Prediction of the Molecular Boundary and Evolutionary History of Novel Viral AlkB Domains Using Homology Modeling and Principal Component Analysis

By

Clayton Moore

A Thesis
presented to
The University of Guelph

In partial fulfilment of requirements
for the degree of
Master of Science
in
Bioinformatics

Guelph, Ontario, Canada
© Clayton Moore, September, 2018
AlkB (Alkylation B) proteins are ubiquitous among cellular organisms where they act to reverse the damage in RNA and DNA due to methylation. The AlkB protein family is so significant to all forms of life that it was recently discovered that an AlkB domain is encoded as part of the replicase (poly)protein in a small subset of single-stranded, positive-sense RNA viruses belonging to five families. Interestingly, these AlkB-containing viruses are mostly important pathogens of perennials such as fruit crops. As a newly identified protein domain in RNA viruses, the origin, molecular boundary of the viral AlkB domain as well as its function in viral replication and infection are unknown. The poor sequence conservation of this domain has made traditional bioinformatic approaches unreliable and inconclusive. Here we apply principal component analysis, homology modelling, and traditional sequence based techniques to propose a molecular boundary and function to this novel viral domain.
ACKNOWLEDGEMENTS

This research project was supported by a grant from the Discovery Grant program (Project RGPIN-2014-05306) of the Natural Science and Engineering Research Council of Canada.

First and foremost I would like to thank my advisor, Dr. Baozhong Meng, for all of his patience and guidance throughout my entire M.Sc. program. I am tremendously grateful for his guidance, expertise, and his encouragement to help me achieve my highest potentials.

I would like to additionally thank the members of my advisory committee, Drs. Steffen Graether, Lewis Lukens and Jinzhong Fu, for advice and scientific input throughout this project.

I would like to thank my friends, family, and lab members as without their support throughout this entire endeavor.
TABLE OF CONTENTS

Abstract ........................................................................................................................................... ii
Acknowledgements ......................................................................................................................... iii
Table of Contents ............................................................................................................................ iv
List of Tables .................................................................................................................................. vi
List of Figures .................................................................................................................................. vii
List of Abbreviations ....................................................................................................................... viii

1: Introduction ................................................................................................................................. 1

1.1 Grapevine history ....................................................................................................................... 1
1.2 Canadian grapevine industry and its economic impact .......................................................... 2
1.3 AlkB ........................................................................................................................................... 3
   1.3.1 AlkB in bacteria .................................................................................................................. 7
   1.3.2 Mammalian AlkB .............................................................................................................. 9
1.4 Sequence conservation and structural features of cellular AlkB proteins... ....................................... 11
1.5 An overview of the current state of knowledge of viral AlkB domains .... 13
   1.5.1 Closteroviridae ................................................................................................................. 15
   1.5.2 Betaflexiviridae .............................................................................................................. 18
   1.5.3 Alphaflexiviridae ............................................................................................................ 21
   1.5.4 Proposed functions of vAlkB .......................................................................................... 21
   1.5.5 Function of AlkB in viruses remains largely unknown and understudied despite a single paper.......................................................... 23
   1.5.6 Proposed origin and evolutionary history of the viral AlkB domain 28
   1.5.7 Recent evidence suggests that plant infecting viruses can hijack their hosts' AlkB .............................................................................................................................. 32
   1.5.8 Conclusion of biological studies of the vAlkB domain ............................................ 33
1.6 Methodologies used to study the novel vAlkB ........................................................................ 33
   1.6.1 Multiple sequence analysis ............................................................................................... 34
   1.6.2 Homology modeling .......................................................................................................... 35
1.6.3 Principal component analysis .......................................................... 35
1.6.4 Shannon diversity index .................................................................. 36
1.7 Significance ......................................................................................... 37
1.8 Hypotheses and objectives ................................................................. 38

2: MATERIALS AND METHODS ................................................................. 40
2.1 Sequence preparations ......................................................................... 40
2.2 Protein modelling ................................................................................ 41
2.3 Principal component analysis .............................................................. 42
2.4 Shannon diversity index ....................................................................... 43

3: RESULTS ................................................................................................ 45
3.1 AlkB-containing viruses are unique in that they are mostly pathogens of perennial plants and are phloem-limited .......................................................... 45
3.2 Multiple bioinformatic analyses suggest a functional viral AlkB of between 150 and 170 amino acids in size ................................................................. 50
3.3 The first viral AlkB was likely acquired by an ancestral virus of the family Alphaflexiviridae .................................................................................. 60
3.4 Identification of additional conserved amino acid residues in viral AlkB proteins ....................................................................................... 64

