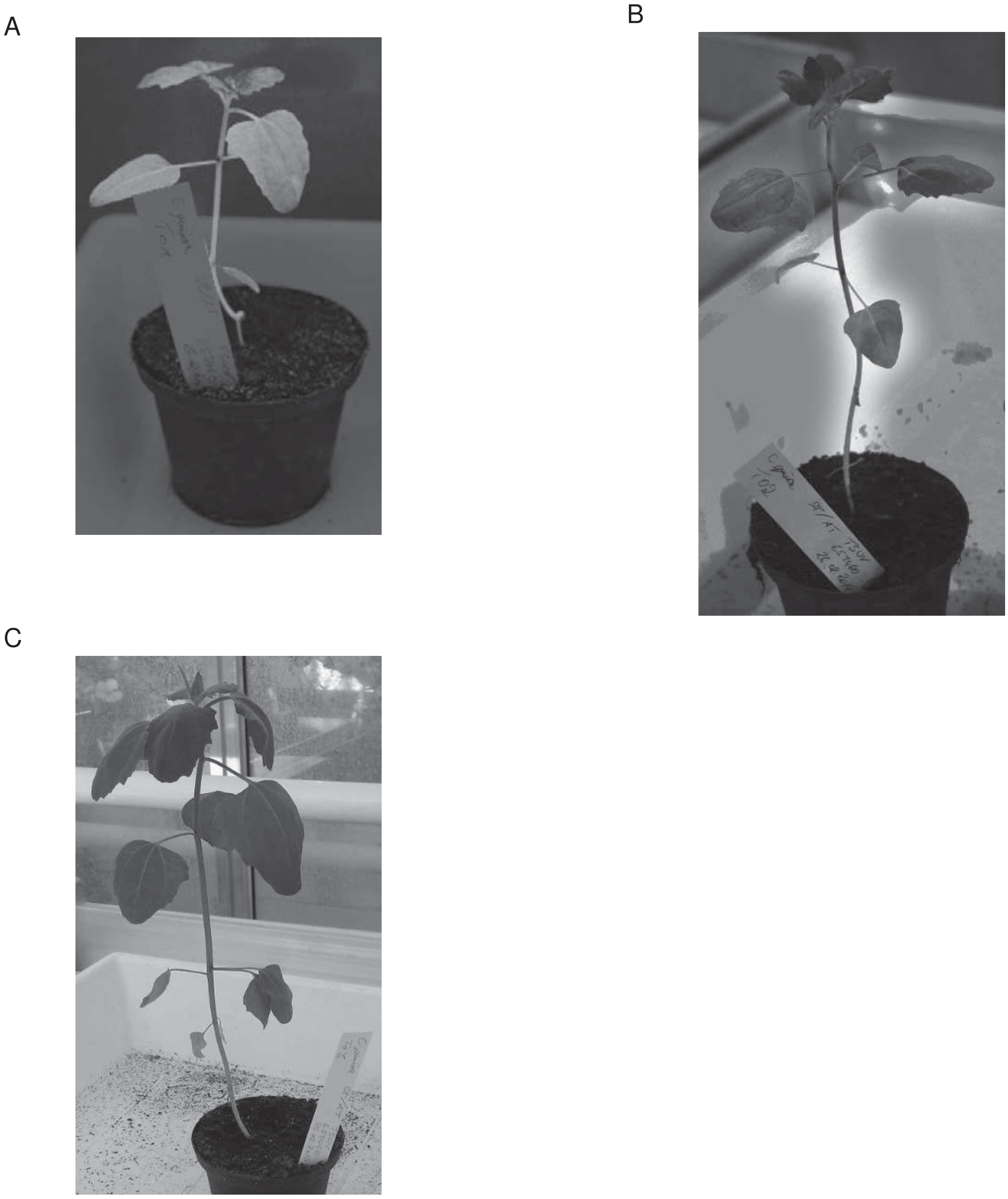


Figure 13: *Chenopodium quinoa* samples infected with TSWV harvested 0 dpi (A), 9 dpi (B) and 14 dpi (C) are displayed. The plants present no discernible symptoms associated with infection.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

3.1 ELISA-detection

In this section the results of the ELISA detections used in section 2.4 are shown. As described the propagated viruses were detected using a DAS ELISA for ArMV and CLRV and a TAS ELISA for TSWV.

3.1.1 DAS-ELISA for the detection of ArMV

Figure 14 shows that the inoculation and subsequent propagation of ArMV was successful. All but one inoculated samples had clearly detectable levels of the virus. The highest optical density (OD) for ArMV isolate E53152 was measured in sample 3.1, which was then used to inoculate the *Arabidopsis thaliana* and *Chenopodium quinoa* samples.

The DAS ELISA detection for ArMV on all samples is shown in figure 15. Using equation 2.3 a cutoff of 0.13 was calculated. The buffer and negative controls are below that cutoff. The positive control as well as samples A01, A02, A03, A2.1, A2.2, A3.1 and A3.2 are above the cutoff.

3.1.2 DAS-ELISA for the detection of TSWV

The DAS ELISA was used to detect TSWV in the propagation samples. However, the results reveal (figure 16) that the method did not work as expected, since the positive control was measured as negative and one of the negative controls was defined as positive. Further analysis revealed that the antibody-enzyme complex used for the reaction showed almost no activity.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

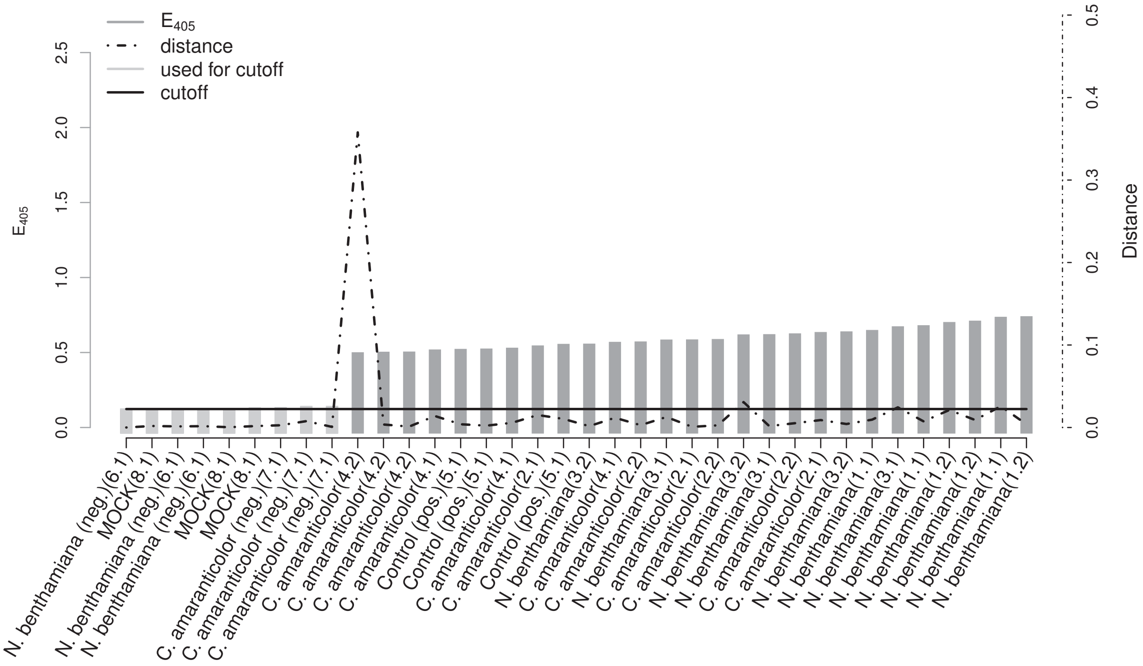


Figure 14: The ELISA results for the ArMV propagation are shown. The sorted E_{405} values as well as their distance to one another is shown. A clear peak is visible in the distance graph (dot-dashed line), showing that values 1 through 9 can be used for cutoff calculation (light gray). The cutoff is 0.1229 (solid line). Sample 3.1 was chosen for further use in inoculation. Samples 6.1 and 7.1 are negative controls and 5.1 is the positive control.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

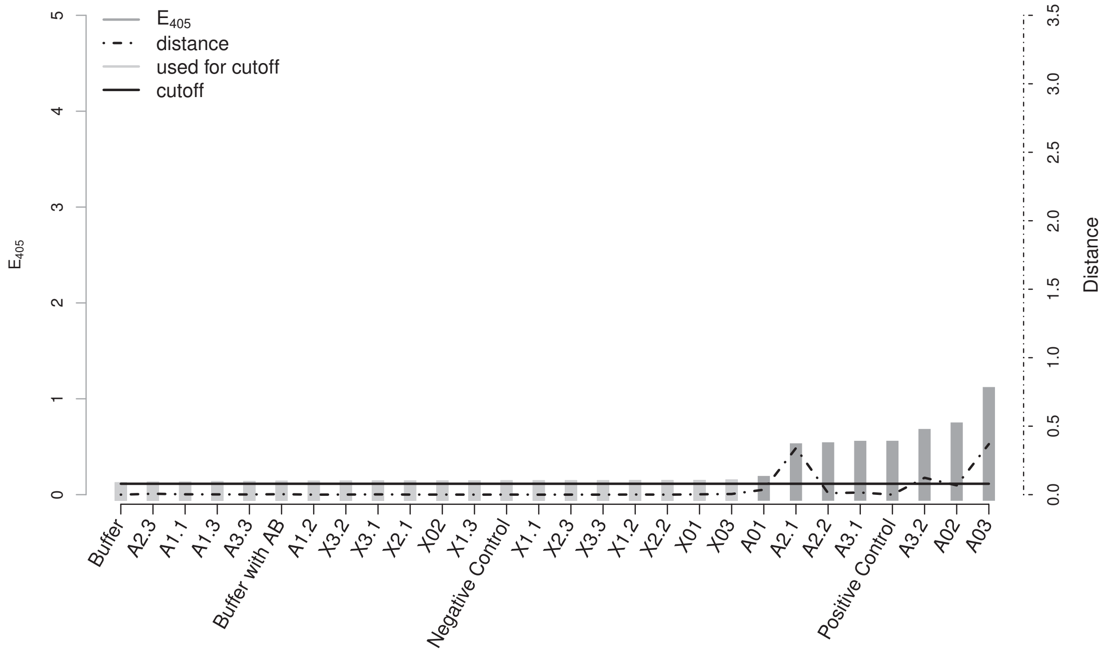


Figure 15: The ELISA results for ArMV are shown. The sorted E_{405} values as well as their distance to one another is shown. A clear peak is visible in the distance graph (dot-dashed line), showing that all values up to A01 can be used for cutoff calculation (light gray). The cutoff is 0.13 (solid line). The buffer shows the lowest E_{405} value. The positive control and samples A01, A02, A03, A2.1, A2.2, A3.1 and A3.2 are above the cutoff.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

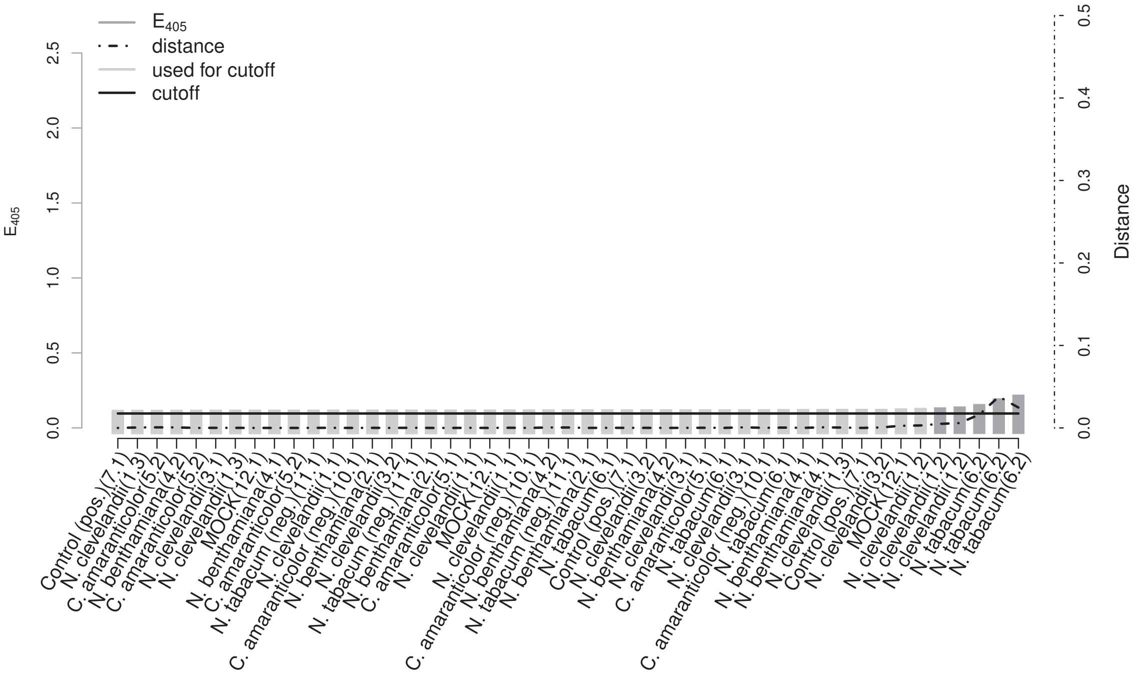


Figure 16: The sorted E_{405} measurements resulting from the DAS ELISA TSWV detection including their respective distances are shown. In sample point MOCK (12.1) a jump in distance can be seen, thus sample points 1 through 46 have been used to calculate a cutoff of 0.0961 (solid line). The positive control (sample 7.1) is below the cutoff.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

3.1.3 DAS-ELISA for the detection of CLRV

Figure 17 shows the results of the DAS ELISA detection for CLRV on all samples. A cutoff of 0.78 could be calculated using equation 2.3. The buffer and negative controls are below that cutoff. The positive control as well as samples X1.2, C02 and C03 are above the cutoff. The positive control as well as the samples C02 and C03 are *Chenopodium quinoa* samples and show a significantly higher extinction value compared to sample X1.2, which, while classified as positive, has an E_{405} of 1.01 compared to 4.25 (positive control), 4.27 (C02) and 4.43 (C03). The distance between respective sample extinction values show multiple smaller peaks. Since the first peak is between the buffer and the sample with the lowest E_{405} value (X01) and further peaks are between control samples and the negative control, thus indicating a significant auto fluorescence, the distance that clearly surpassed those smaller peaks was chosen as linearity boundary.

3.1.4 TAS-ELISA for the detection of TSWV

The TAS ELISA results for the TSWV detection after propagation are summarized in figure 18. The positive and negative controls were categorized correctly and revealed sample 6.1 to be a good candidate for further use. Although the propagation samples were chosen from the *Nicotiana* and the *Chenopodium* families, only *Nicotiana* samples are represented among the positives.

The TAS ELISA detection for TSWV on all samples is shown in figure 19. Using equation 2.3 a cutoff of 0.418 was calculated. The buffer and negative controls are below that cutoff. The positive control is above the cutoff. Since the buffer, the buffer with antibody, the negative control and the positive control have been categorized correctly and the highest extinction value after the positive control is buffer with antibody, the ELISA test did perform within the expected parameters and shows that the infection with TSWV did not succeed in any sample.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

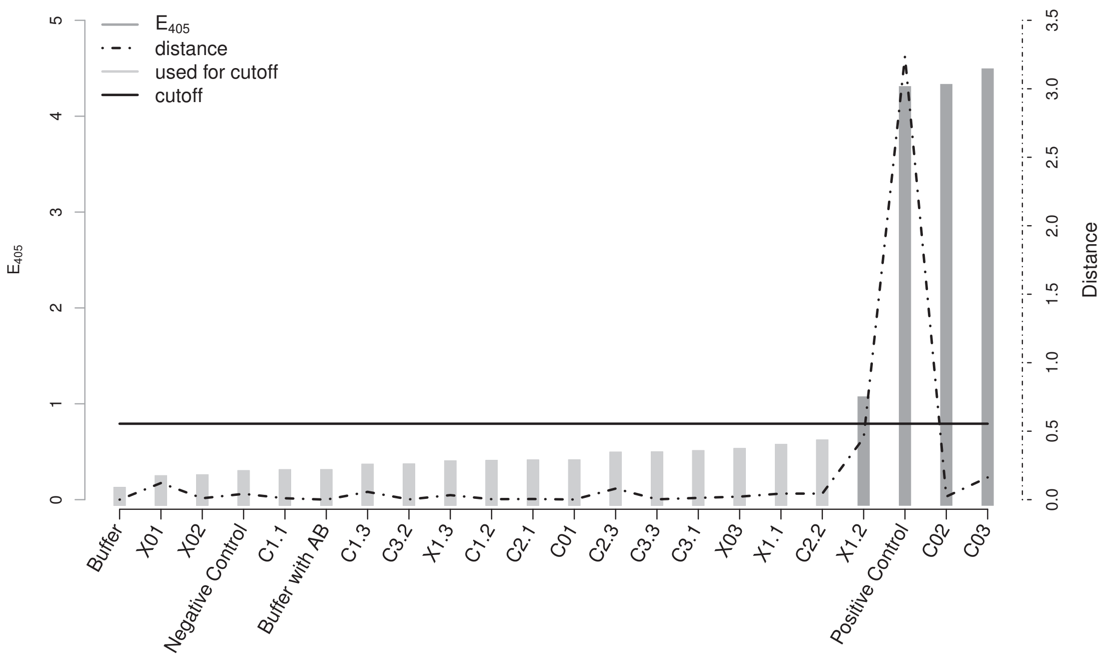


Figure 17: The ELISA results for CLRV are shown. The sorted E_{405} values as well as their distance to one another is shown. A clear peak is visible in the distance graph (dot-dashed line), showing that all values up to X1.2 can be used for cutoff calculation (light gray). The cutoff is 0.78 (solid line). The buffer shows the lowest E_{405} value. The positive control and samples X1.2, C02 and C03 are above the cutoff.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

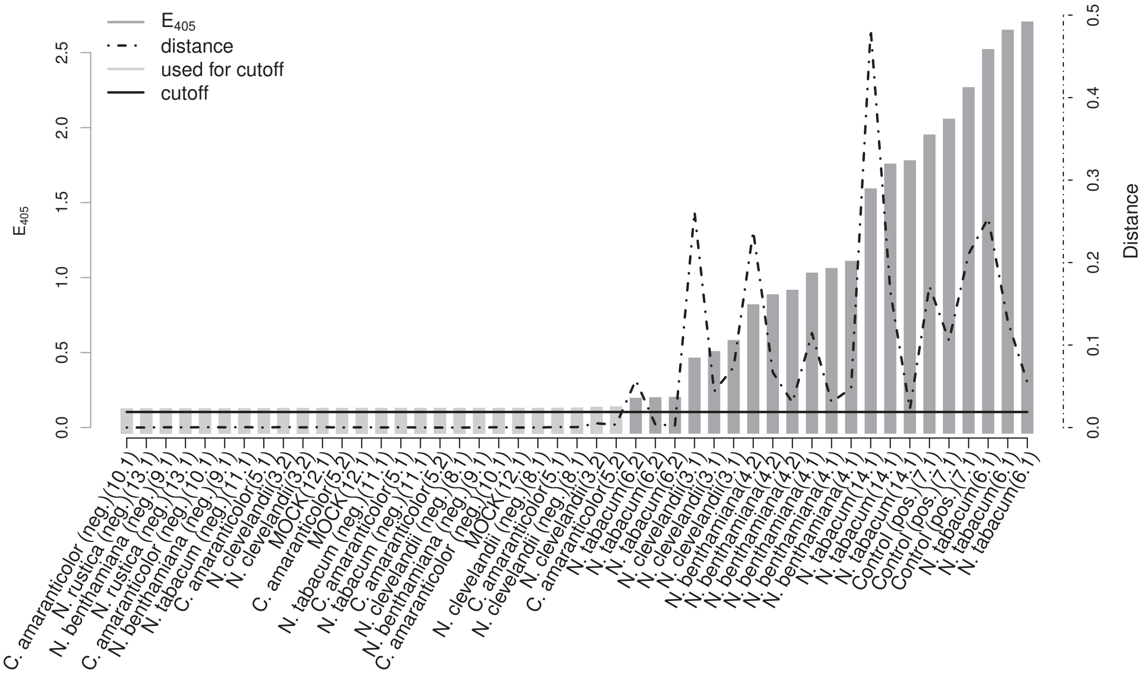


Figure 18: The results of the TAS ELISA TSWV detection after virus propagation is shown. The E_{405} measurements are shown in sorted order. Also the distance between the sample points is shown. The distance between sample points 5.2 and 6.2 shows the first peak, thus the first 27 sample points were used to calculate a cutoff of 0.1037 (solid line). Sample 7.1 is the positive control, samples 8.1, 9.1, 10.1 and 11.1 are negative controls. Sample 6.1, showing the highest E_{405} value, was used for further inoculation.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

