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).

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

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-

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

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).

2 Materials and Methods

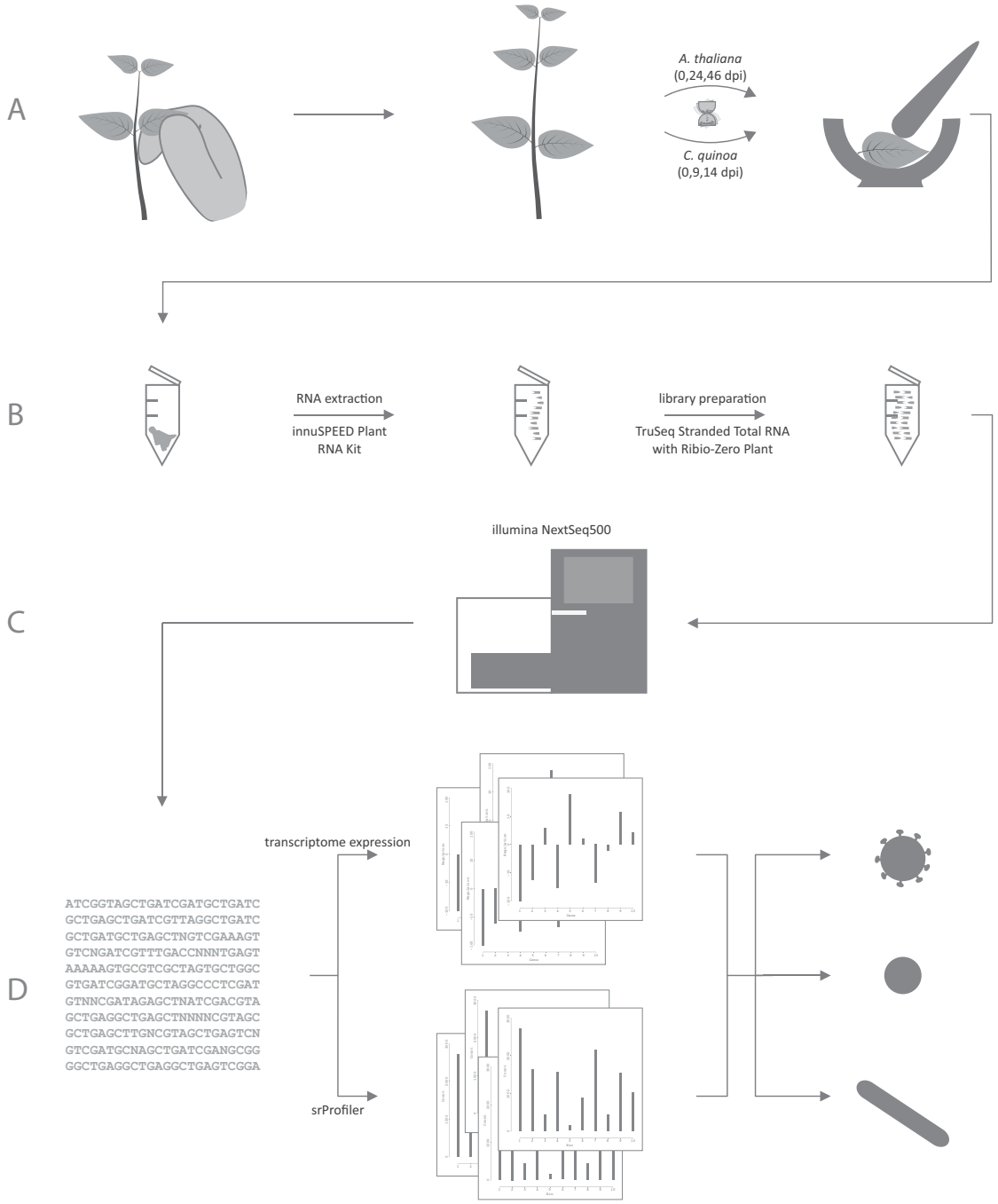

Figure 1: The experimental design is shown from inoculation and harvesting (A) over sample preparation for sequencing (B), sequencing (C) and subsequent data analysis resulting in profiles which classify the pathogens (D).

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).

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.

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*