4: DISCUSSION ........................................................................................... 66
4.1 Homology modeling and principal component analysis suggest functional vAlkBs to comprise 150-170 amino acids ............................................. 66
4.2 Results of in silico analysis point to the acquisition of AlkB by an ancestral virus of the family Alphaflexiviridae ......................................................... 70
4.3 Identification of expended conserved functional motifs in viral AlkBs. ... 72
4.4 Viral AlkB is unique to RNA viruses that are pathogens of perennial plants with a tissue tropism to the phloem ...................................................... 73
4.5 Conclusion ............................................................................................. 75

5: Works Cited ............................................................................................ 77
LIST OF TABLES

Table 1: AlkB containing viruses .................................................................47
LIST OF FIGURES

Figure 1.1: Examples of oxidative damaged repaired by AlkB proteins ................................. 4

Figure 1.2: The overall tree of the AlkB family with representative members from cellular and viral organisms .................................................................................................................. 7

Figure 1.3: Structural aspects of E. coli AlkB protein ................................................................ 12

Figure 1.4: Symptoms of Grapevine leafroll disease (GLD) on a red cultivar ...................... 16

Figure 1.5: Genome map of Grapevine leafroll-associated virus 3 (GLRaV-3) ............... 17

Figure 1.6: Genome map of (Grapevine rupestris stem pitting-associated virus (GRSPaV) ................................................................................................................................. 19

Figure 1.7: Genome map of Grapevine Virus E ........................................................................ 20

Figure 1.8: Viral AlkB multiple sequence alignment of viral AlkB proteins used in the Van den Born (2008) study ........................................................................................................ 25

Figure 1.9: Pairwise distance comparison of three essential viral (poly)protein domains compared to viral AlkB .......................................................................................................... 27

Figure 1.10: Phylogeny of predicted viral AlkB origins (Martelli et al., 2007) ............... 29

Figure 1.11: Pairwise distance alignment of viral AlkBs (Bratlie and Drabløs, 2005) .. 31

Figure 3.1: Multiple sequence alignment of viral AlkBs from members of each viral family ................................................................................................................................. 49

Figure 3.2: Multiple sequence alignment of Grapevine virus B isolates ........................ 51-52

Figure 3.3: Homology modelling results for one of the viruses tested through phage reactivation assays by van den Born et al. (2008) ........................................................................ 54-55

Figure 3.4: Molecular boundary prediction of viral AlkBs by principal component analysis and Shannon entropy scores ........................................................................................................ 58-59

Figure 3.5: Phylogeny and principal component analysis of viral AlkBs to derive potential origins ................................................................................................................................. 62-63

Figure 3.6: Viral AlkB predicted size and uniquely viral conserved residues .............. 65
# LIST OF ABBREVIATIONS

## General Terms

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>m1A</td>
<td>1-methyl-adenine</td>
</tr>
<tr>
<td>m3C</td>
<td>3-methylcytosine</td>
</tr>
<tr>
<td>AlkB</td>
<td>Alkylation B</td>
</tr>
<tr>
<td>atalkbh9b</td>
<td>Arabidopsis thaliana AlkB homolog 9B</td>
</tr>
<tr>
<td>DOPE</td>
<td>Discrete Optimization of Potential Energy</td>
</tr>
<tr>
<td>EcAlkB</td>
<td>E. coli AlkB</td>
</tr>
<tr>
<td>FTO</td>
<td>Fat mass and obesity-associated protein</td>
</tr>
<tr>
<td>GLD</td>
<td>Grapevine leafroll disease</td>
</tr>
<tr>
<td>HEL</td>
<td>Helicase</td>
</tr>
<tr>
<td>ALKBH[1-8]</td>
<td>Human AlkB Homolog [1-8]</td>
</tr>
<tr>
<td>MTR</td>
<td>Methyltransferase</td>
</tr>
<tr>
<td>MSA</td>
<td>Multiple sequence analysis</td>
</tr>
<tr>
<td>PCA</td>
<td>Principal Component Analysis</td>
</tr>
<tr>
<td>RdRp</td>
<td>RNA-dependent RNA polymerase</td>
</tr>
<tr>
<td>RMSD</td>
<td>Root mean square deviation</td>
</tr>
<tr>
<td>RW</td>
<td>Rugose Wood</td>
</tr>
<tr>
<td>(+)ssRNA</td>
<td>Single-stranded positive-sense RNA</td>
</tr>
<tr>
<td>vAlkB</td>
<td>Viral AlkB</td>
</tr>
<tr>
<td>Virus</td>
<td>Description</td>
</tr>
<tr>
<td>---------------</td>
<td>--------------------------------------------------</td>
</tr>
<tr>
<td>AMV</td>
<td>Alfalfa mosaic virus</td>
</tr>
<tr>
<td>BVY</td>
<td>Blackberry virus Y</td>
</tr>
<tr>
<td>BlScV</td>
<td>Blueberry scorch Virus</td>
</tr>
<tr>
<td>CMV</td>
<td>Cucumber mosaic virus</td>
</tr>
<tr>
<td>GLRaV-1</td>
<td>Grapevine leafroll-associated virus 1</td>
</tr>
<tr>
<td>GLRaV-2</td>
<td>Grapevine leafroll-associated virus 2</td>
</tr>
<tr>
<td>GLRaV-3</td>
<td>Grapevine leafroll-associated virus 3</td>
</tr>
<tr>
<td>GRSPaV</td>
<td>Grapevine rupestris stem pitting-associated virus</td>
</tr>
<tr>
<td>GVA</td>
<td>Grapevine virus A</td>
</tr>
<tr>
<td>GVB</td>
<td>Grapevine virus B</td>
</tr>
<tr>
<td>LChV2</td>
<td>Little cherry virus 2</td>
</tr>
<tr>
<td>SVX</td>
<td>Shallot virus X</td>
</tr>
</tbody>
</table>
1 Introduction