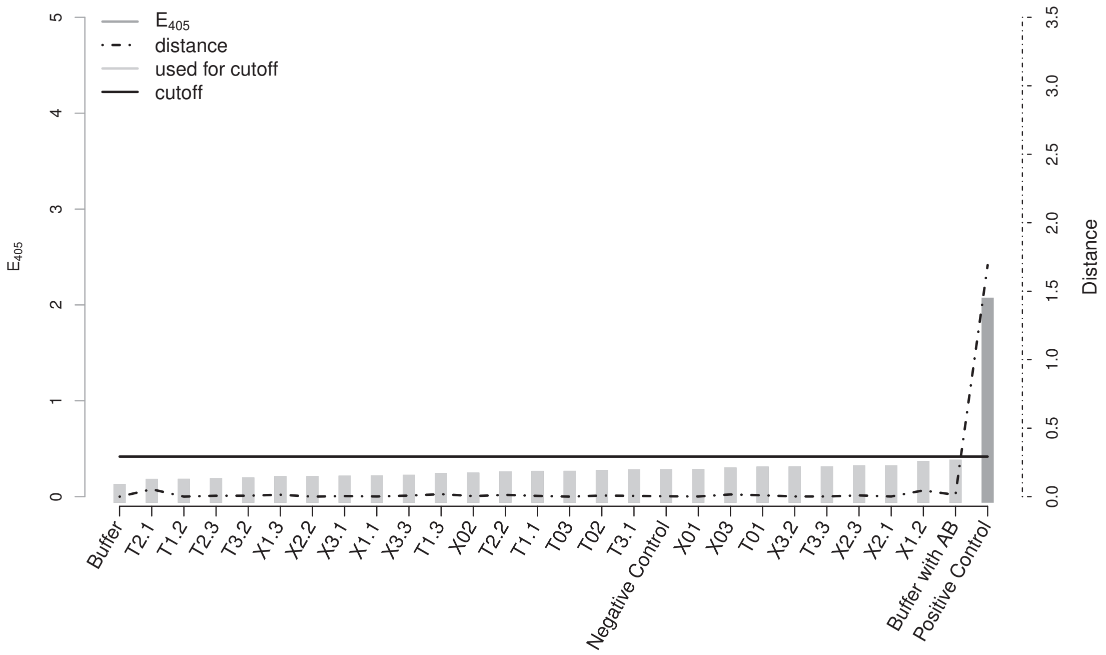


Figure 19: The ELISA results for TSWV are shown. The sorted E_{405} values as well as their distance to one another is shown. A clear peak is visible in the distance graph (dot-dashed line), showing that all values up to the positive control can be used for cutoff calculation (light gray). The cutoff is 0.418 (solid line). The buffer shows the lowest E_{405} value. The positive control is above the cutoff. The remaining samples are below the cutoff.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

3.2 Sequencing

262.281.477 paired end reads were sequenced, ranging up to 14 million reads per sample with an average read length of 133.5 bases (table 9). After all *Arabidopsis thaliana* samples were aligned, the *Chenopodium quinoa* samples yielded a four fold decrease in UMRs and a three fold decrease in MMRs in the alignment against said reference genome, compared to *Arabidopsis thaliana*.

3.3 Bioinformatical Analysis

The PCA maps being shown in the next subsections are based on two remaining dimensions. The components which were used to arrive at these maps were chosen based on the lowest 25 pValues from the training set (solid shapes in the maps). The pValues were not corrected against multiple testing, because the value of the pValue was not considered, just the order which is not effected by multiple testing. Using those components from the test set allowed the projection of those samples into the existing maps (hollow shapes in the maps).

3.3.1 Transcriptome Analysis

The expression analysis was performed on *arabidopsis thaliana core v29.82.10*, which contains 33.602 gene models (protein coding genes, transposable elements and rcRNA). The expression value for each of those 33.602 entities was used to construct the sample vectors. Those were used to calculate the PCA maps shown in figures 20 through 26.

The map in figure 20 shows the samples grouped by their species. The best 25 genes for this classification, having a pValue smaller than or equal to $1.96 \cdot 10^{-4}$, were used to construct this classifier. Two distinct and narrowly spread point clouds

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

can be seen. The *Chenopodium quinoa* samples are located to the right of the plot and range from quadrant I to quadrant IV. The point cloud formed by the *Arabidopsis thaliana* samples is located in the left part of the plot intersecting with quadrants II,III and IV. The test samples and respective training samples show a small distance. According to the STRESS plot, the three dimensional representation is the ideal trade off. Using two dimensions a STRESS of smaller than 0.3 and an r^2 value of 0.85 can be calculated.

In figures 21 through 23 the PCA maps of the age classification can be seen. A multi stage approach shows the classification of Time Point (TP) one and two (S1: figure 21), one and three (S2: figure 22) and two and three (S3: figure 23). The S1 PCA map shows two, partly overlapping, narrowly spread point clouds. The majority of TP1 samples are located in quadrant II and III, while most of the TP2 samples are spread through quadrants I and IV. The greatest overlap is among samples belonging to *Chenopodium quinoa*. The corresponding STRESS plot shows an ideal trade off using three principal components rather than two. The best 25 genes being used for the construction of this classifier had a pValue of smaller or equal to $1.16 \cdot 10^{-01}$. S2 shows two sample groups with minimal overlap in quadrant I. The TP3 samples are located in the left area of the plot. The TP1 samples can be found in the right area of the plot. The test and training samples are positioned relatively close to one another. The STRESS using two dimensions is above 0.3 and the corresponding r^2 value is 0.82. A pValue equal to or smaller than $9.25 \cdot 10^{-03}$ was calculated for the 25 used genes. The map for S3 shows a STRESS of greater than 0.3 and an r^2 value of 0.84. The TP2 samples are located close to the positive side of the PC1 axis. The TP3 samples are wide spread through quadrant II and III. An overlap of three points and their respective test samples are located within the TP2 grouping. The genes used to construct this classifier have a pValue of $2.59 \cdot 10^{-02}$ or smaller.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

The classification of infecting virus is based on a multi stage construct as well. The classifiers differentiate between ArMV infected samples and all others (B1: figure 24), CLRV infected samples and all others (B2: figure 25) and infected samples against all others (B3: figure 26). A subset of samples (cross in figures) has been removed from the initial calculations, since, due to the partially conflicting results of ELISA and pathogen alignment, the true infection state was uncertain. Those samples were considered additional test samples in this scope. B1 shows the ArMV infected samples in the east quadrants, mainly in quadrant IV. The not ArMV infected samples are spread over quadrants I, II and III mainly. Two training points and three test points of the ArMV infected samples are located very close or within the area occupied by the other group. Among those points is the test sample A03, however the training counter part is located in the lower right corner with a great distance. A pValue equal to or smaller than $3.53 \cdot 10^{-02}$ was calculated for the 25 genes constructing this classifier. The r^2 value using two dimensions is 0.87 with a corresponding STRESS of less than 0.3. The classifier B2 shows the CLRV infected samples at the outer right edge of the point cloud produced by the not CLRV infected samples. The genes used for this classifier have a pValue smaller than or equal to $7.0 \cdot 10^{-02}$. The STRESS for this map is lower than 0.25 and the r^2 value is 0.9. The classification of not infected vs. infected samples shows two wide spread point clouds. The not infected samples are located at the right end of the plot, while the infected samples can be found in the center and left area of the plot. The STRESS is greater than 0.3 and the r^2 value is 0.835. The 25 best genes showed a pValue of smaller than or equal to $1.04 \cdot 10^{-01}$.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

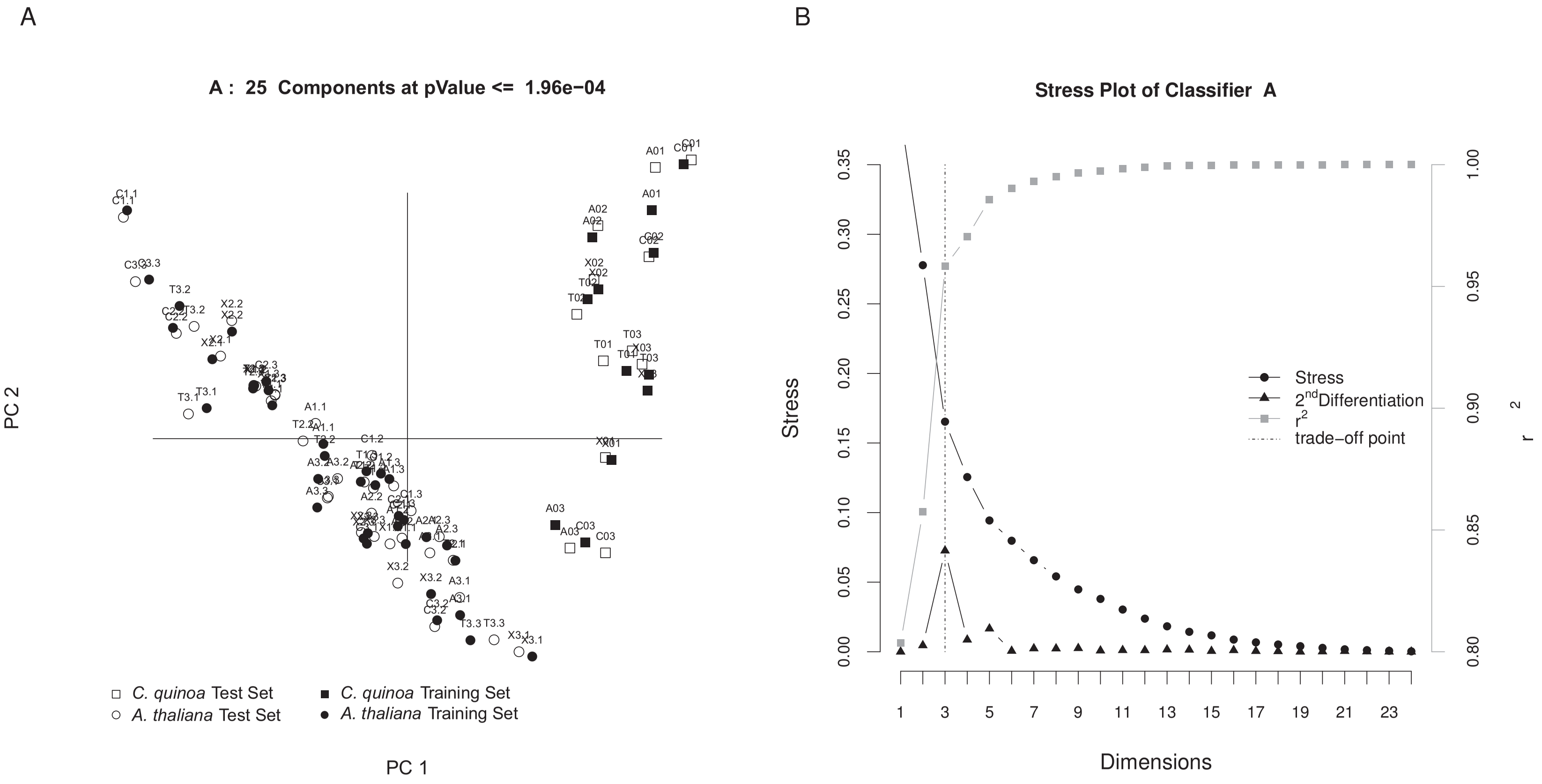


Figure 20: The species classifier (A) is shown. A shows the PCA of *Arabidopsis thaliana* (circles) vs. *Chenopodium quinoa* (squares). 25 components were used to calculate the map, all of which had a pValue smaller or equal to $1.94 \cdot 10^{-4}$. The data of *Arabidopsis thaliana* are spread mainly through the west quadrants in a narrow point cloud. The *Chenopodium quinoa* points are positioned to the right of the graph, also in a narrow cloud. The projection points (hollow shapes) are positioned within their respective grouping. The STRESS plot shown in B shows an ideal trade off point (dot-dashed line) at three dimensions, however the STRESS at two dimensions is below 0.3 and the r^2 value is 0.85.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

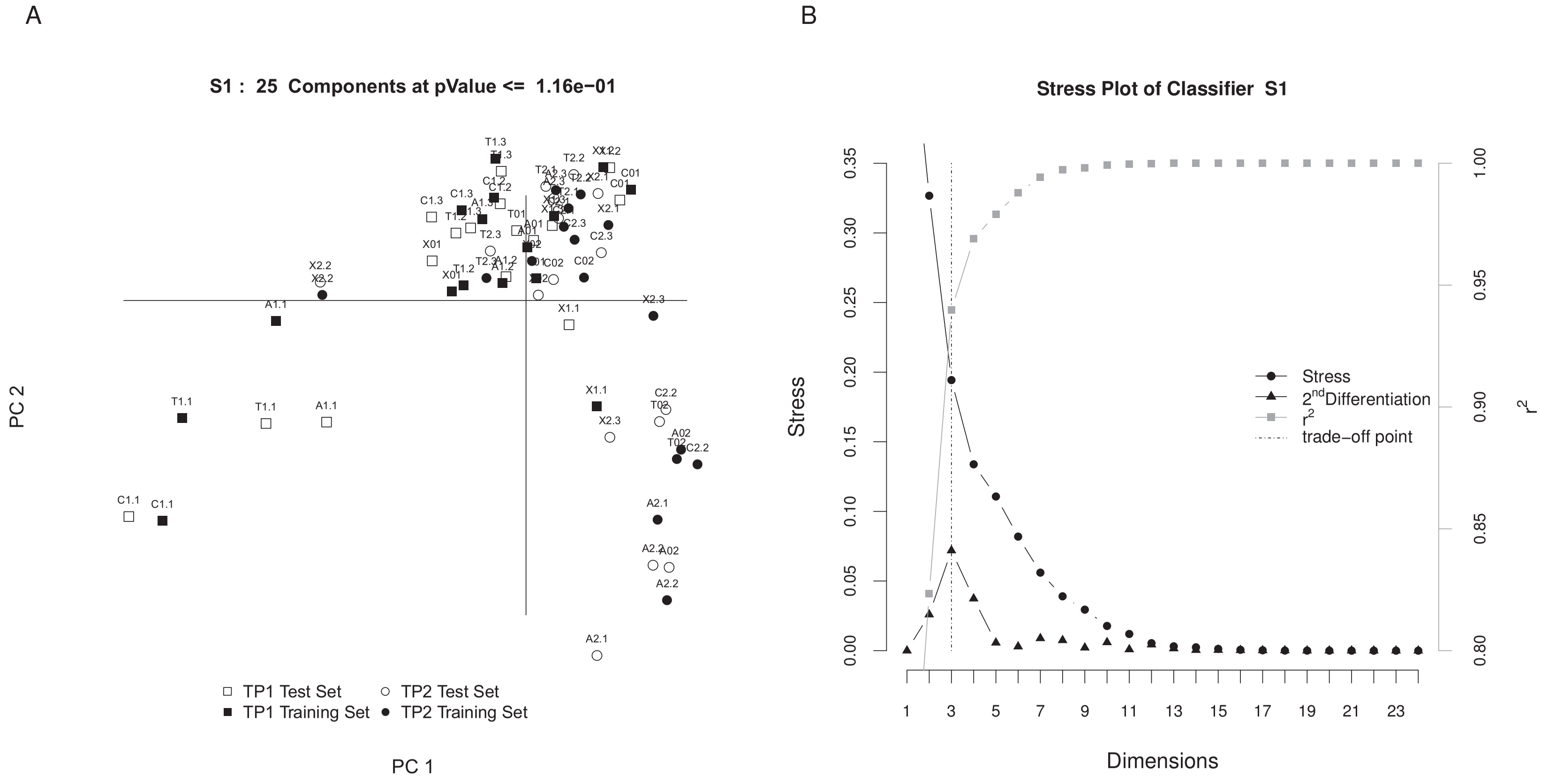


Figure 21: The age classifier (S1) differentiating between TPs 1 (squares) and 2 (circles) is shown in A. 25 components were used to derive this map having a pValue equal to or smaller than $1.16 \cdot 10^{-01}$. The majority of TP1 points are located in the quadrants II and III. The points of TP2 are mainly positioned in quadrant I and IV. Most points are clustered in quadrant I and II. All projection points (hollow shapes) can be found in relative proximity to their respective training points (solid shapes). The STRESS plot (B) shows an ideal tradeoff at three dimensions. At two dimensions the STRESS is above 0.3 and the r^2 value is 0.815.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

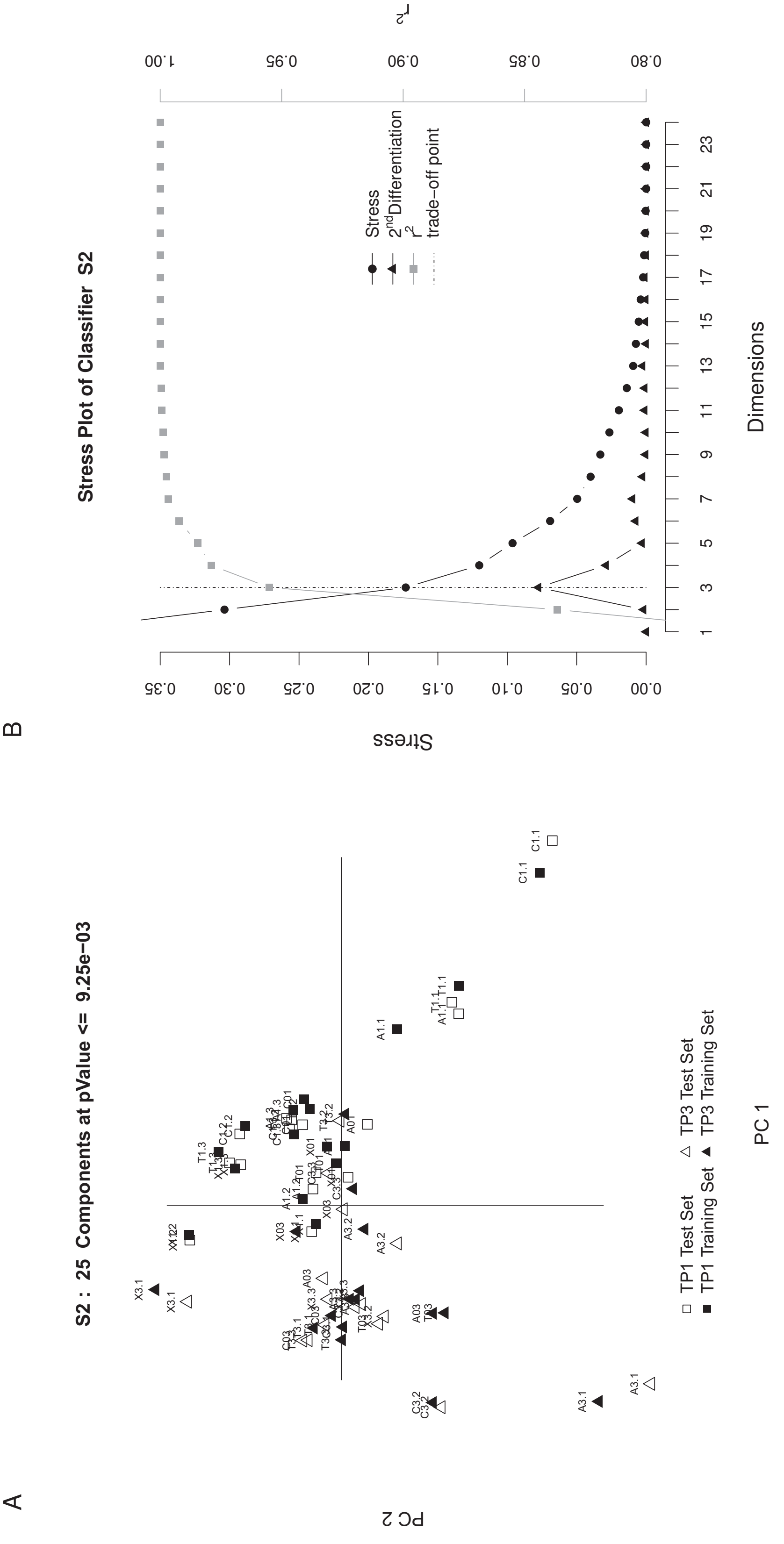


Figure 22: The age classifier (S2) differentiating between TPs 1 (squares) and 3 (triangles) is shown in A. 25 components were used to derive this map having a pValue equal to or smaller than $9.25 \cdot 10^{-03}$. The majority of TP1 points are located in the quadrants I and IV. The points of TP3 are mainly positioned in quadrant II and III. All test sample points (hollow shapes) can be found in relative proximity to their respective training points (solid shapes). The STRESS plot (B) shows an ideal tradeoff at three dimensions. At two dimensions the STRESS is above 0.3 and the r^2 value is 0.82.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

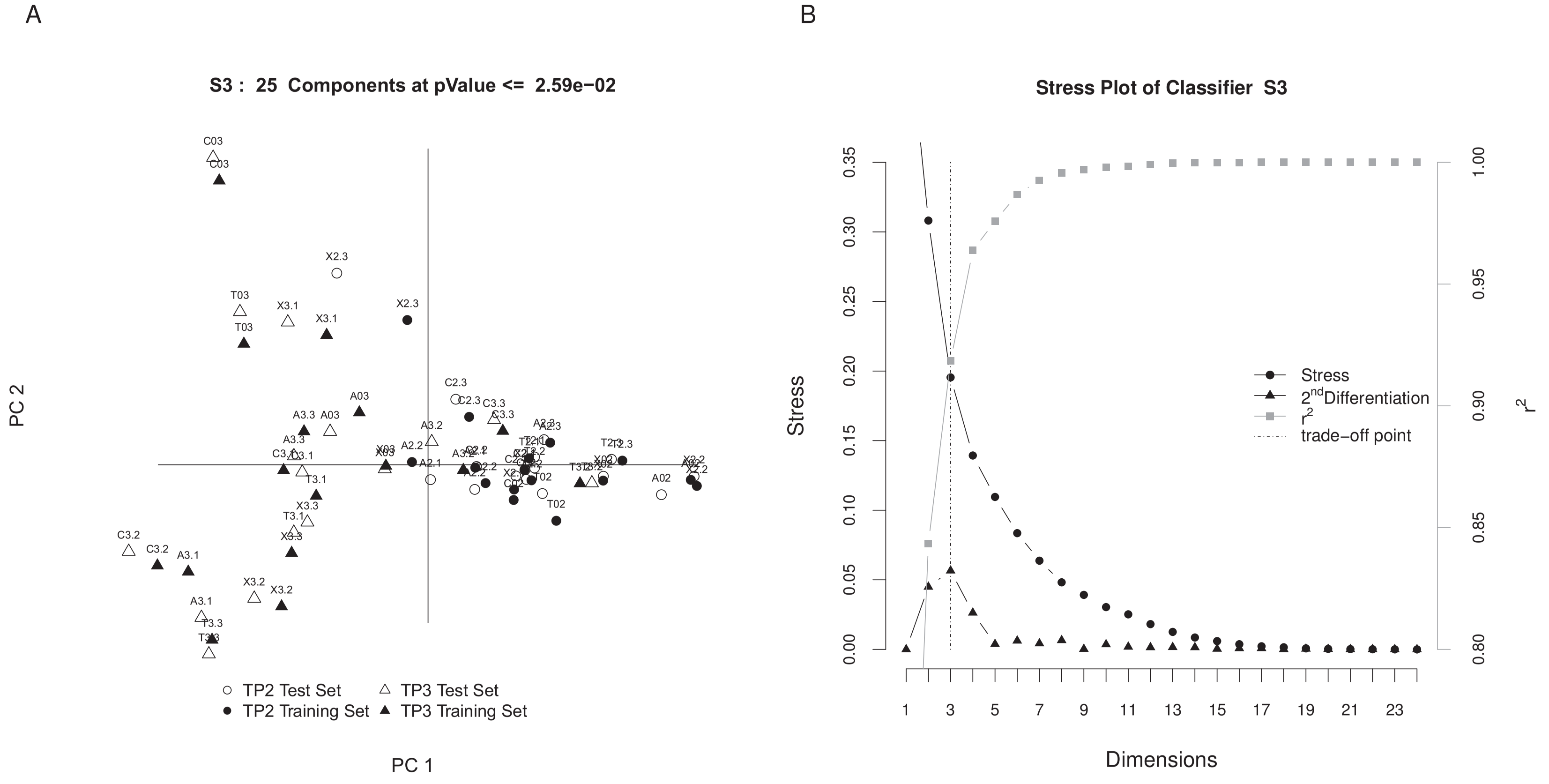


Figure 23: The age classifier (S3) differentiating between TPs 2 (circles) and 3 (triangles) is shown in A. 25 components were used to derive this map having a pValue equal to or smaller than $2.59 \cdot 10^{-02}$. The majority of TP2 points are located along the positive PC1 axis. The points of TP3 are mainly spread through quadrants II and III. Three TP3 points and their respective test set points are located among the TP2 points. All test sample points (hollow shapes) can be found in relative proximity to their respective training points (solid shapes). The STRESS plot (B) shows an ideal tradeoff at three dimensions. At two dimensions the STRESS is above 0.3 and the r^2 value is 0.84.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

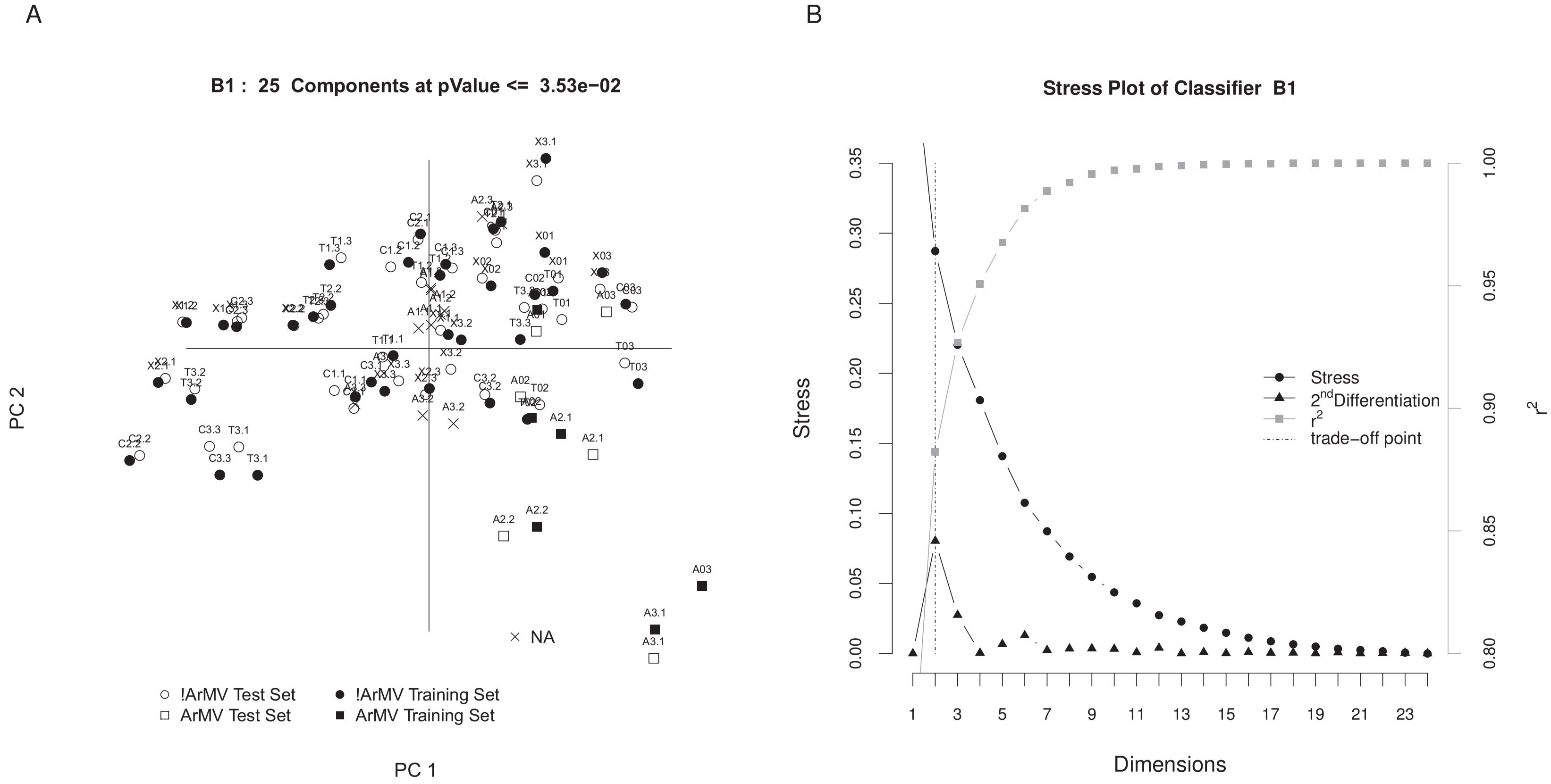


Figure 24: B1, the classifier differentiating between ArMV infected samples (squares) and not ArMV infected samples (circles) (all other samples) is shown. The map in A shows a wide spread point cloud of not ArMV infected samples through the quadrants I, II and III. The ArMV infected samples are, apart from one training point (solid squares) and two test points (hollow squares) positioned in quadrant IV. Samples not used for the construction of the classifier are represented as crosses. The test samples can be found close to the respective training points. A03 shows a very large distance between test (quadrant I) and training sample (lower right corner). The two dimensional representation is the ideal tradeoff according to the STRESS plot in B. The STRESS is less than 0.3 and the r^2 value is 0.87.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

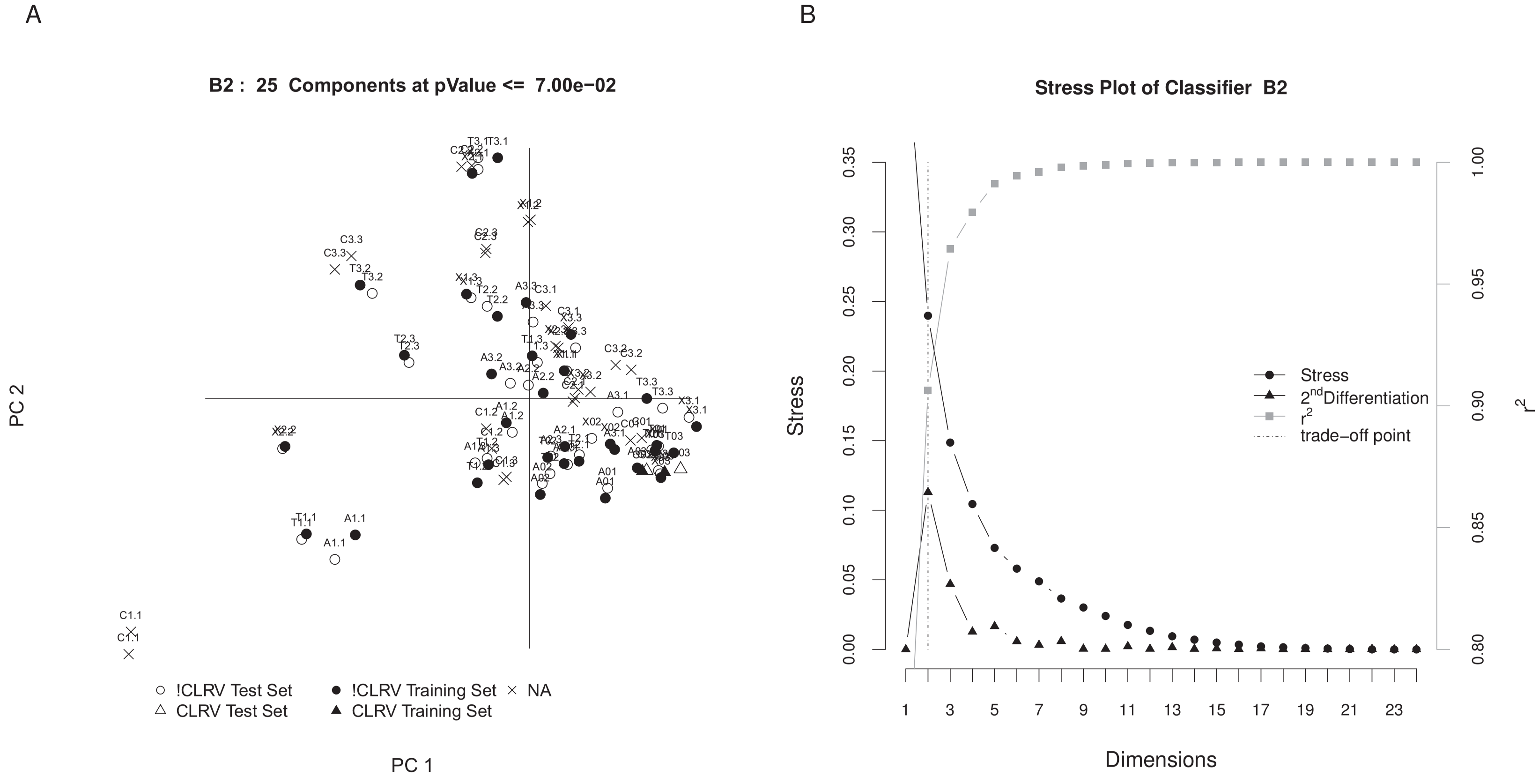


Figure 25: B2, the classifier differentiating between CLRV infected samples (triangles) and not CLRV infected samples (circles) (all other samples) is shown. The map in A shows a wide spread point cloud of not CLRV infected samples through all quadrants. The CLRV infected samples are close to one another at the lower right edge of the not infected sample point cloud. Samples not used for the construction of the classifier are represented as crosses. The test samples can be found close to the respective training points. The two dimensional representation is the ideal tradeoff according to the STRESS plot in B. The STRESS is less than 0.2 and the r^2 value is 0.9.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

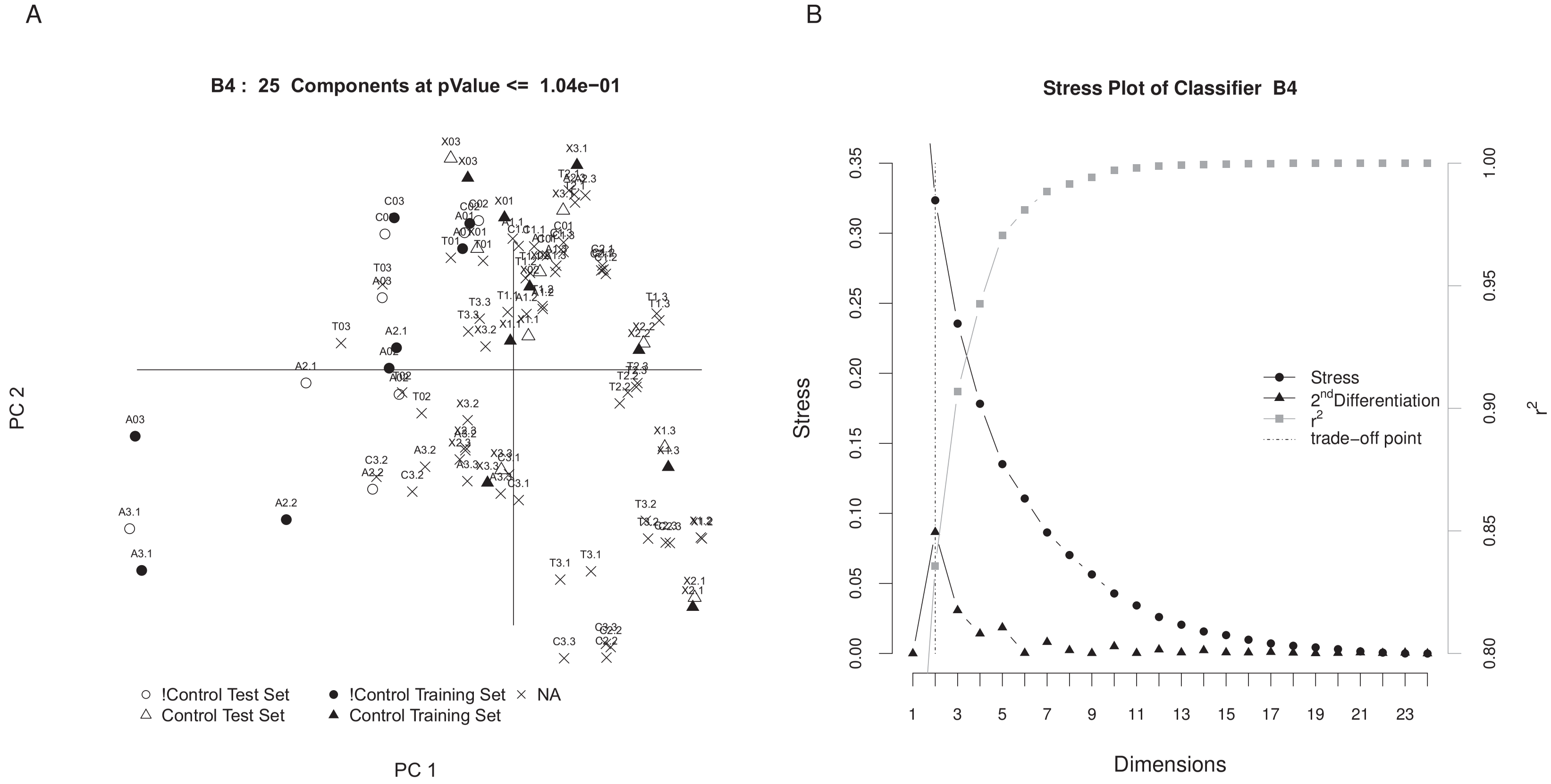


Figure 26: B4, the classifier differentiating between infected samples (circles) and not infected samples (triangles) is shown. The map in A shows a wide spread of all points. The not infected samples are located to the right of the plot in a horizontally confined area. The infected samples are located in the left area of the graph with a large horizontal spread. Samples not used for the construction of the classifier are represented as crosses. The test samples can be found close to the respective training points, apart from samples A2.1 and A03 which show a greater distance between trainings and test points. The two dimensional representation is the ideal tradeoff according to the STRESS plot in B. The STRESS is greater than 0.3 and the r^2 value is 0.835.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

3.3.2 n-Gram Based Sequence Profiling

The sequencing data was transformed into 4.194.304 profiles of size 11. The count set of each profile was used to arrive at a 4.194.304 dimensional sample vector. These sample vectors were the basis for the analyses resulting in the PCA maps depicted in figures 27 through 33.

The classification in figure 27, grouping *Chenopodium quinoa* and *Arabidopsis thaliana*, shows two distinctly different point clouds. The points representing *Arabidopsis thaliana* samples (circles) are widely spread over all quadrants with a relatively big distance amongst individual points. The points representing the *Chenopodium quinoa* samples (squares) are tightly grouped in a narrow band located to the far center right of the graph. One sample (C1.1) is erroneously located in this grouping. All other sample points, training and test, can be found within the area clearly defined by their respective group. With a pValue smaller than or equal to $2.56 \cdot 10^{-13}$ 25 components were used to construct the depicted two dimensional PCA map with a STRESS of under 0.1 and a r^2 value of 0.985.

In figures 28 through 30 the PCA maps of the age classification can be seen. A multi stage approach shows the classification of TP one and two (S1: figure 28), one and three (S2: figure 29) and two and three (S3: figure 30). S1 (figure 28) shows the majority of TP2 samples in quadrant II, while the TP1 samples are spread mainly over quadrants I and IV. The test samples are located relatively close to the training samples and are within the area defined by the corresponding group. The 25 utilized components had a pValue of smaller than or equal to $8.88 \cdot 10^{-05}$. The STRESS at the ideal trade off of two dimensions is below 0.25 and the r^2 value is 0.915. Figure 29 shows the PCA map based on the best 25 components (pValue: $1.16 \cdot 10^{-06}$ or smaller) for classifier S2. The TP1 samples show a wide vertical spread through quadrants II and III. The TP3 samples are spread horizontally around the PC1 axis. An overlap can be seen in samples A1.2 (test), T1.2 (test) and T3.2 (test). The

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

STRESS is 0.16 and the r^2 value is 0.96 at the ideal trade off of two dimensions. The classification shown in figure 30 has been produced by the 25 best components for classifier S3 (pValue: $1.08 \cdot 10^{-04}$). The TP3 samples are spread in an arch over quadrants II and III with sample X3.1, located at the bottom of the graph, having the greatest distance within the group. The TP2 samples are positioned in two distinct groupings. The majority of samples can be found at center right of the graph, however a subset, made up of all *Chenopodium quinoa* samples, is located at the top middle of the graph. An overlap of two test samples (T02 and A02) can be seen in this subset. The STRESS of 0.225 and r^2 value of 0.92 can be seen at two dimensions, however the ideal trade off is calculated to be at three dimensions.