Alkylation B (AlkB) homologs are diverse and ubiquitous proteins that can be found in virtually all forms of life (Bratlie and Drabløs, 2005; Van den Born et al., 2008). Their ubiquity speaks to their universal importance. Surprisingly, AlkB homologs have even been identified in a small subset of positive-sense single-stranded RNA viruses (Bratlie and Drabløs, 2005); this observation is surprising given the limited coding capacity of these small RNA viruses. It is additionally striking that the majority of these AlkB containing viruses affect economically significant woody perennial plants. In this thesis, I focus primarily on grapevines and their viruses due to the economic and scientific importance in Canada and globally.

1.1 Grapevine history

Grapes have been a quintessential part of human history and culture for as many as 8,000 years, and has been a significant driver of economy, culture, and research ever since (This et al., 2006; Naidu et al., 2014). The cultivation of *Vitis vinifera*, the most common commercial grapevine, is believed to have originated somewhere between the Black Sea region and Iran (This et al., 2006). Subsequently, the spread of grapevines across Europe was heavily influenced by the Roman Empire which brought considerable development of viticulture and viniculture between the first and the end of the second century AD (Maliogka et al., 2015; Terral et al., 2010). From the fourth century onwards, while the Christian faith spread its influence throughout Europe,
viticulture and viniculture maintained its geographical expansion, eventually extending its reach worldwide (Terral et al., 2010).

Today, the grapevine industry continues to flourish globally. There are approximately eight million hectares of vineyard worldwide which produce an estimated 75 million tons of grapes. The majority of the production goes towards wine production (55%), while the remaining 45% is used for fresh consumption as table grapes, dried raisins, or is processed into non-alcoholic juices (FAO, 2016). All of this has cemented grapes and grapevines among the most important global crops.

1.2 Canadian grapevine industry and its economic impact

The Canadian grapevine industry is a significant driver to the economy. For example, the 2016 harvest alone produced a crop of 70,851 tonnes of grapes with an estimated raw value of at least $95.3 million (Grape Growers of Ontario, 2017). However, the economic impact of grapevines extends well beyond the value of grapes produced alone; as such the total economic impact of grapevines in Ontario was an estimated $9.08 billion in 2015 (Rimerman and Eyler, 2017), a $2.2 billion increase from the previous survey conducted in 2011 (Rimerman and Eyler, 2017). The Ontario wine industry shows no signs of decreasing its economic advancements, as Ontario has consistently been the most productive province in the nation with 180 wineries that produce about 71% of total Canadian wine volume (Grape Growers of Ontario, 2017).

However, Ontario is not the only province in Canada expanding its wine industry. Overall, the Canadian wine industry has grown substantially from 476 wineries in 2011, to 604 in 2015 – a 27% increase (Rimerman and Eyler, 2017). These wineries were
serviced by approximately 1,770 grape growers operating in Canada with a combined acreage of 31,100 grape-bearing acres (Rimerman and Eyler, 2017).

As climate change allows grapevines to be grown increasingly further north, we are already seeing with the success of vineyards in Prince Edward County, ON, and other regions further north than Niagara. It is clear that the Canadian grapevine industry will continue to thrive and even grow over the coming years. However, climate change has also caused several significant changes in viticultural practices. These challenges are further exacerbated by the growing spread of pathogens. Consequently, it is evident that it is economically beneficial for Canada to invest in our growing grape and wine industry to better understand the looming threats that are already presenting themselves.