The classification of the infecting virus is based on a multi stage construct as well. The classifiers differentiate between ArMV infected samples and all others (B1: figure 24), CLRV infected samples and all others (B2: figure 25) and infected samples against all others (B3: figure 26). Here too, a subset of samples (cross in figures) has been removed from the initial calculations because of an uncertain infection state due to the partially conflicting results of ELISA and pathogen alignment. Figure 31 represents the two dimensional PCA map, with a STRESS of less than 0.25 and a r^2 value of 0.95, of the 25 best components (pValue: $1.95 \cdot 10^{-14}$) of B1. The not ArMV infected samples are grouped into two point clouds. The majority of samples are spread vertically over quadrants II and III. A subset, made up of *Chenopodium quinoa* samples exclusively, can be found in a more condensed area in the center of the map. The ArMV infected samples are positioned in a confined area to the far center right of the map. Two test samples (A01 and A3.1) have a big relative distance to their group. The independent sample A3.2 is located relatively close to the ArMV infected sample set. The PCA map in figure 32 is based on the 25 best components (pValue: $2.96 \cdot 10^{-15}$) derived from classifier B2 and has a STRESS value less than 0.25 and a r^2 value of 0.955. The CLRV infected samples are located to the center left of the map. All other samples are spread mostly in quadrants I and

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

IV. The samples X1.1, T03, X03 and T02 (test and training) follow a narrow band, crossing quadrants II and III and show a relatively big distance to the other samples. The classifier differentiating between infected and not infected samples (B4) is represented in figure 33. The PCA has a STRESS value of less than 0.3 and a r^2 value of 0.925. The majority of not infected samples are spread over quadrants I and IV. Sample X03 is relatively far away from the other samples within this set. The infected samples are located in two groupings. Most of the samples are in a tight cluster to the far left of quadrant II. Samples C02, C03 and A01 make up a distinct grouping located around the left PC1 axis.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

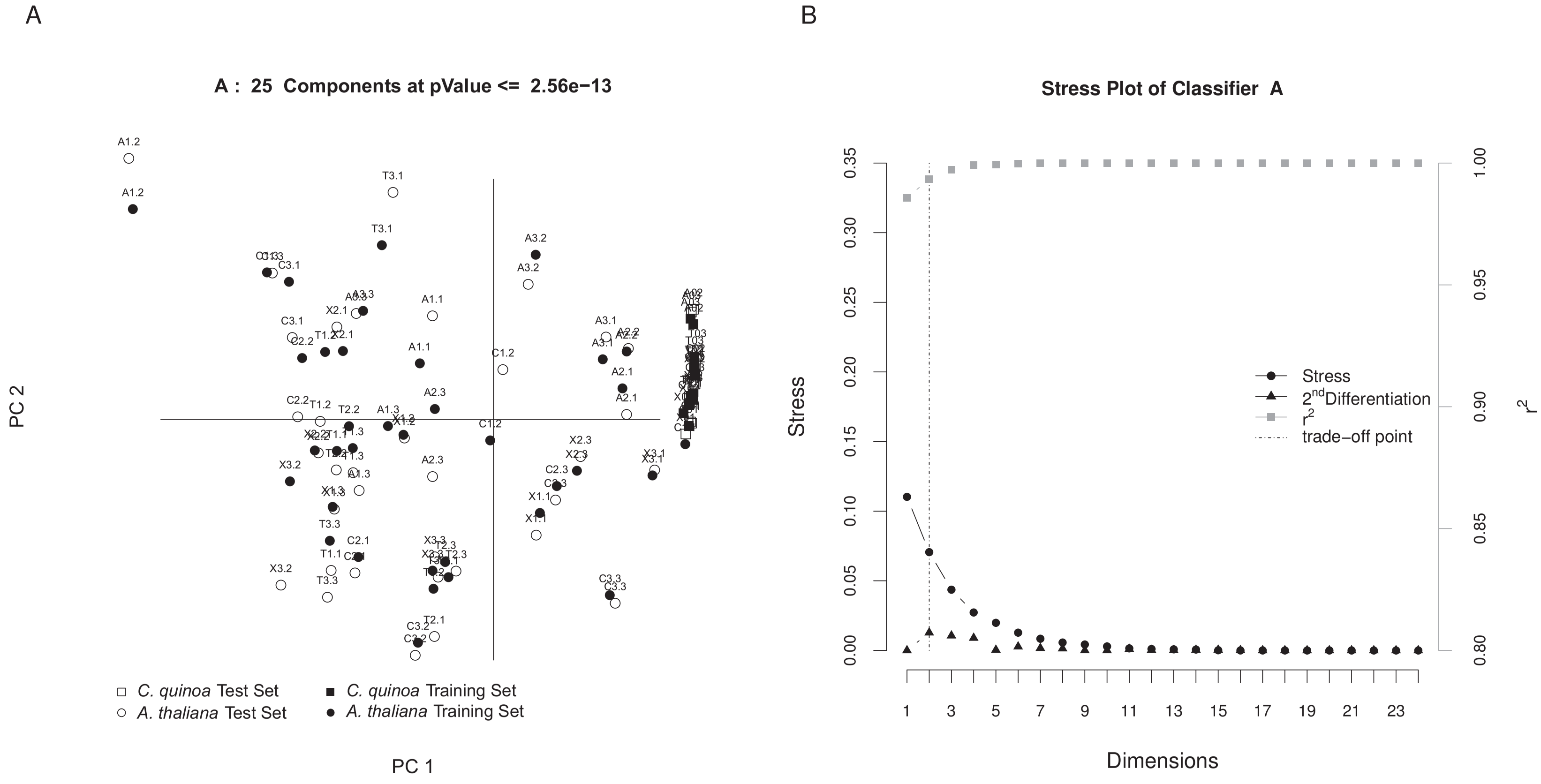


Figure 27: The species classifier (A) based on the `srProfiler`-approach is shown. A shows the PCA of *Arabidopsis thaliana* (circles) vs. *Chenopodium quinoa* (squares). 25 components were used to calculate the map, all of which had a pValue smaller or equal to $2.56 \cdot 10^{-13}$. The sample clouds of the respective species show a distinctly different behavior. The *Arabidopsis thaliana* samples are spread widely through all quadrants. The *Chenopodium quinoa* samples are tightly grouped at the far right edge of the graph. Sample C1.1, in test and training, overlaps with the *Chenopodium quinoa* samples. All other samples are positioned in the area of their respective group. The STRESS plot shown in B shows an ideal trade off point (dot-dashed line) at two dimensions with the STRESS being below 0.15 and the r^2 value being 0.985.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

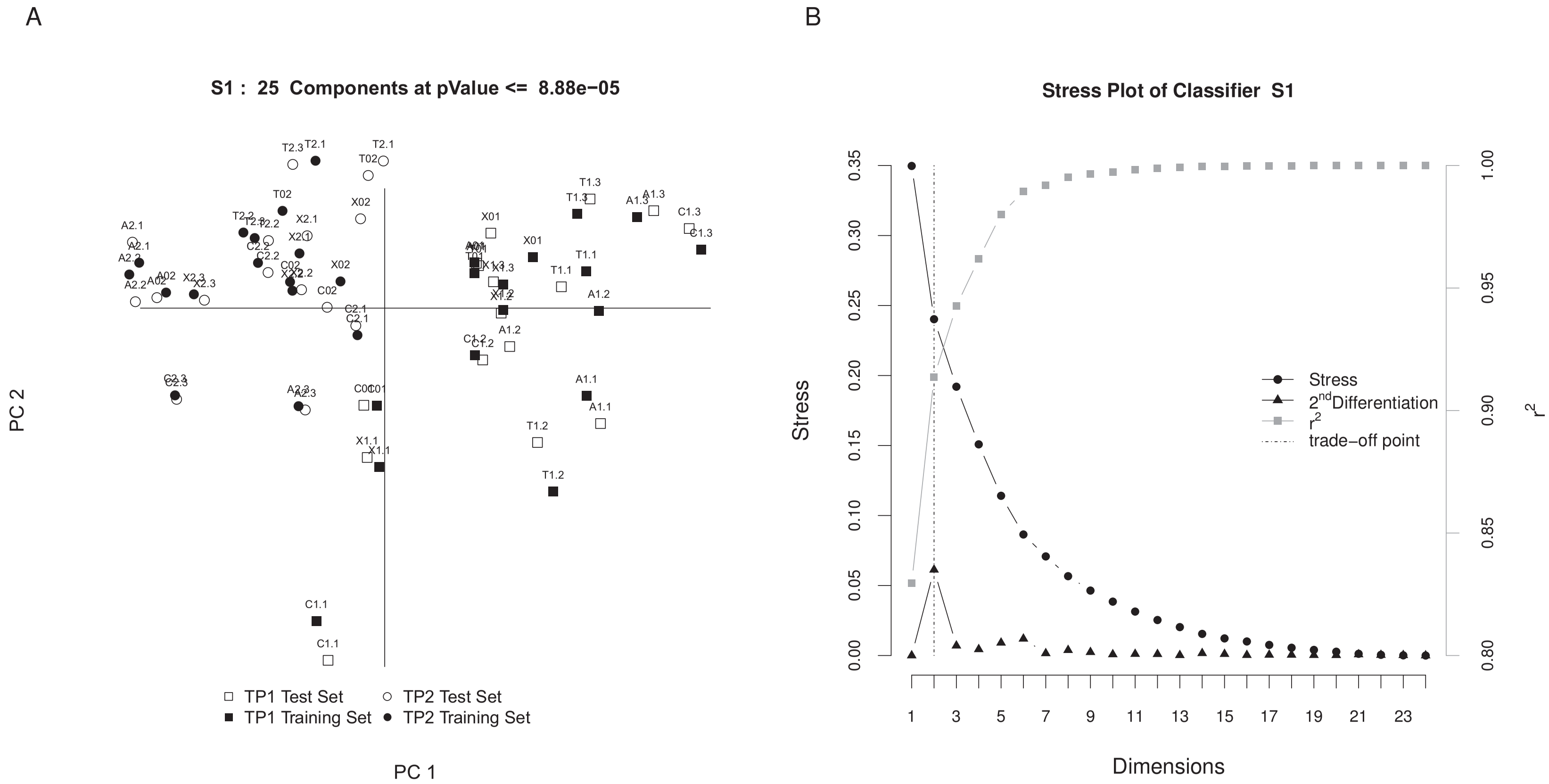


Figure 28: The age classifier (S1), differentiating between TPs 1 (squares) and 2 (circles), based on the `srProfiler`-approach is shown. The PCA in A shows the majority of TP2 samples in quadrant II. Most of the TP1 samples are spread through quadrants I and IV. 25 components with a pValue smaller than or equal to $8.88 \cdot 10^{-05}$ have been used to calculate the PCA. The STRESS at two dimensions is below 0.25 and the corresponding r^2 value is 0.915 as can be seen in B.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

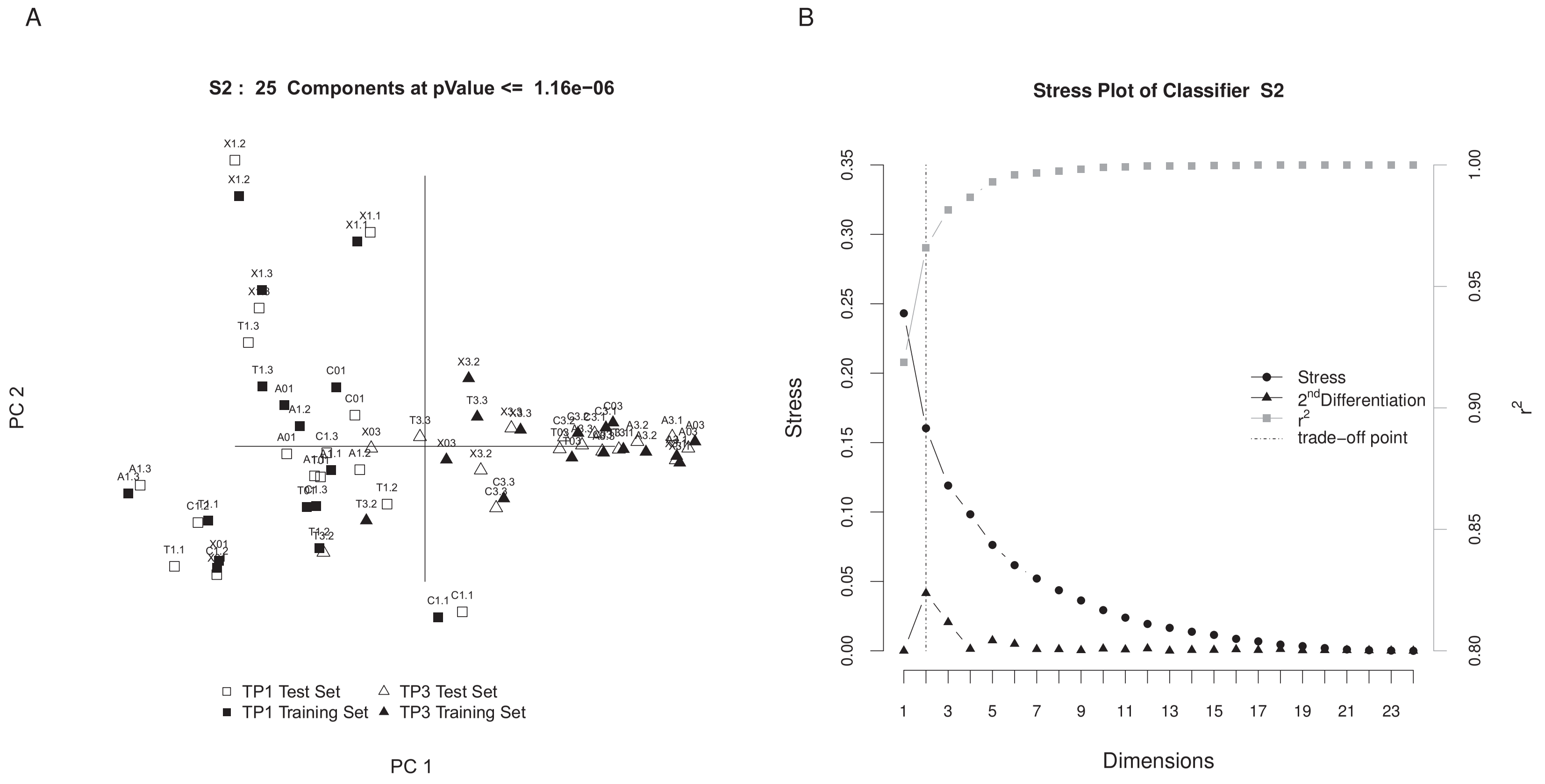


Figure 29: The age classifier (S2), differentiating between TPs 1 (squares) and 3 (triangles), based on the `srProfiler`-approach is shown. The PCA in A shows the TP1 points vertically spread through quadrants II and III. The TP3 samples are spread horizontally along the PC1 axis. A minor overlap involving the samples X03 (test) and T3.2 (training and test) can be seen in quadrant III. The test set samples (hollow shapes) are located in relative proximity to the respective training set counterpart. The map was calculated using 25 components with a pValue smaller than or equal to $1.16 \cdot 10^{-06}$. B shows an trade off at two dimensions with STRESS of 0.16 and a r^2 value of 0.96.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

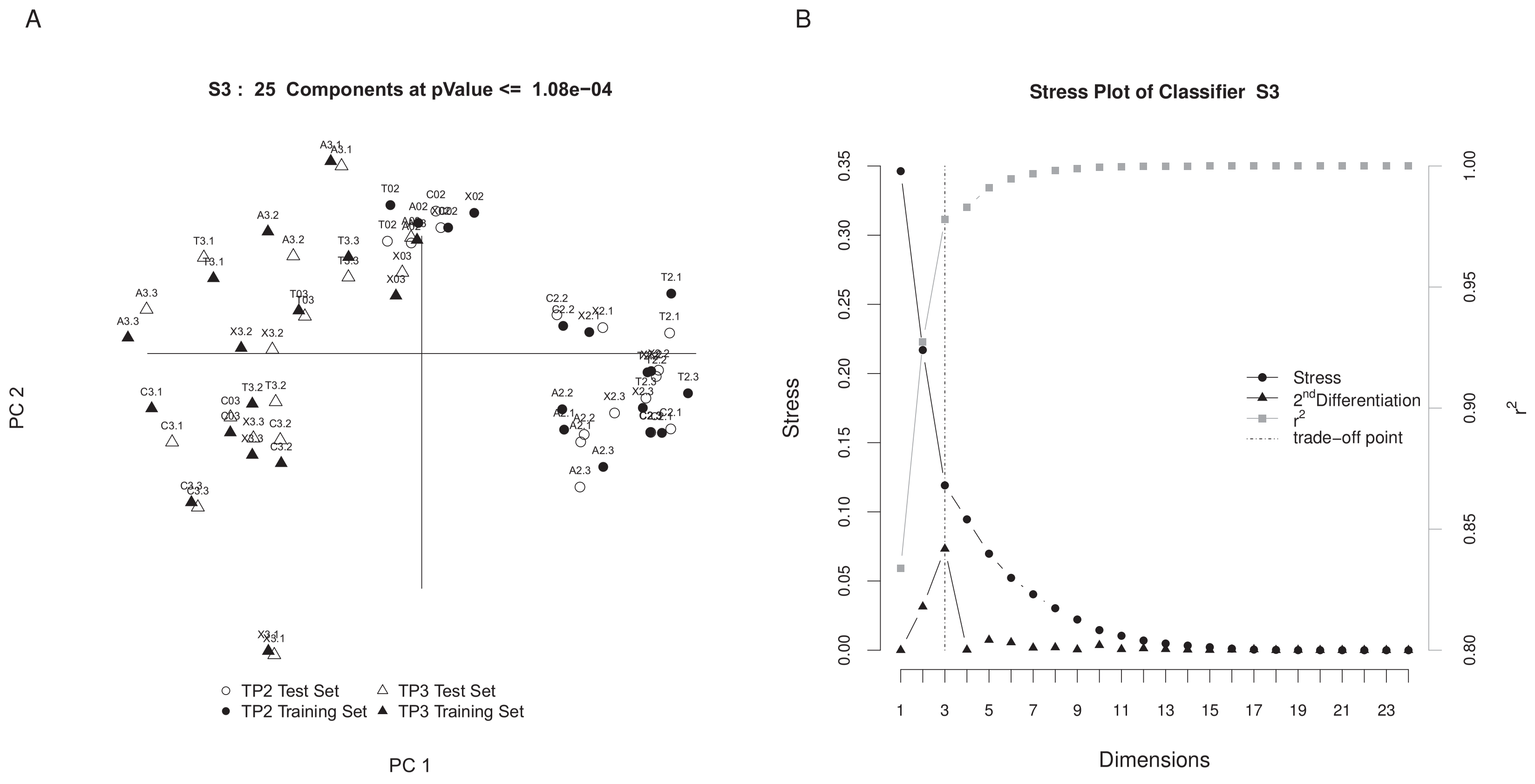


Figure 30: The age classifier (S3), differentiating between TPs 2 (circles) and 3 (triangles), based on the `srProfiler`-approach is shown. The PCA in A shows the TP3 points spread widely over quadrants II and III. The majority of TP2 samples are located center right of the graph. A subgroup of TP2 samples involving all *Chenopodium quinoa* samples is located center top of the graph, overlapping a TP3 sample (A03). The test and training samples are in relative proximity to one another. The pValue of the utilized 25 components is smaller than or equal to $1.08 \cdot 10^{-04}$. The STRESS plot in B shows an ideal trade off at three dimensions while showing a STRESS of 0.225 and a r^2 value of 0.92 at two dimensions.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

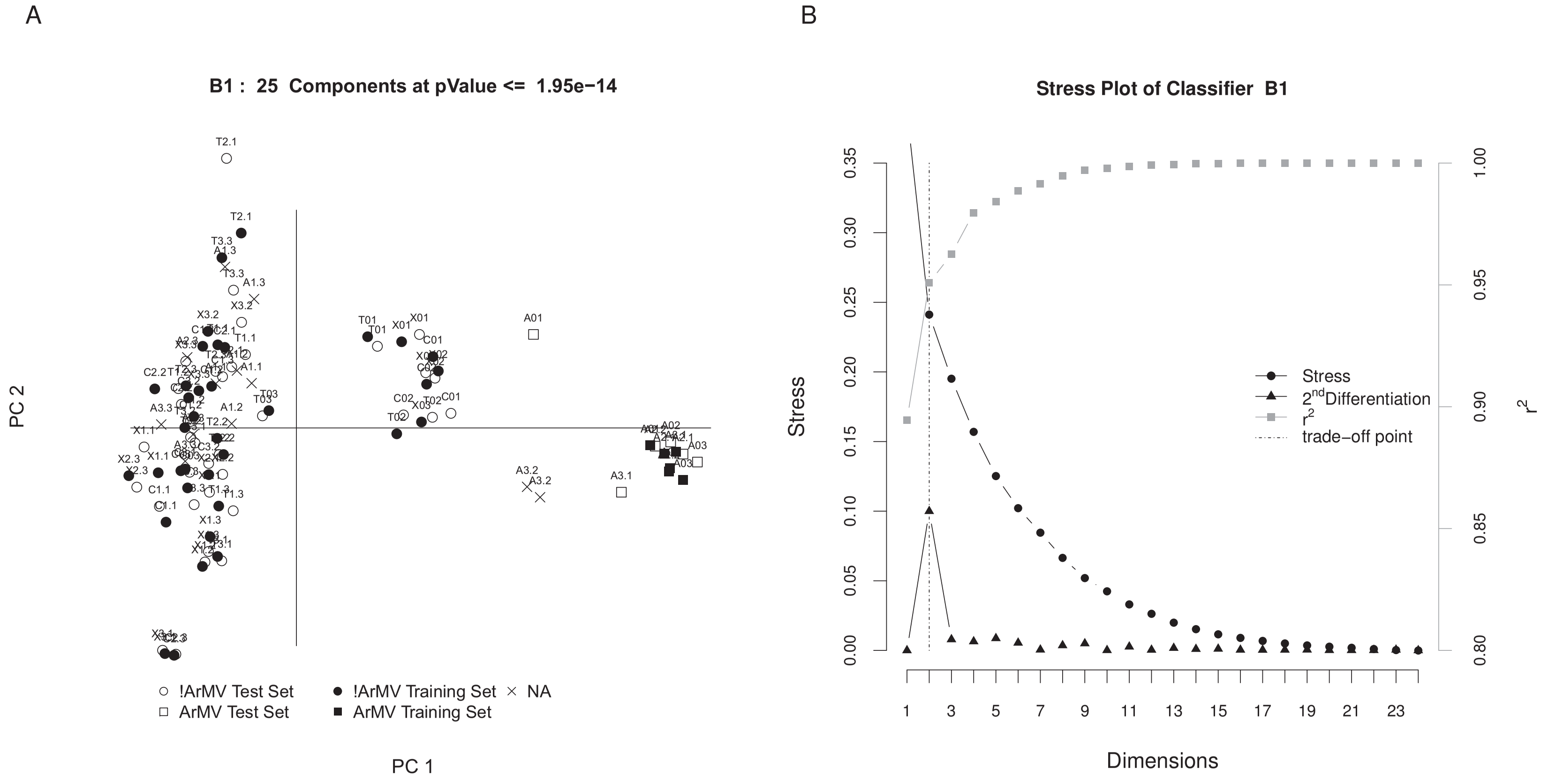


Figure 31: B1, the classifier differentiating between ArMV infected samples (squares) and not ArMV infected samples (circles) (all other samples) is shown. The map in A was constructed using 25 components with a pValue smaller than or equal to $1.95 \cdot 10^{-14}$ and shows a wide spread point cloud of not ArMV infected samples through the quadrants II and III. The ArMV infected samples are, apart from one test point (hollow squares), positioned in quadrant IV. Samples not used for the construction of the classifier are represented as crosses. The test samples can be found close to the respective training points. A01 shows a very large distance between test (quadrant I) and training sample (quadrant IV). A subset of the not ArMV infected samples builds a distinct group made up of the majority of *Chenopodium quinoa* samples. Sample A3.2, which was not used for the construction of this classifier, is located in close proximity to the ArMV infected sample group. The two dimensional representation is the ideal trade off according to the STRESS plot in B. The STRESS is less than 0.25 and the r^2 value is 0.95.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

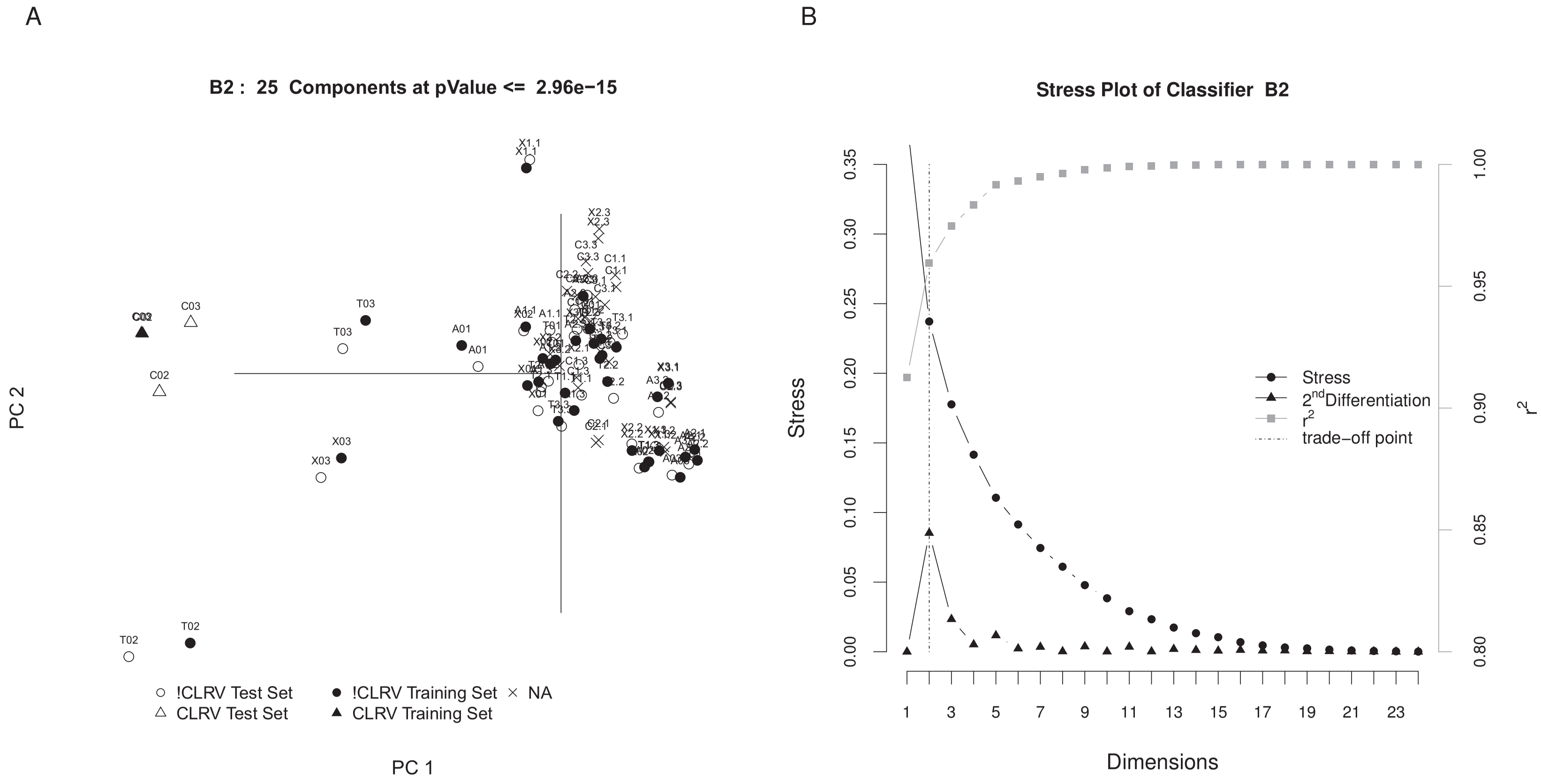


Figure 32: B2, the classifier differentiating between CLRV infected samples (triangles) and not CLRV infected samples (circles) (all other samples) is shown. The map in A was constructed using 25 components with a pValue smaller than or equal to $2.96 \cdot 10^{-15}$ and shows the not CLRV infected samples spread mainly over quadrants I and IV. Samples X1.1, T03, X03 and T02 (test and training) follow a narrow, wide spread band distinctly different from the other samples in this group. The CLRV infected samples are positioned at center left of the graph with a large distance to all other points. Points represented as crosses were not involved in the construction of this classifier. The test samples (hollow shapes) can be found close to the respective training points (solid shapes). The two dimensional representation is the ideal trade off according to the STRESS plot in B. The STRESS is less than 0.25 and the r^2 value is 0.955.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

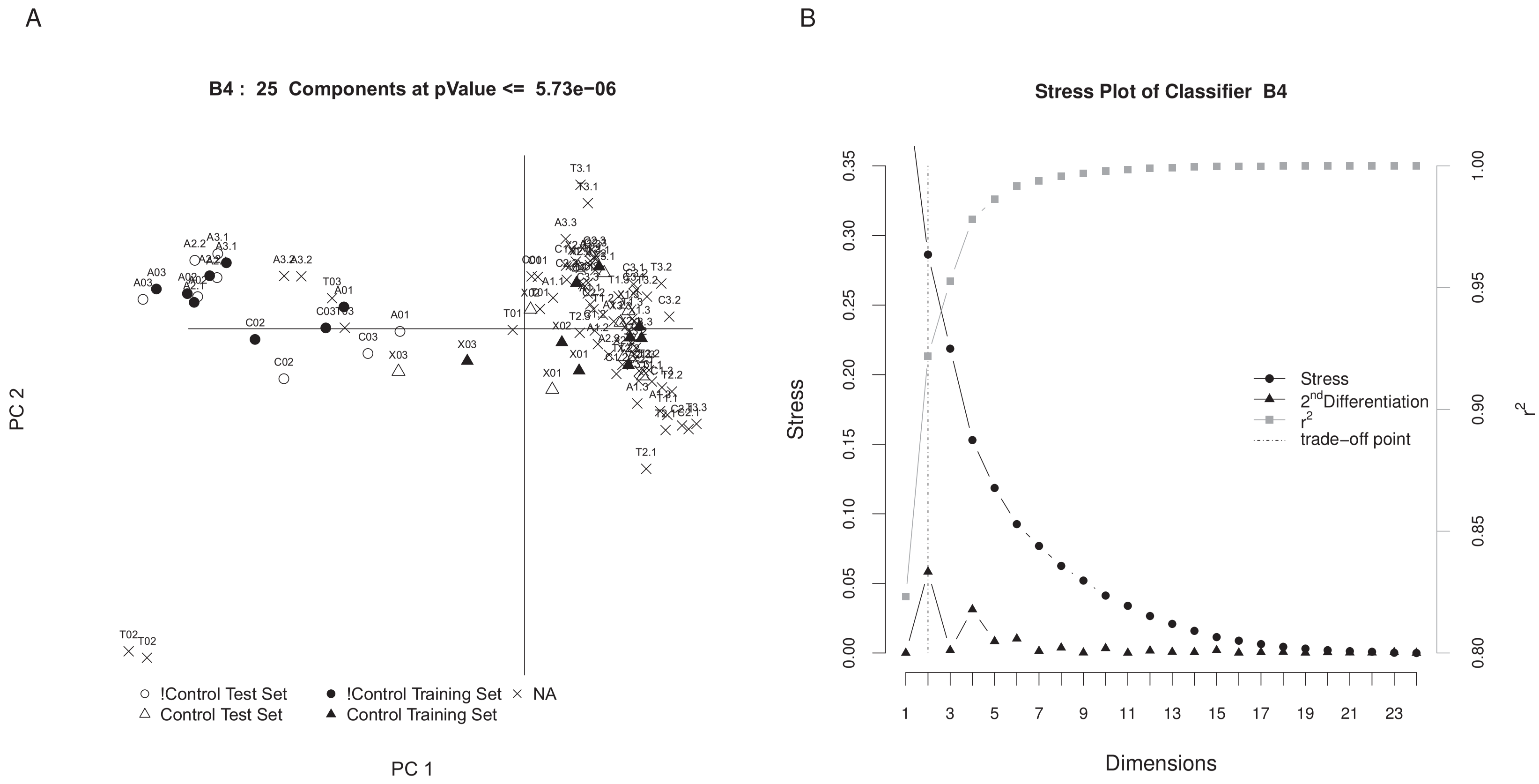


Figure 33: B4, the classifier differentiating between infected samples (circles) and not infected samples (triangles) is shown. The map in A was constructed using 25 components with a pValue smaller than or equal to $5.73 \cdot 10^{-06}$. The infected samples are positioned center left and the not infected samples are mainly spread over quadrants I and IV. Sample X03 (test and training) shows a larger distance to the other samples in its group. The infected samples are located in two groupings. A tightly grouped set to the far left of the plot and a wider spread set, made up of C02, C03 and A01 (test and training), further to the right. Points represented as crosses were not involved in the construction of this classifier. The test samples (hollow shapes) can be found close to the respective training points (solid shapes). The two dimensional representation is the ideal trade off according to the STRESS plot in B. The STRESS is less than 0.3 and the r^2 value is 0.925.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

3.4 Validation by means of Pathogen Alignment

As described in section 2.8 the sequencing results were aligned against the reference genomes of the three used viruses to measure the viral load in each sample at the time of harvesting. Figure 34 shows the results of this alignment by depicting how much of the sequenced reads belong to the respective virus (see table 10 for per sample values). The changes over time in *Arabidopsis thaliana* and *Chenopodium quinoa* are similar considering that the *Arabidopsis thaliana* samples were grown over a much longer period. For ArMV the viral load increases and reaches a plateau after a time, in the *Arabidopsis thaliana* graph a slight reduction can be seen. The CLRV concentration in *Chenopodium quinoa* rises over the first time period and decreases after that. In *Arabidopsis thaliana* the concentration first declines slightly and then drops more significantly over the second time period. The TSWV concentration could not be measured in *Arabidopsis thaliana* and in *Chenopodium quinoa* the concentration drops significantly during the first period and cannot be measured at all after the second period. The results suggest that the inoculation with ArMV was successful in *Arabidopsis thaliana* and *Chenopodium quinoa*. However the declining measurements of CLRV in *Arabidopsis thaliana*, which are orders of magnitude smaller than the load measured for ArMV suggests that the inoculation of CLRV was successful in *Chenopodium quinoa* only. Furthermore the TSWV inoculation appears to have been unsuccessful in both species, since no load could be measured in *Arabidopsis thaliana* and the load in *Chenopodium quinoa* dropped significantly from measurement one towards measurement two.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

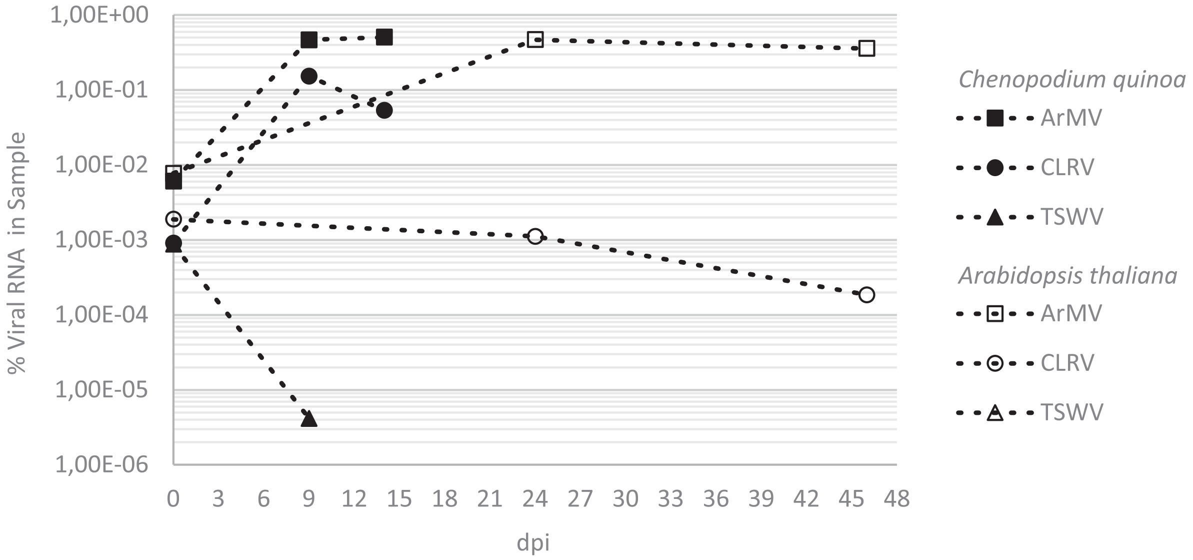


Figure 34: The viral load, as measured by means of sequence alignment against the individual viral reference genomes of ArMV, CLRV and TSWV in %RNA of *Chenopodium quinoa* and *Arabidopsis thaliana* samples over time is shown. The measurements of the *Chenopodium quinoa* samples were taken 0, 9 and 14 dpi. The *Arabidopsis thaliana* measurements were taken 0, 24 and 46 dpi. TSWV could be measured in *Chenopodium quinoa* samples only at time points 1 and 2, in *Arabidopsis thaliana* samples TSWV was not detectable.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

4 Discussion

A major issue of transcriptome analysis is the requirement of highly specialized personnel and a lot of time, especially if a reference genome to the sample in question is not provided (Baker, 2012). The introduced `srProfiler` approach addresses this issue and shows that, in addition to being faster and easily handled, the results produced are similar and even better in direct comparison with the conventional methods (section 3.3).

Biological Assay The proliferation and movement of a virus introduced to a plant strongly depend on the initial titer. The more viral particles were successfully introduced into a plant, the faster the virus reaches a high titer and the faster it spreads through the plant. Localized defensive systems like the hypersensitve response of *Chenopodium quinoa* reduce the titer and the subsequent spread of a virus by selective apoptosis of infected cells (Pontier *et al.*, 1998). Therefore the distribution of a given virus within a sample plant is uneven. While in one leaf the titer is very high and can be measured clearly, a different leaf of the same plant can show no measurable titer (Hull, 2013). Similarly a defensive response of the host can be highly localized. The tissue amount required for NGS is comparatively small, a 0.5 cm^2 piece is enough to run the sequencing (section 2.5). Considering the uneven distribution of the virus and the response markers, as well as the small tissue amount used for analysis, rather than using random leaves as input - as was done here - or even using only leaves showing strong symptoms, using multiple parts of the sample plant, homogenizing those parts and using an aliquot of the resulting homogenate would significantly reduce the possibility of under or over representation of viral particles or response markers during the analyses.

The results of Rumbou *et al.* (2009) suggest that CLRV isolate E395 has a comparatively low infection rate in *Arabidopsis thaliana* compared to other isolates. The

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

results of this work agree with those findings, since neither of the nine *Arabidopsis thaliana* samples inoculated with CLRV could be identified as infected, while two of the three inoculated *Chenopodium quinoa* samples could be clearly classified as such (section 3.1). A different possible reason for the unsuccessful inoculation is the use of different source materials for the inoculum, due to the different dates the inoculation was performed on *Chenopodium quinoa* and *Arabidopsis thaliana* respectively (section 2.3). The source plant was the same, however different leaves were used, possibly being subject to the same uneven dissemination of viral particles as described earlier. A different titer in the inocula could account for the varying success of the inoculations. The apparent CLRV infection of an *Arabidopsis thaliana* control sample (X1.2), as shown by the ELISA (section 3.1), would dispute this theory, however this apparent infection could not be corroborated by the pathogen alignment (section 3.4). Furthermore the E_{405} value of this sample is considerably lower compared to other positive samples. The biological replicates of X1.2 show no signs of infection. Considering those contradictions a false positive measurement is probable.

The ELISA (section 3.1) as well as the pathogen alignment (section 3.4) show that the inoculation with TSWV did not succeed in any of the samples. While the propagation worked well, as shown by the TAS ELISA performed on the proliferation samples (figure 18), subsequent inoculation using material from the propagation sample showing the greatest titer failed entirely. It can be seen from said ELISA that although the proliferation of the virus was successful, it was so in *Nicotiana spp.* samples only. The TSWV isolate was used on multiple propagation samples including *Chenopodium spp.* samples, none of which could be infected successfully. Additionally, the origin host of the isolate was a *Nicotiana rustica* (Menzel, 2016). This could be an indicator that this isolate (PC-0182(L3)) is of a particular strain which has a very low infection rate in samples other than *Nicotiana spp.*.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

It is also possible that, even though the material for the inoculum showed strong signs of infection, the viral titer was too low for further infection. Since, however, two distinct inoculi were used (one for *Chenopodium quinoa* samples and later another for *Arabidopsis thaliana* samples) (section 2.3), the probability of choosing a low-titer-leaf twice, given a sample so clearly infected, is low.

With three RNAs TSWV is more complicated to spread using manual mechanical inoculation, since all three RNAs have to be introduced into the cell. Due in part to this problem Mandal *et al.* (2008) proposed a new method of applying the inoculum using a high pressured (4.1 bar) spray. However, while the probability of successful infection of a sample using mechanical inoculation is lower compared to ArMV or CLRV, it is not impossible as could be seen during propagation. Therefore it is unlikely, even given this circumstance, that none of the 12 samples inoculated would be infected.

Genetic Analyses All transcriptome based classifications were performed without a reference for *Chenopodium quinoa*, which would have allowed for a better feature selection. This decreased the accuracy of the transcriptome based classifiers, since only genes being expressed in *Chenopodium quinoa* and *Arabidopsis thaliana* could be used. *Chenopodium quinoa* has no reference as of this time. In order to overcome this obstacle a reference could have been assembled. A time consuming process, which requires a deep sequencing coverage (Baker, 2012; Sims *et al.*, 2014), which was not provided here. Furthermore, building a reference every time a new plant is analyzed, updating local references and exchanging those references with all collaborating labs is a strong reason not to use NGS in a routine environment (Boonham *et al.*, 2014), especially considering reproducibility using unstable, temporary references. In order to reproduce realistic circumstances during the analyses a reference was used which fit very well to one species and only partly to the second species (Collins *et al.*, 2008; Ward *et al.*, 2012).

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

The species classification performs very well using the expression profiles (section 3.3). The `srProfiler` profiles also perform well using components with a much lower maximal pValue. The two dimensional map using a reference and expression values shows a better, more condensed, clustering of the two groups (figure 20), while the profiles without any reference background produce minimally overlapping clusters with a substantially different behavior (figure 27). While the clustering using the expression profiles appears to be better, the underlying methodology produces a markable bias in this case. Since the reference for *Arabidopsis thaliana* was used to discover differentiating genes in both species and the expression of any given gene had to be greater than zero, the method can only utilize genes which are homologous in both species. The impact of this constraint can be seen in the higher STRESS and r^2 values in the expression map compared to the `srProfiler` map. A further result of this bias can be seen in the high spread of the *Arabidopsis thaliana* samples compared to the *Chenopodium quinoa* samples in the `srProfiler` species map. While the input to the expression map is based solely on homologous genes which were expressed by both species, thus resulting in the representation of few large entities of the host expression profile of all samples being covered (section 2.6), the input for the `srProfiler` map is based on many small entities being covered (section 2.9). If a certain gene is diversely expressed in one species, its impact is comparatively minor since only one value is representing said gene. On the other hand, the same gene using the `srProfiler` approach is covered many times, resulting in multiple values representing this diversity. Therefore the *Arabidopsis thaliana* samples, having three times the input amount compared to the *Chenopodium quinoa* samples, are represented with a much larger diversity. This over representation of one species results in the bias in the `srProfiler` map. With increasing sample numbers, resulting in an equal representation of all species, the map will show a similar spread area for the represented species (Whitley *et al.*, 2002). This bias is only present in the species classification, since in all other classifications the species properties are immaterial and are thus removed during the dimensionality reduction (section 2.10).

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

The age related classifications show mostly well defined clusters throughout all maps. The `srProfiler` maps, however, are clearer. The overlaps are fewer in number and the pValues are smaller. In the S3 classification, dividing samples harvested at time points two and three respectively, the time point two samples are spread in two distinct clusters (figure 30). The clusters are made up of only *Chenopodium quinoa* samples and only *Arabidopsis thaliana* samples respectively. This behavior indicates a species related difference not present in time points one and three. Since, however, the species related properties have been removed for this classification, the clustering must be caused by an age related property. The *Chenopodium quinoa* samples were harvested 9 dpi (TP2) and 14 dpi (TP3) (section 2.3), having a temporal distance of five days or 36%. The *Arabidopsis thaliana* samples were harvested 24 dpi (TP2) and 46 dpi (TP3), having a temporal distance of 22 days or 47%. The subgrouping of *Chenopodium quinoa* samples is closer to the time point two samples, which can be caused by the smaller temporal distance of 36% between samples harvested at time point two and time point three compared to 47% temporal distance in *Arabidopsis thaliana* samples. Figure 35 illustrates the difference of the two time spans.

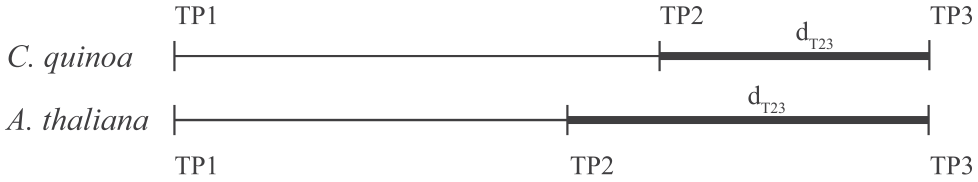


Figure 35: The dates of harvesting, referenced by TP1, TP2 and TP3, are shown relative to the absolute duration of the experiments. The distance between TP2 and TP3, referenced as d_{T23}, is highlighted.

In the B classifications, differentiating between samples infected with a given pathogen and all other samples, an additional grouping of samples was introduced. During the analysis of the samples conflicting results were provided by ELISA (section 3.1) and pathogen alignment (section 3.4). While, for instance, sample X1.2 was diagnosed as infected by CLRV using ELISA, no significant amount of reads aligned to the

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

pathogen reference. To ensure that the classifications were calculated without a possible bias, only samples which could be diagnosed by both validating methodologies as being in agreement with their respective treatments, were used. Samples not being in agreement with the underlying treatment, or producing conflicting results during validation were disregarded from the training set and placed in a second “test” set, the actual state of which is unknown.

The B classifications are significantly better using the `srProfiler` profiles. The classification between ArMV infected and other samples (figure 31) shows a tight cluster of infected samples which is positioned a great distance from the other samples. In the expression map (figure 24) two overlapping clouds can be seen overlapping mainly in samples of TP1 and TP2, with a distance increasing apparently depending on host-virus interaction time. However, the test set A03 seems to dispute this theory. Since this point behaves remarkably different from the other test set samples, which are positioned relatively close to their training set counterparts, it is probable that A03 is an outlier in this scope. The independent sample A3.2 is very close to the not infected samples. While it is classified correctly the distance to the not infected samples should be considerably greater. Considering the amount of reads aligning to the pathogen (table 10) a relatively small distance to A3.1 is expected. The `srProfiler` map shows a different behavior. Sample A3.2 is clearly positioned in relative proximity to the other infected samples and closest to sample A3.1.

The classification of samples infected with CLRV and those that are not infected with CLRV did work poorly using expression profiles (figure 25). A differentiation is not recognizable. Using the `srProfiler` profiles on the other hand (figure 32), the map shows a great distance between infected samples and all others. Considering however, that only two samples could be used in the CLRV infected group to construct the classifiers, both resulting maps must be regarded as highly circumstantial. The small amount of information usable, could have resulted in a classifier which differentiates *Chenopodium quinoa* samples being older than a few days and any

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

other sample. This is in part corroborated by the relative small distance to other *Chenopodium quinoa* samples of TP2 and TP3 (T03, T02, X03).
The classification of infected versus uninfected samples worked well in both scenarios (figure 26, 33). The maps show a clear split of uninfected samples to the right and the infected samples to the left. Moreover, both maps show that samples with prolonged host-virus interaction have a larger distance to the uninfected samples. In these maps a clear difference in response quality can be seen regarding the independent sample. In the expression map A3.2 is very close to the uninfected group, which is in agreement with the expression ArMV classification. However, considering the results of ELISA and pathogen alignment this sample should be clearly in the infected group. In this case too the `srProfiler` map is considerably closer to the expectation. The sample is positioned well within the infected group and relatively close to sample A3.1. So, while the individual maps are in agreement with the respective independently calculated ArMV classifications, the `srProfiler` maps show a much clearer result, which behaves analogously to the expectation.

In all classifications which were unrelated to species (S1, S2, S3, B1, B2 and B4) the `srProfiler` outperforms the expression profiles. This can be seen in table 8, which lists the sensitivity, specificity and accuracy for all classifiers. Furthermore, the clustering is cleaner and the overlaps are less. The pValues of the used entities are significantly smaller using the `srProfiler` profiles. Additionally the STRESS plots show that the two dimensional representation is more strained for the expression profiles. A higher dimensional space is needed for the expression based maps to achieve similar scores to the `srProfiler` based maps. The better performance of the `srProfiler` approach can be credited to the much higher amount of usable entities. The pool of genes which can be used to perform the classification, genes having an expression greater than zero in all training samples, is much smaller than the pool of bins which could be used, bins showing a count of greater than zero in all training samples. The differentiation using genes is less significant than the

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

`srProfiler` method. Moreover, while the expression values are very host specific, due to the specificity of the reference, the `srProfiler` can be regarded as a generic system, much like NGS itself (Selvarajan *et al.*, 2016). While the `srProfiler` uses all sequenced reads, including NMR and MMR, the transcriptome approach uses only UMR. This ensured a consistently higher amount of usable data in every classification. Figure 36 shows the distribution of entities covered by any given number of samples. It can be seen that only very few transcripts are covered by all samples, whereas the vast majority of bins is covered by all samples.

Table 8: The table lists the sensitivity, specificity and accuracy of each classifier using either the transcriptome approach (TA) or the `srProfiler` approach (n-Gram).

Classifier	Classifier ID	Sensitivity		Specificity		Accuracy	
		TA	n-Gram	TA	n-Gram	TA	n-Gram
Species	A	1,000	1,000	1,000	0,972	1,000	0,981
Age (TP1 vs TP2)	S1	0,688	1,000	0,875	1,000	0,781	1,000
Age (TP1 vs TP3)	S2	0,875	0,938	0,938	0,875	0,906	0,906
Age (TP2 vs TP3)	S3	0,813	1,000	0,938	0,875	0,875	0,938
Infectant (ARMV)	B1	0,500	1,000	1,000	1,000	0,929	1,000
Infectant (CLRV)	B2	0,000	1,000	1,000	1,000	0,943	1,000
Infectant (Control)	B4	1,000	1,000	0,875	1,000	0,938	1,000

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

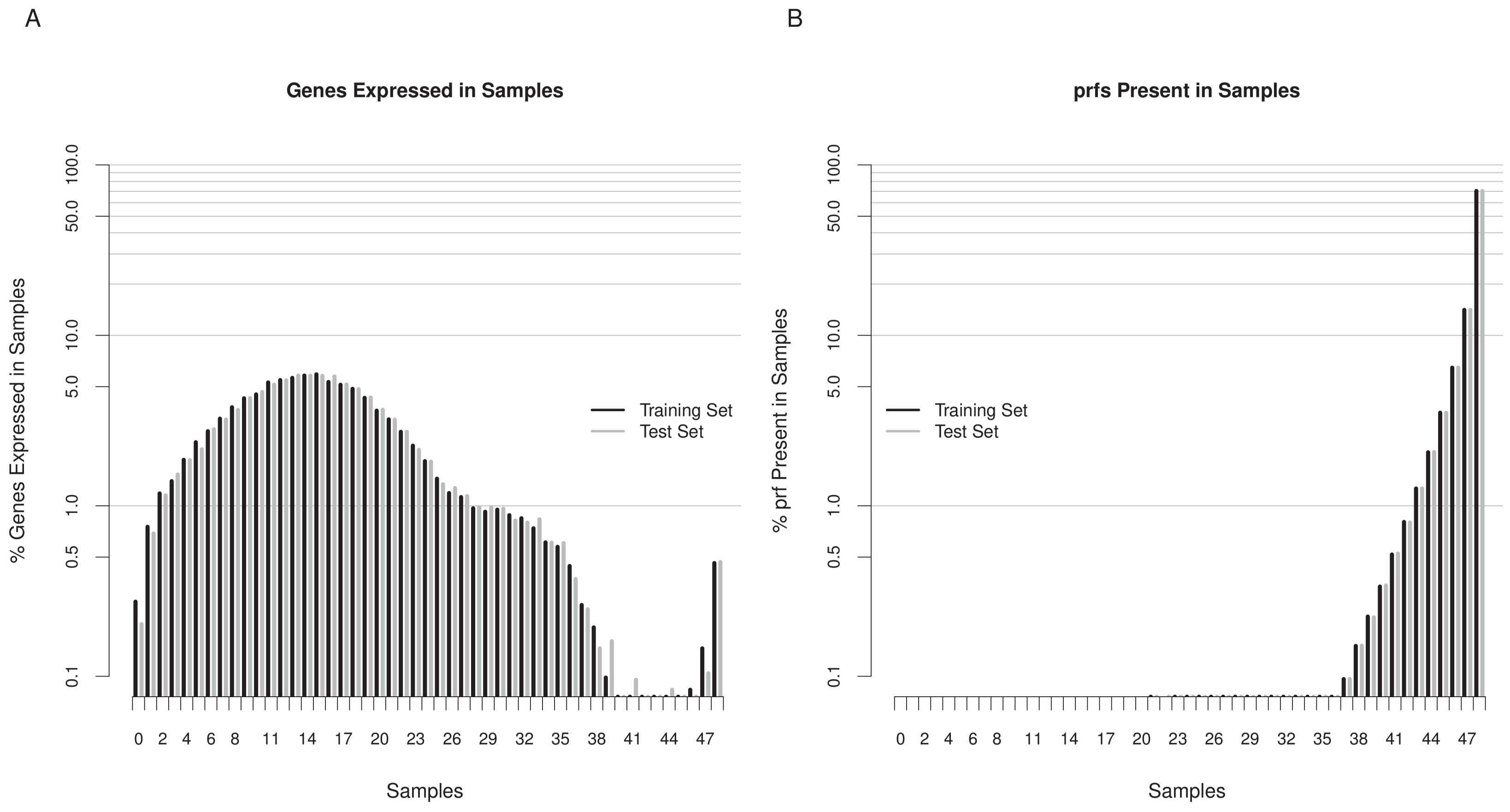


Figure 36: The amount of entities covered only by a given number of samples in either training (black) or test (gray) set is shown. A illustrates the percentage of genes covered by an increasing number of samples. B depicts the percentage of bins covered by an increasing number of samples. While only 0.5% of genes are covered by all samples, over 70% of bins are covered by all samples.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

Gold Standard Comparison The information, which can be gained from a transcriptome analysis is many times greater than the information acquired from either ELISA or qPCR, however the resources, in form of time, personnel and money, needed to properly analyze the numerous reads are vastly higher compared to the gold standard methods (Boonham *et al.*, 2014).

In the field of routine diagnostics ELISA has been established as the gold standard. It is cheap, comparatively easy to use, robust, scalable, highly automizable and provides the user with a well interpretable result. However, the test has inherent shortcomings. It answers only one query at a time, it produces false negatives if the antigen is slightly changed (Capobianchi *et al.*, 2013; Adams *et al.*, 2013) and the design and establishment of new antibodies is expensive both financially and temporally (Boonham *et al.*, 2014; Büttner *et al.*, 2013). On many routine diagnostics those shortcomings have very little impact. In these areas ELISA will continue to be the method of choice (Büttner *et al.*, 2013). However, in some areas or for some queries the impact of those shortcomings is substantial. For instance, a symptomatic sample is analyzed, yet the underlying pathogen is not obvious (Büttner *et al.*, 2013). Multiple ELISA tests must be run to find the pathogen, presuming that the analyzing lab has the corresponding antibodies. The price per sample rises significantly if the pathogens cannot be limited to only a few candidates. If the pathogen is new or a new quasi species of a known strain, a different method must be used or a new antibody must be designed. The sample could be analyzed using qPCR or NGS. At this point multiple tests were run without any usable information being produced. Those cases are well suited for NGS analyses because after the investment the information provided will most likely include the underlying pathogen (Boonham *et al.*, 2014). The problem arising now is carving the information from the plethora of data without a clear idea what the underlying pathogen might be. An assembly must be performed and the resulting contigs must be compared to multiple pathogens to discover the perpetrator (Baker, 2012). This method assumes that the pathogen

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

has been sequenced along with the host and was not removed prior. Otherwise a search for markers within the data begins hinting to the pathogen. Instead of multiple ELISA tests followed by NGS, time and money can be saved starting with NGS. After the sample has been sequenced, the data can be screened for multiple properties before starting the time consuming, in depth analyses. This screening, performed by the `srProfiler` tool, can severely reduce the number of possibilities a subsequent analysis has to consider. In cases of quasi species, the tool being less specific than ELISA (Capobianchi *et al.*, 2013; Adams *et al.*, 2013), could provide the answer right away, further providing the bases to assemble the new pathogen (Prabha *et al.*, 2013).

In the near future the price for an NGS run will reach the level of the gold standard tests (Wetterstrand, 2013), at which point it would be preferable to run NGS because the resulting data can be used for many independent analyses afterwards and the run itself will be faster than the ELISA test. If, of course, the subsequent analysis requires a lot more time by comparison, the advantages of the superior technology will pale (Boonham *et al.*, 2014). The `srProfiler` method circumvents this problem. Using this tool the sequenced sample can be run against every pathogen of interest in the same time an ELISA test is run on a single pathogen. Moreover, the tool can provide information about more general properties such as the species of the pathogen (i.e. RNA virus, fungi, bacteria) or the age of the sample, reducing the complexity of the following analyses.

A strong automation is a helpful aspect in a routine environment (Boonham *et al.*, 2014). ELISA and NGS provide such an automation for all wet-lab steps. The subsequent analyses strongly depend on the queries. While the analyses and interpretation of ELISA plates are standardized and straight forward (the plate is given into a reader which is connected to a computer running a software that can interpret the extinction information and provide the user with a clear answer), the analyses and corresponding interpretations of NGS runs are less so. In this work the tran-

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

scriptome analysis was used, following a more or less strict work flow (Baker, 2012). A different possibility would be an exome analysis, following its own work flow. This multitude of analyses which can be based on any NGS run, are the reasons why substantial bioinformatic expertise is needed throughout the process (Boonham *et al.*, 2014; Baker, 2012).

The `srProfiler` method, in contrast, can be fully automatized, allowing for a standardized routine workflow. Moreover, because it is classification based, the tool can work with all data NGS produces as long as the classifiers are designed accordingly. The results produced are of an easy to interpret nature. The output is a list of probabilities, stating the likelihood of the analyzed sample belonging to any provided class. A cutoff calculation as for ELISA (BIOREBA AG, 2014) is not necessary.

Data-reusability is a minor aspect in diagnostic analyses. The data produced by ELISA is discarded once the results are produced because all information within the extinction data is depleted. NGS presents a very different form of data. The information within the data is massive and can be accessed even years after the initial run (Capobianchi *et al.*, 2013). This is a very positive property. A sample analyzed with ELISA cannot be analyzed by any other procedure again, no further information can be gained. A negative side effect is the increasing need for storage to save the NGS reads for future analyses. While the `srProfiler` profiles can not entirely solve this need, they are substantially smaller than either basecall or `fastQ` files (section 2.9). Like the original files the profiles do retain the information for future analyses and can be run against any kind of classifiers designed in the future.

A major disadvantage of NGS analyses are the requirements regarding computational infrastructure. Storage is not the only hardware needed. In order to run any analyses in an acceptable time frame multiple strong CPUs and a lot of memory is required. Millions of reads being aligned to multiple very long fragments preferably stored in memory is a computationally expensive process (Langmead *et al.*, 2009; 2012). ELISA is the opposite in this regard. Apart from the plate reader and a con-

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

ventional computer, it requires no substantial infrastructure. Even if the data would be stored, the original files are very small. The `srProfiler` again presents itself as a middle path. As described earlier it requires storage and the analyses are more expensive than ELISA analyses. Those requirements are however significantly less compared to the standard NGS methods. While all bins of a profile are stored for further use, only very few are used for classification (25 in this work). So instead of handling millions of reads, a computer running a `srProfiler` analysis handles only 25 bins at a time, comparable in bit size to a read comprised of four bases. The tool could effectively be used on a standard laptop in the field.

The greatest advantage the `srProfiler` method has compared to ELISA is that it can run many classifications simultaneously in a very short time. Even if after an analysis new information should come to light, demanding further or different analyses, the tool can simply be run with the corresponding classifiers. While for every query a new ELISA test has to be run, a new NGS run or new preprocessing of the data is not required.

The usability of the `srProfiler` based classifications is dependent on the existence of classifiers, much like ELISA is dependent on the existence of specific antibodies (Büttner *et al.*, 2013) or qPCR is dependent on the existence of specific reference primers. While the design and production of antibodies is a time consuming and expensive process which needs to be run independently of the analyses that require them (Boonham *et al.*, 2014), the construction of new classifiers can be done using the same NGS data that call for them. In these cases a manual classification is needed to train a new classifier for the desired properties, which can be used for further routine analyses subsequently. Classifier design is automated and requires, depending on the input sample amount, up to a few hours. This process is much faster than creating new primers for qPCR or designing new antibodies for an ELISA. Possibly most importantly, the creation of a new classifier is financially neglectable. This possibility to create classifiers quickly and cheaply can be used in highly versatile screening processes, which can be quickly adopted to new challenges.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

Further Work Due to time and financial constraints the amount of samples for this work was kept comparatively small. It would be very beneficial if further analyses would be run using a much larger sample set (Whitley *et al.*, 2002). A greater number of individual samples would allow for the creation of biologically independent test sets and training sets. Currently the sets are designed using different reads from all samples. While this methodology results in sufficiently independent sets for the comparison of two NGS based analyses, as was done here, further insights could be gained concerning the classification of entirely independent samples. During the analyses performed in this work, a single sample (A3.2) was treated as biologically independent (sections 2.7) and the results (section 3.3) suggest that the `srProfiler` approach classifies such a sample well and better than the transcriptome approach. However, these results, being produced with only one sample, cannot be regarded as statistically significant and require further research on a much larger sample set. Using more samples would also allow for time series analyses. While single samples were taken at different time points, the amount of samples for each time point is too small for time series analyses. Having a larger amount of samples per time point would provide the means to analyze the changes of the samples during the course of infection. The results of the pathogen alignment (section 3.4) suggest that the course of infection is different among different viruses. Assuming, therefore, that specific bins change differently over the courses of different infections, Hidden Markov Models (HMMs) (Baum *et al.*, 1970) could be designed, emitting the observed bins per state. Using those virus specific HMMs, it would be possible to not only classify the current state of the host but rather the complex changes over time undergone by the host in response to a specific infection. Further time series analyses could be run using one single host, for instance a member of the *Betula spp.*. This host would be sampled over a prolonged period of time, giving insights into the infection specific changes over time within one distinct host.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

The tool itself is dependent on the classifiers provided. Further experiments should be undertaken to construct more classifiers. Using public NGS data, it would be possible to design classifiers for a multitude of pathogens without the high initial sequencing costs. Consolidating new and improved classifiers in a public database would allow the scientific community to use and improve the `srProfiler` approach on a multitude of different experiments.

In order to validate the functionality of the `srProfiler` approach in a human diagnostic setting, the experiments should be repeated using mammalian hosts as samples or NGS patient data released for research purposes.

Future Prospects The potential of a generic diagnostic tool, that can utilize the huge amount of data produced by NGS without simultaneously being dependent on reference information such as reference genome or gene annotation, is enormous. In the areas of quality control, food security and market transparency a tool running many, in part custom, diagnoses in a single run is particularly interesting (Fox *et al.*, 2015; Mumford *et al.*, 2016; Büttner *et al.*, 2013). However, those fields that, to this day, suffer from the limited number of publicly accessible reference information and are confronted with numerous changing and new pathogens regularly, can utilize the enormous potential of NGS only slowly and with great effort. Using NGS based diagnostics will become continuously less expensive (Wetterstrand, 2013) and, utilizing technologies like minION (Jain *et al.*, 2016), accessible directly in the field. At this point preliminary analyses must be applicable quickly and directly, without prior in depth preprocessing and without the need for substantial bioinformatic expertise (Boonham *et al.*, 2014). Analyzing the quality of plant material should ideally result in a list of multiple pathogens tested in short succession after sequencing. This list should be quickly adaptable and new pathogens should be easily addable to be constantly up to date. The in depth analyses which will be required in the minority of

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

samples, i.e. assembling newly discovered pathogens, is done afterwards in a properly equipped lab with fewer time and infrastructure constraints. The `srProfiler` is the tool with which those preliminary, quickly accessible results can be generated. It will allow the personnel in the field to quickly screen many samples for many properties simultaneously and send only those samples to further analyses which fall into specific classes. The tool can be updated quickly with new classifiers, which were designed to screen for further properties of interest. Those new classifiers can be introduced to the tool directly in the field using, for instance, an email attachment. A public database holding an ever growing number of classifiers will increase the productivity and efficiency of the tool.

The amount of queries being run during an analysis is solely dependent on the amount of classifiers used. This scalability and modularity makes this tool usable for different areas, i.e. agricultural and pharmaceutical industry as well as basic research. While in an agricultural screening process the focus might lie on properties such as the resistance to certain chemicals, a pharmaceutical screening would be concentrated on the production of certain substances. A screening for basic research could be based on the search for new pathogens, thus finding all those samples infected with something that can not be classified to any known pathogen. Eventually the construction of classifiers for different fields and queries might result in the routine screening of any sample for every known pathogen.

Moreover, the generic nature of the tool makes it applicable to any genetic diagnostic, entirely independent of species. Designing specific classifiers will allow the screening of plants, while the same tool provided with different classifiers can be used to screen human patients for a multitude of properties, in a fast and straight forward manner.

Conclusion The PCAs show that the classifications can be done as well with the `srProfiler`, in some cases even better than using transcriptome expression pro-

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

files, mainly due to the use of all sequenced reads including NMR and MMR and the utter independence from host or pathogen related references. This proves that the `srProfiler` can be used to arrive at similar information as the transcriptome approach, while being entirely generic, faster and less computationally expensive (sections 2.9 and 3.3.2). The use of `srProfiler` based classifications can be entirely automated, arriving at an easy to use system reporting many information in a short time.

The gold standards for screening processes are ELISA and qPCR because both systems are reliable, versatile and comparatively easy to perform. The interpretation of the results is quick and straightforward (Boonham *et al.*, 2014). Even though NGS can provide a lot more information and the cost per sample is steadily decreasing, so far it is not used for routine analyses. As a major reason Boonham *et al.* (2014) regard the challenge of data analysis. On the other hand Adams *et al.* (2013) showed that in many cases qPCR and ELISA are too specific to discover a virus. This is a problem for new viruses but also for quasi-species of already known viruses (Capobianchi *et al.*, 2013). A fully automated screening process which can be quickly adopted to new information and is capable of general findings, such as the sample is infected with *some* virus, is very useful when the infection state of a sample cannot be directly observed due to the absence of symptoms or the sample being a seed (Fox *et al.*, 2015). In the areas of quality control and food-security, which are based on the routine analysis of many samples per day, this new system can greatly improve the detection of especially new or mutated pathogens which ELISA and qPCR are to specific for.

It was the development of real-time PCR that effectively turned PCR, a technique that was prone to contamination and required experienced molecular biologists, into a routine diagnostic tool that was robust and required a similar level of skill as did ELISA (Boonham *et al.*, 2008). The skills needed for NGS are already similar to ELISA, since ever new and increasingly automated sequencer designs are devel-

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

oped. So far it required experienced specialists to perform the bioinformatic interpretation. The use of the `srProfiler` requires a level of skill also required by real-time PCR and ELISA. Therefore NGS in concert with `srProfiler` could be for NGS what the development of real-time PCR was for PCR and help establish NGS as a robust and straight forward diagnostic tool in a routine setting.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

Bibliography

Adams M.J. and Antoniw J.F. DPVweb: a comprehensive database of plant and fungal virus genes and genomes. *Nucleic acids research*, 34(suppl 1):pages D382–D385. URL `http://www.dpvweb.net/`, 2006.

Adams I.P., Glover R.H., Monger W.A., Mumford R., Jackeviciene E., Navalinskiene M., Samuitiene M. and Boonham N. Next-generation sequencing and metagenomic analysis: a universal diagnostic tool in plant virology. *Molecular plant pathology*, 10(4):pages 537–545, 2009.

Adams I., Miano D., Kinyua Z., Wangai A., Kimani E., Phiri N., Reeder R., Harju V., Glover R. and Hany U. Use of next-generation sequencing for the identification and characterization of *Maize chlorotic mottle virus* and *Sugarcane mosaic virus* causing maize lethal necrosis in Kenya. *Plant Pathology*, 62(4):pages 741–749, 2013.

Al Rwahnih M., Daubert S., Golino D. and Rowhani A. Deep sequencing analysis of RNAs from a grapevine showing Syrah decline symptoms reveals a multiple virus infection that includes a novel virus. *Virology*, 387(2):pages 395–401, 2009.

Amberg R. private Communication, 2014.

Analytik Jena AG. *innuSPEED Tissue RNA Kit Manual*. Analytik Jena AG, hb_ks-2560_e_100303 edition, 2010.

Baker M. De novo genome assembly: what every biologist should know. *Nature methods*, 9(4):page 333, 2012.

von Bargen S., Langer J., Robel J., Rumbou A. and Büttner C. Complete nucleotide sequence of *Cherry leaf roll virus* (CLRV), a subgroup C nepovirus. *Virus research*, 163(2):pages 678–683, 2012.

Barzon L., Lavezzo E., Militello V., Toppo S. and Palù G. Applications of Next-Generation Sequencing Technologies to Diagnostic Virology. *International journal of molecular sciences*, 12(11):pages 7861–7884, 2011.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

Baum L.E., Petrie T., Soules G. and Weiss N. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The annals of mathematical statistics*, 41(1):pages 164–171, 1970.

Berardini T.Z., Reiser L., Li D., Mezheritsky Y., Muller R., Strait E. and Huala E. The Arabidopsis Information Resource: Making and mining the "gold standard" annotated reference plant genome. *genesis*, 53(8):pages 474–485, 2015.

BIOREBA AG. ELISA Data Analysis. Technical report, BIOREBA AG. URL `http://www.bioreba.ch/files/Tecnical_Info/ELISA_Data_Analysis.pdf`, 2014.

Bonham-Carter O., Steele J. and Bastola D. Alignment-free genetic sequence comparisons: a review of recent approaches by word analysis. *Briefings in bioinformatics*, page bbt052, 2013.

Boonham N., Glover R., Tomlinson J. and Mumford R. Exploiting generic platform technologies for the detection and identification of plant pathogens. *European Journal of Plant Pathology*, 121(3):pages 355–363, 2008.

Boonham N., Kreuze J., Winter S., van der Vlugt R., Bergervoet J., Tomlinson J. and Mumford R. Methods in virus diagnostics: From ELISA to next generation sequencing. *Virus research*, 186:pages 20–31, 2014.

Brown T.A. *Genomes*. Garland science, 3ed edition, 2006.

Büttner C., von Bargen S., Bandte M. and Mühlbach H.P. Forest diseases caused by viruses. *Infectious forest diseases*, pages 50–75, 2013.

Capobianchi M., Giombini E. and Rozera G. Next-generation sequencing technology in clinical virology. *Clinical Microbiology and Infection*, 19(1):pages 15–22, 2013.

Chen J., Zhang H., Feng M., Zuo D., Hu Y. and Jiang T. Transcriptome analysis of woodland strawberry (Fragaria vesca) response to the infection by Strawberry vein banding virus (SVBV). *Virology journal*, 13(1):page 128, 2016.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

Clark M.F. and Adams A. Characteristics of the microplate method of enzyme-linked immunosorbent assay for the detection of plant viruses. *Journal of general virology*, 34(3):pages 475–483, 1977.

Cock P.J., Fields C.J., Goto N., Heuer M.L. and Rice P.M. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic acids research*, 38(6):pages 1767–1771, 2010.

Collins L.J., Biggs P.J., Voelckel C. and Joly S. AN APPROACH TO TRANSCRIPTOME ANALYSIS OF NON-MODEL ORGANISMS USING SHORT-READ SEQUENCES. *Genome informatics*, 21:pages 3–14, 2008.

Cox A.J. ELAND: Efficient large-scale alignment of nucleotide databases. Illumina, San Diego, 2007.

De Haan P., Kormelink R., de Oliveira Resende R., Van Poelwijk F., Peters D. and Goldbach R. Tomato spotted wilt virus L RNA encodes a putative RNA polymerase. *Journal of General Virology*, 72(9):pages 2207–2216, 1991.

De Haan P., Wagemakers L., Peters D. and Goldbach R. The S RNA segment of tomato spotted wilt virus has an ambisense character. *Journal of General Virology*, 71(5):pages 1001–1007, 1990.

Dijkstra J. and de Jager C.P. *Practical Plant Virology: Protocols and Exercises*, chapter Mechanical Inoculation of Plants, pages 5–13. Springer Berlin Heidelberg, Berlin, Heidelberg. ISBN 978-3-642-72030-7. doi:10.1007/978-3-642-72030-7_1. URL `http://dx.doi.org/10.1007/978-3-642-72030-7_1`, 1998.

EFSA. Scientific Opinion on the pest categorisation of *Cherry leafroll virus*. *EFSA Journal*, 12(10):3848:page 23 pp. doi:10.2903/j.efsa.2014.3848, 2014.

Engvall E. and Perlmann P. Enzyme-linked immunosorbent assay (ELISA) quantitative assay of immunoglobulin G. *Immunochemistry*, 8(9):pages 871–874. doi:10.1016/0019-2791(71)90454-x. URL `http://dx.doi.org/10.1016/0019-2791(71)90454-X`, 1971.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

Ewing B. and Green P. Base-Calling of Automated Sequencer Traces Using *Phred*. II. Error Probabilities. *Genome research*, 8(3):pages 186–194, 1998.

Fisher R.A. *Statistical Methods for Research Workers*. Genesis Publishing Pvt Ltd, 1925.

Fox A., Adams I., Hany U., Hodges T., Forde S., Jackson L., Skelton A. and Barton V. The application of Next-Generation Sequencing for screening seeds for viruses and viroids. *Seed Science and Technology*, 43(3):pages 531–535, 2015.

Gogol-Döring A. and Chen W. An Overview of the Analysis of Next Generation Sequencing Data. *Next Generation Microarray Bioinformatics: Methods and Protocols*, pages 249–257, 2012.

Hadidi A., Barba M., Candresse T. and Jelkmann W. *Virus and virus-like diseases of pome and stone fruits*. The American Phytopathological Society, 2011.

Hadidi A., Flores R., Candresse T. and Barba M. Next-Generation Sequencing and Genome Editing in Plant Virology. *Frontiers in Microbiology*, 7, 2016.

Hollings M. HOST RANGE STUDIES WITH FIFTY-TWO PLANT VIRUSES. *Annals of Applied Biology*, 47(1):pages 98–108, 1959.

Hollings M. RECENT ADVANCES IN VIRUS DETECTION AND IDENTIFICATION BY BIOASSAY AND SEROLOGICAL TESTS. In *III International Symposium on Virus Diseases of Ornamental Plants 36*, pages 23–34, 1972.

Hull R. *Plant virology*. Academic press, 2013.

illumina Inc. Illumina Sequencing Technology Highest data accuracy, simple workflow, and a broad range of applications. Technical report, illumina Inc. URL `http://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf`, 2010.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

illumina Inc. Sequencing Technology Video. URL `http://www.illumina.com/SBSvideo`, 2014.

illumina Inc. TruSeq Stranded Total RNA Library Preparation Kit with Ribo-Zero Plant. Technical Report 770-2013-006, 2015.

illumina Inc. NextSeq® 500 System Guide. Technical Report 15046563_v02, illumina Inc., 2016.

Jain M., Olsen H.E., Paten B. and Akeson M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology*, 17(1):page 239, 2016.

Jalkanen R., Büttner C. and Bargen S.v. *Cherry leaf roll virus* abundant on *Betula pubescens* in Finland. *Silva Fennica*, 41(4):pages 755–762, 2007.

Khan J.A. and Dijkstra J. *Plant Viruses As Molecular Pathogens*. CRC Press, 2001.

Kormelink R., De Haan P., Meurs C., Peters D. and Goldbach R. The nucleotide sequence of the M RNA segment of tomato spotted wilt virus, a bunyavirus with two ambisense RNA segments. *Journal of General Virology*, 73(11):pages 2795–2804, 1992.

Kreuze J.F., Perez A., Untiveros M., Quispe D., Fuentes S., Barker I. and Simon R. Complete viral genome sequence and discovery of novel viruses by deep sequencing of small RNAs: a generic method for diagnosis, discovery and sequencing of viruses. *Virology*, 388(1):pages 1–7, 2009.

Kruskal J.B. Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, 29(2):pages 115–129, 1964.

Langmead B. and Salzberg S.L. Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9(4):pages 357–359, 2012.

Langmead B., Trapnell C., Pop M. and Salzberg S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biol*, 10(3):page R25, 2009.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

Li H. and Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14):pages 1754–1760, 2009.

Lodish H., Berk A., Zipursky S.L., Matsudaira P., Baltimore D. and Darnell J. *Molecular Cell Biology*, volume 4. WH Freeman New York, 2000.

Mandal B., Csinos A., Martinez-Ochoa N. and Pappu H. A rapid and efficient inoculation method for *Tomato spotted wilt tospovirus*. *Journal of virological methods*, 149(1):pages 195–198, 2008.

Meinke D.W., Cherry J.M., Dean C., Rounsley S.D. and Koornneef M. *Arabidopsis thaliana*: A Model Plant for Genome Analysis. *Science*, 282(5389):pages 662–682, 1998.

Menzel W. DSMZ Plant Virus Collection. private Communication, 2016.

Mumford R., Macarthur R. and Boonham N. The role and challenges of new diagnostic technology in plant biosecurity. *Food Security*, 8(1):pages 103–109, 2016.

Nagalakshmi U., Wang Z., Waern K., Shou C., Raha D., Gerstein M. and Snyder M. The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. *Science*, 320(5881):pages 1344–1349, 2008.

Nagano A.J., Honjo M.N., Mihara M., Sato M. and Kudoh H. Detection of Plant Viruses in Natural Environments by Using RNA-Seq. *Plant Virology Protocols: New Approaches to Detect Viruses and Host Responses*, pages 89–98, 2015.

Pearson K. On Lines and Planes of Closest Fit to Systems of Point in Space. *Philosophical Magazine*, 2(11):pages 559–572, 1901.

Pontier D., Balagué C. and Roby D. The hypersensitive response. A programmed cell death associated with plant resistance. *Comptes Rendus de l'Académie des Sciences-Series III-Sciences de la Vie*, 321(9):pages 721–734, 1998.

Posnette A. and Cropley R. Leaf roll: a virus disease of Cherry. *Report. East Malling Research Station*, pages 126–127, 1955.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

Prabha K., Baranwal V. and Jain R. Applications of Next Generation High Throughput Sequencing Technologies in Characterization, Discovery and Molecular Interaction of Plant Viruses. *Indian Journal of Virology*, 24(2):pages 157–165, 2013.

Rhee S.Y., Beavis W., Berardini T.Z., Chen G., Dixon D., Doyle A., Garcia-Hernandez M., Huala E., Lander G. and Montoya M. The *Arabidopsis* Information Resource (TAIR): a model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. *Nucleic acids research*, 31(1):pages 224–228, 2003.

Rumbou A., von Bargen S. and Büttner C. A model system for plant-virus interaction–infectivity and seed transmission of Cherry leaf roll virus (CLRV) in Arabidopsis thaliana. *European journal of plant pathology*, 124(3):pages 527–532, 2009.

Rumbou A., von Bargen S., Demiral R., Langer J., Rott M., Jalkanen R. and Büttner C. High genetic diversity at the inter-/intra-host level of *Cherry leaf roll virus* population associated with the birch leaf-roll disease in Fennoscandia. *Scandinavian Journal of Forest Research*, pages 1–15, 2016.

Rupert J.TOMATO SPOTTED WILT VIRUS. *Advances in virus research*, 13:page 65, 1968.

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

Schmelzer K. Studies on viruses of ornamental and wild woody plants. 2nd Part. Viroses of Forsythia, Lonicera, Ligustrum, and Laburnum. *Phytopathologische Zeitschrift*, 46(2):pages 105–138, 1962.

Selvarajan R. and Balasubramanian V. Cutting-Edge Technologies for Detection of Plant Viruses in Vegetatively Propagated Crop Plants. In *Plant Viruses: Evolution and Management*, pages 53–71. Springer, 2016.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

Sherwood J.L., German T.L., Moyer J.W., Ullman D.E. and Whitfield A. Tomato spotted wilt virus. *Encyclopedia of Plant Pathology. OC Maloy and TD Murray, eds. John Wiley & Sons, New York*, pages 1030–1031, 2000.

Sims G.E., Jun S.R., Wu G.A. and Kim S.H. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proceedings of the National Academy of Sciences*, 106(8):pages 2677–2682, 2009.

Sims D., Sudbery I., Ilott N.E., Heger A. and Ponting C.P. Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics*, 15(2):pages 121–132, 2014.

Smith K. and Markham R. Two new viruses affecting Tobacco and other plants. *Phytopathology*, 34(3):pages 324–329, 1944.

Song K., Ren J., Reinert G., Deng M., Waterman M.S. and Sun F. New developments of alignment-free sequence comparison: measures, statistics and next-generation sequencing. *Briefings in bioinformatics*, page bbt067, 2013.

Student. THE PROBABLE ERROR OF A MEAN. *Biometrika*, pages 1–25, 1908.

Studholme D.J., Glover R.H. and Boonham N. Application of High-Throughput DNA Sequencing in Phytopathology. *Annual review of phytopathology*, 49:pages 87–105, 2011.

Voller A., Bidwell D. and Bartlett A. Enzyme immunoassays in diagnostic medicine: Theory and practice*. *Bulletin of the World Health Organization*, 53(1):page 55, 1976.

Wang Z., Gerstein M. and Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):pages 57–63, 2009.

Ward J.A., Ponnala L. and Weber C.A. Strategies for transcriptome analysis in non-model plants. *American Journal of Botany*, 99(2):pages 267–276, 2012.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

Weemen B.V. and Schuurs A. IMMUNOASSAY USING ANTIGEN-ENZYME CONJUGATES. *FEBS Letters*, 15(3):pages 232–236. doi:10.1016/0014-5793(71)80319-8. URL `http://dx.doi.org/10.1016/0014-5793(71)80319-8`, 1971.

Welch B.L. THE GENERALIZATION OF 'STUDENT'S' PROBLEM WHEN SEVERAL DIFFERENT POPULATION VARIANCES ARE INVOLVED. *Biometrika*, 34(1/2):pages 28–35, 1947.

Wetterstrand K. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). URL `http://www.genome.gov/sequencingcostsdata`. Accessed 2017/02/25, 2013.

Wetzel T., Beck A., Wegener U. and Krczal G. Complete nucleotide sequence of the RNA 1 of a grapevine isolate of *Arabis mosaic virus*. *Archives of virology*, 149(5):pages 989–995, 2004.

Wetzel T., Meunier L., Jaeger U., Reustle G. and Krczal G. Complete nucleotide sequences of the RNAs 2 of German isolates of Grapevine fanleaf and Arabis mosaic nepoviruses. *Virus Research*, 75(2):pages 139–145, 2001.

Whitley E. and Ball J. Statistics review 4: Sample size calculations. *Critical care*, 6(4):page 335, 2002.

Wilhelm B.T., Marguerat S., Watt S., Schubert F., Wood V., Goodhead I., Penkett C.J., Rogers J. and Bähler J. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*, 453(7199):pages 1239–1243, 2008.

Wilson K. and Walker J. *Principles and Techniques of Biochemistry and Molecular Biology*. Cambridge university press, 2010.

Yanagisawa H., Tomita R., Katsu K., Uehara T., Atsumi G., Tateda C., Kobayashi K. and Sekine K.T. Combined DECS Analysis and Next-Generation Sequencing

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

Enable Efficient Detection of Novel Plant RNA Viruses. *Viruses*, 8(3):page 70, 2016.

Yates A., Akanni W., Amode M.R., Barrell D., Billis K., Carvalho-Silva D., Cummins C., Clapham P., Fitzgerald S. and Gil L. Ensembl 2016. *Nucleic acids research*, 44(D1):pages D710–D716. Ensembl Plants release 35 - April 2017, 2016.

Yockteng R., Almeida A.M., Yee S., Andre T., Hill C. and Specht C.D. A method for extracting high-quality RNA from diverse plants for next-generation sequencing and gene expression analyses. *Applications in plant sciences*, 1(12):page 1300070, 2013.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

A Appendix

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

A.1 Sequencing

Table 9: The alignment results for *Arabidopsis thaliana* to the reference genome tair v10.29 is shown in this table. The results are listed in number of UMR, MMR and NMR, as well as in percent of total reads respectively.

Name	UMR	UMR in %	MMR	MMR in %	NMR	NMR in %	Sum
A1.1	52.463.330	17%	31.560.087	10%	219.029.238	72%	303.052.655
A1.2	73.740.146	25%	36.853.748	12%	185.532.901	63%	296.126.795
A1.3	75.725.380	28%	36.219.236	13%	161.424.024	59%	273.368.640
A2.1	12.461.697	2%	5.705.483	1%	511.905.996	97%	530.073.176
A2.2	13.453.419	2%	5.964.504	1%	719.708.087	97%	739.126.010
A2.3	52.544.655	24%	30.842.715	14%	134.111.913	62%	217.499.283
A3.1	18.569.940	3%	13.282.522	2%	545.470.901	94%	577.323.363
A3.2	32.619.365	5%	21.274.844	3%	563.032.476	91%	616.926.685
A3.3	56.522.308	22%	26.839.853	10%	173.516.618	68%	256.878.779
C1.1	15.875.660	10%	10.241.962	6%	139.989.290	84%	166.106.912
C1.2	83.355.670	26%	43.714.195	14%	190.418.592	60%	317.488.457
C1.3	163.156.701	28%	84.466.277	15%	327.300.126	57%	574.923.104
C2.1	459.755.478	29%	202.836.209	13%	931.306.716	58%	1.593.898.403
C2.2	63.264.212	26%	28.656.381	12%	147.713.692	62%	239.634.285
C2.3	208.088.567	17%	87.332.317	7%	955.009.158	76%	1.250.430.042
C3.1	72.544.036	25%	36.617.816	13%	181.048.660	63%	290.210.512
C3.2	77.616.586	26%	40.529.599	13%	183.171.189	61%	301.317.374
C3.3	33.721.224	14%	21.268.476	9%	184.858.749	77%	239.848.449
T1.1	89.910.858	26%	53.414.613	16%	200.100.103	58%	343.425.574
T1.2	75.070.564	24%	43.933.038	14%	197.920.255	62%	316.923.857
T1.3	360.464.019	25%	302.074.747	21%	807.378.439	55%	1.469.917.205
T2.1	80.214.378	27%	36.088.188	12%	178.605.715	61%	294.908.281
T2.2	82.494.540	29%	42.959.748	15%	154.170.298	55%	279.624.586
T2.3	75.848.653	28%	36.925.567	14%	157.176.279	58%	269.950.499
T3.1	51.023.679	15%	100.261.013	29%	188.885.988	56%	340.170.680
T3.2	74.796.657	26%	40.775.523	14%	174.307.564	60%	289.879.744
T3.3	111.349.465	31%	44.703.945	12%	205.971.416	57%	362.024.826
X1.1	37.953.382	15%	21.842.188	9%	188.538.775	76%	248.334.345
X1.2	369.548.125	27%	168.318.725	12%	837.907.544	61%	1.375.774.394
X1.3	341.714.416	29%	166.149.012	14%	673.161.760	57%	1.181.025.188
X2.1	96.027.393	28%	47.005.115	14%	202.207.308	59%	345.239.816
X2.2	285.816.275	29%	152.681.793	16%	535.155.728	55%	973.653.796
X2.3	19.796.827	13%	10.632.604	7%	123.647.707	80%	154.077.138
X3.1	93.767.839	11%	43.113.707	5%	735.066.260	84%	871.947.806
X3.2	93.575.348	28%	45.251.537	13%	198.864.784	59%	337.691.669
X3.3	85.937.189	28%	35.944.937	12%	188.040.680	61%	309.922.806

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

A.2 Pathogen Alignment

Table 10: The table shows the results of the pathogen alignment. The percentages of reads being sequenced per sample which aligned to the respective pathogens is given.

Sample	Inoculation Virus	% Reads aligned in ArMV	% Reads aligned in TSWV	% Reads aligned in CLRV
A01	ArMV	0,00611271	0	3,5782E-08
A02	ArMV	0,46591602	0	3,6049E-07
A03	ArMV	0,50614392	0	9,6058E-09
A1.1	ArMV	0,02850755	0	2,0793E-05
A1.2	ArMV	0,00772373	0	8,5821E-06
A1.3	ArMV	0,0043374	0	9,8622E-06
A2.1	ArMV	0,46875624	0	0
A2.2	ArMV	0,50531034	0	9,3302E-06
A2.3	ArMV	0,00094264	0	2,8106E-06
A3.1	ArMV	0,45243521	0	1,2041E-05
A3.2	ArMV	0,35900543	0	3,9933E-05
A3.3	ArMV	0,00173678	0	2,3458E-06
C01	CLRV	0,00060436	0	0,00090573
C02	CLRV	0,00020776	0	0,15294738
C03	CLRV	0,00018586	0	0,0534406
C1.1	CLRV	0,00031066	0	0,00094895
C1.2	CLRV	9,8583E-05	0	0,00461201
C1.3	CLRV	0,00017594	0	0,00188847
C2.1	CLRV	0,0003358	3,7915E-05	0,0044042
C2.2	CLRV	0,00030455	0	0,00038027
C2.3	CLRV	5,425E-05	0	0,00111832
C3.1	CLRV	0,00016775	0	0,00013372
C3.2	CLRV	0,00018474	0	0,00142288
C3.3	CLRV	0,00502864	0	0,00018522
T01	TSWV	1,1824E-05	0,00087498	0,00015192
T02	TSWV	2,2715E-05	4,1535E-06	3,3941E-05
T03	TSWV	0,0008638	0	0,0001178
T1.1	TSWV	0,00024195	1,1765E-06	0,00020649
T1.2	TSWV	3,066E-05	0	1,6786E-06
T1.3	TSWV	4,8818E-05	0	7,4245E-06
T2.1	TSWV	6,4237E-05	0	3,9656E-06
T2.2	TSWV	6,523E-05	0	1,4165E-08
T2.3	TSWV	9,357E-05	0,00011489	1,4918E-06
T3.1	TSWV	0,00020004	0	0
T3.2	TSWV	0,00043954	0	1,3009E-06
T3.3	TSWV	0,00059402	0	2,001E-06
X01	Control	5,7861E-05	0	0,00144467
X02	Control	0,00011902	0	5,8034E-05
X03	Control	7,3236E-05	0	4,5475E-05
X1.1	Control	0,00151415	0	1,0103E-05
X1.2	Control	0,0001017	0	8,8079E-07
X1.3	Control	0,00020818	0	3,8235E-07
X2.1	Control	0,00031411	0	1,2323E-05
X2.2	Control	8,8172E-05	0	4,5033E-07
X2.3	Control	0,00048675	0	3,4232E-06
X3.1	Control	0,0011842	0	1,5163E-05
X3.2	Control	0,00124124	0	3,8237E-08
X3.3	Control	0,00030074	0	4,9458E-09

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

Eidesstattliche Erklärung

Hiermit versichere ich, dass ich die vorliegende Dissertation selbständig und ohne die Zusammenarbeit mit gewerblichen Promotionsberatern, auf der Grundlage der angegebenen Hilfsmittel und Hilfen unter Kenntnisnahme der zu Grunde liegenden Promotionsordnung der Lebenswissenschaftlichen Fakultät, veröffentlicht im Amtlichen Mitteilungsblatt der Humboldt Universität zu Berlin Nr. 12/2015 vom 05.03.2015, verfasst habe. Die Dissertation wurde weder in Teilen noch als Ganzes bei einer anderen wissenschaftlichen Einrichtung eingereicht, angenommen oder abgelehnt. Ich versichere weiterhin, dass ich mich nicht anderwärts um einen Doktorgrad beworben habe, bzw. einen entsprechenden Doktorgrad besitze und dass die Grundsätze der Humboldt Universität zu Berlin zur Sicherung guter wissenschaftlicher Praxis eingehalten wurden.

Datum, Ort Steffen Pallarz

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

In der Reihe *Berliner ökophysiologische und phytomedizinische Schriften* sind bisher erschienen:

Band 01: Mohammad Mahir Uddin (2009)
Chemical ecology of mustard leaf beetle Phaedon cochleariae (F.).
ISBN 978-3-89959-848-3.

Band 02: Ilir Morina (2009)
Entwicklung von Verfahren zur Rekultivierung der Aschedeponie des Braunkohlekraftwerks in Prishtina (Kosovo).
ISBN 978-3-89959-872-8.

Band 03: Melanie Wiesner (2009)
Veränderungen gesundheitsrelevanter Inhaltsstoffe in *Parthenium* hysterophorus L. in Abhängigkeit von der Pflanzengröße und Klimafaktoren.
ISBN 978-3-89959-880-3.

Band 04: Fransika Rohr (2009)
Variabilität aliphatischer Glucosinolate in *Arabidopsis thaliana*-Ökotypen und deren Einfluss auf die Wirtspflanzeneignung von zwei folivoren Insektenarten.
ISBN 978-3-89959-884-9.

Band 05: Jutta Buchhop (2009)
Characterization of phylogenetically diverse CLRV-isolates by RFLP and research into identification of two isometric viruses.
ISBN 978-3-89959-929-9.

Band 06: Nora Koim (2010)
Urban sprawl, land cover change and forest fragmentation – Case study Pereira, Colombia.
ISBN 978-3-89959-955-8.

Band 07: Nadja Förster (2010)
Eignung unterschiedlicher salicylathaltiger Salix-Klone für die Arzneimittelindustrie.
ISBN 978-3-89959-964-0.

Band 08: Jana Gentkow (2010)
Cherry leaf roll virus (CLRV): Charakterisierung ausgewählter Virusisolate unter besonderer Berücksichtigung des viralen Hüllproteins.
ISBN 978-3-89959-976-3.

Band 09: Ahmad Fakhro (2010)
Interaction of Pepino mosaic virus (PepMV) and fungal root endophytes with tomato hosts (*Lycopersicum esculentum* Mill.).
ISBN 978-3-89959-995-4.

Band 10: Stefan Irrgang (2010)
Mikro- und makroskopische Untersuchungen an Veredelungsstellen von Straßenbäumen im Hinblick auf die Beeinflussung ihrer Bruchsicherheit.
ISBN 978-3-89959-998-5.

Band 11: Julia Jahnke (2010)
Guerilla Gardening anhand von Beispielen in New York, London und Berlin.
ISBN 978-3-86247-001-3.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

Band 12: Astrid Karoline Günther (2010)
Analysen zur Intensität der Pflanzenschutzmittel-Anwendung und Aufklärung ihrer Einflussfaktoren in ausgewählten Ackerbaubetrieben.
ISBN 978-3-86247-005-1.

Band 13: Milena A. Dimova (2010)
Untersuchungen zur Epidemiologie von *Pythium aphanidermatum* in Abhängigkeit von den Umgebungsbedingungen bei der Gewächshausgurke (*Cucumis sativus* L.).
ISBN 978-3-86247-033-4.

Band 14: Claudia Patricia Pérez-Rodríguez (2010)
Physiologische Veränderungen in Früchten der Solanaceaengewächse in Abhängigkeit von physikalischen Elicitoren während der Produktion und nach der Ernte.
ISBN 978-3-86247-066-2.

Band 15: Charles Adarkwah (2010)
Integrated management of the stored-product pest insects *Corcyra cephalonica*, *Cadra cautella*, *Sitophilus zeamais* and *Tribolium castaneum* by use of the parasitic wasps *Habrobracon hebetor*, *Venturia canescens*, *Lariophagus distinguendus* and neem seed oil.
ISBN 978-3-86247-077-8.

Band 16: Christoph von Studzinski (2010)
Angewandte Methoden der xenovegetativen Vermehrung.
ISBN 978-3-86247-088-4.

Band 17: Tanja Mucha-Pelzer (2011)
Amorphe Silikate – Möglichkeiten des Einsatzes im Gartenbau zur physikalischen Schädlingsbekämpfung.
ISBN 978- 3-86247-106-5.

Band 18: Diego Miranda (2011)
Effect of salt stress on physiological parameters of cape gooseberry, *Physalis peruviana* L.
ISBN 978- 3-86247-119-5

Band 19: Franziska Beran (2011)
Host preference and aggregation behavior of the striped flea beetle, *Phyllotreta striolata*.
ISBN 978- 3-86247-188-1

Band 20: Mohammed Abul Monjur Khan (2011)
Induced biochemical changes and gene expression in *Brassica oleracea* and *Arabidopsis thaliana* by drought stress and its consequences on resistance to aphids.
ISBN 978- 3-86247-203-1.

Band 21: Sandra Lerche (2012)
Untersuchungen zur Anwendung, Praxiseinführung und molekularen Identifizierung von Stamm V24 des entomopathogenen Pilzes *Lecanicillium muscarium* (Petch) Zare & W. Gams.
ISBN 978- 3-86247-248-2.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

Band 22: Carsten Richter (2012)
Entwicklung und Überprüfung eines gasdichten Küvettensystems für Experimente unter hochgradig kontrollierten Bedingungen mit Gaswechselmessungen.
ISBN 978- 3-86247-271-0.

Band 23: Aksana Grineva (2012)
Influence of the two stored grain pest insects *Sitophilus granarius* and *Oryzaephilus surinamensis* on temperature, relative humidity, moisture content, and mould growth in stored triticale.
ISBN 978- 3-86247-279-6.

Band 24: Carmen Büttner & Christian Ulrichs (2012)
Aktuelle Themen in Landwirtschaft und Gartenbau am Beispiel von Südtirol.
ISBN 978- 3-86247-279-6.

Band 25: Juliane Langer (2012)
Molecular and epidemiological characterisation of Cherry leaf roll virus (CLRV).
ISBN 978- 3-86247-279-6.

Band 26: Franziska Rohr-Doucet (2012)
AOP-Variabilität in *Arabidopsis thaliana*-Kreuzungslinien – Auswirkungen auf die Resistenz gegenüber verschieden spezialisierten Lepidopteren-Arten.
ISBN 978- 3-86247-329-8.

Band 27: Vanessa Hörmann (2012)
Lignin als biologische Barriere gegen Schimmelpize in Innenräumen.
ISBN 978- 3-86247-330-4.

Band 28: Jacqueline Kurth (2013)
Auswirkungen verschiedener Düngerzusammensetzungen auf den Ertrag bei Schnittrosen unter Berücksichtigung des Anbauverfahrens.
ISBN 978- 3-86247-336-6.

Band 29: Juliane Langer, Carmen Büttner & Christian Ulrichs (2014)
Kolumbien – klimatische und politische Voraussetzungen für eine landwirtschaftliche Produktion.
ISBN 978- 3-86247-430-1.

Band 30: Heike Luisa Dieckmann (2014)
Detection of the European mountain ash ringspot associated virus (EMARaV) in Sorbus aucuparia L. in several European contries.
ISBN 978- 3-86247-441-7.

Band 31: Rima Marion Baag (2014)
Analyse von trans-Resveratrol in historischen Rebsorten der Weinanbaugebiete Sachsen und Saale-Unstrut.
ISBN 978- 3-86247-488-2.

Band 32: Ayesha Rahmann (2014)
Study of the protective effects of nano-structured silica and plant derived biomolecules on nuclear polyhedrosis virus affected silkworm larvae at the behavioral and molecular level.
ISBN 978- 3-86247-495-0.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

Band 33: Bettina Gramberg (2015)
Weiterentwicklung eines elektrochemischen Biosensors zum Nachweis von Pflanzenviren und Insektiziden.
ISBN 978- 3-86247-512-4.

Band 34: Wilhelm van Husen (2015)
Artspezifische Aufnahme und Verteilung von Cadmium bei indigenen afrikanischen Gemüsearten und daraus abzuleitende Ernährungsempfehlungen.
ISBN 978- 3-86247-523-0.

Band 35: Jenny Roßbach (2015)
European mountain ash ringspot-associated viras (EMARaV): diversity and geographic distribution in Europe.
ISBN 978- 3-86247-547-6.

Band 36: Silke Steinmöller (2015)
Risikominderung der Verbreitung von Quarantäneschadorganismen der Kartoffel durch hygienisierende Maßnahmen.
ISBN 978- 3-86247-550-6.

Band 37: Christin Siewert (2016)
Genomic and functional analysis of species within the Acholeplasmataceae – Phytoplasmas and Acholeplasmas.
ISBN 978- 3-86247-579-7.

Band 38: Angela Köhler (2016)
Untersuchungen zur Phenolglycosidkonzentration ausgewählter intra- und interspezifischer Kreuzungen salicinreicher Biomasseweiden.
ISBN 978- 3-86247-581-0.

Band 39: Nicolas Meyer (2016)
Vergleichende ökophysiologische Untersuchung verschiedener Baumarten zur Verwendung als Straßenbegleitgrün in Berlin.
ISBN 978- 3-86247-586-5.

Band 40: Stefanie Schläger (2017)
Identification of variation within sex pheromone blends of various Maruca vitrata populations for refining pheromone lures and traps in Asia.
ISBN 978- 3-7369-9570-3.

Band 41: Elisha Otieno Gogo (2017)
Pre- and postharvest treatments for the quality assurance of African indigenous leafy vegetables.
ISBN 978-3-7369-9650-2.

Band 42: Luise Dierker (2017)
Interaktion des RNA2-kodierten Transportproteins (MP) des *Cherry leaf roll virus* (CLRV) mit dem viralen Hüllprotein (CP) und pflanzlichen Wirtsfaktoren
ISBN 978-3-7369-9670-0.

Band 43: Nadja Förster (2017)
Antikarzinogenes Potential ausgewählter Glucosinolate von *Moringa oleifera*
ISBN 978-3-7369-9704-2.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.

Dieses Werk ist copyrightgeschützt und darf in keiner Form vervielfältigt werden noch an Dritte weitergegeben werden.
Es gilt nur für den persönlichen Gebrauch.