1.3 AlkB

AlkB homologs are a diverse and ubiquitous protein family represented in most cellular organisms with the exception of Archaea (Bratlie and Drabløs, 2005). Its involvement in reversing methylation damage in DNA and RNA plays a critical role in the adaptive response, thus protecting an organism's genome against deleterious effects resulting from alkylation by both internal and environmental chemical agents (Fedele et al., 2015). Specifically, *Escherichia coli* AlkB (EcAlkB) uses molecular oxygen to oxidize the alkyl groups of damaged nucleotides, such as 1-methyl-adenine (m1A) and 3-methylcytosine (m3C), releasing the oxidized alkyl groups as aldehydes, thus regenerating the undamaged base (Falnes and Rognes, 2003) (Figure 1.1). AlkB's
role is such an essential part of cellular life that mutations to AlkB often result in a nonviable organism. For example, it was observed that the lack of a functional AlkB resulted in sensitivity to SN2-type alkylating agents, an increase in spontaneous mutations, and activation of the SOS response in bacteria (Kataoka et al., 1983; Sikora et al., 2010).

Figure 1.1: Examples of repair mechanisms of AlkB on methylated bases (Fedeles et al., 2015).
The gene *alkB* was first associated with alkylation as early as 1983 when an *E. coli* strain lacking the *alkB* gene was found to be modestly sensitive to SN2 class alkylating substances (Kataoka et al., 1983). Following this observation, the research group cloned the *alkB* gene and isolated the recombinant AlkB protein from *E. coli* (Kataoka and Sekiguchi, 1985). However, following this, research largely stagnated due to the limited knowledge of the AlkB protein.

It was not until a decade later that AlkB was classified as a repair enzyme when it was observed that its overexpression in human cell lines corresponded to a moderately increased resistance to alkylation (Chen et al., 1994). However, very little was still understood of the protein. It was not until two decades later that a breakthrough came when AlkB was classified as a member of the 2-oxoglutarate (2OG; also known as alpha-ketoglutarate) and Fe(II)-dependent dioxygenase superfamily using a bioinformatics approach (Aravind and Koonin, 2001). Shortly thereafter, two independent research groups reconstituted AlkB-mediated DNA repair *in vitro*, which experimentally verified the classification of AlkB as a member of the 2OG-Fe(II) oxygenase superfamily (Falnes et al., 2002; Trewick et al., 2002).

AlkB proteins show a great diversity of substrate specificity including DNA (Jennifer A. Calvo et al., 2012; Cetica et al., 2009; Chen et al., 2010; Dango et al., 2011; Duncan et al., 2002; P. Li et al., 2013; Nay et al., 2012; Sundheim et al., 2006; Zhu and Yi, 2014), RNA (Chandola et al., 2015; D. Fu et al., 2010; Y. Fu et al., 2010; Jia et al., 2011; Muller et al., 2012; Müller et al., 2013; Nordstrand et al., 2010; Ougland et al., 2012; Pan et al., 2008; Pastore et al., 2012; Silvestrov et al., 2012; Van Den Born et al., 2011; Westbye et al., 2008a; Zdzalik et al., 2014), and even proteins (Bjørnstad et al.,...
Its function appears to be so ubiquitous to life that AlkB homologs are represented in most organisms, with the exception of Archaea (Figure 1.2) (Bratlie and Drabløs, 2005). Surprisingly, AlkB homologs are even found in a small subset of single-stranded positive-sense RNA (ssRNA(+)) viruses (Bratlie and Drabløs, 2005; Van den Born et al., 2008).
Figure 1.2: The overall tree of the AlkB family. For clarity, individual viruses are grouped and major groups of bacteria and eukaryotes are collapsed and labelled with the name of the group (van den Born et al., 2008).

1.3.1 AlkB in bacteria

Overstating the ubiquity of AlkB proteins is difficult given that AlkB homologues are found in most bacteria, almost all eukaryotes, and, most surprisingly, even viruses. Of all homologs, the EcAlkB protein is undeniably the best-studied enzyme of the family.
AlkB type proteins are widespread among the kingdom Bacteria - which tend to have one, or at most two, AlkB homologues (Bratlie and Drabløs, 2005). The majority of aerobic bacterial species express AlkB proteins in some form. However, given AlkB’s requirement for molecular oxygen, obligate anaerobic bacteria do not appear to have AlkB type proteins (Sedgwick et al., 2007; Van den Born et al., 2009).

Bioinformatic analysis of bacterial AlkB homologues revealed four phylogenetic groups denoted 1A, 1B, 2A, and 2B (Van den Born et al., 2009). Group 1A proteins are characterized by robust oxidative dealkylation activity and a broad substrate specificity (Van den Born et al., 2009). This group includes the prototypical EcAlkB, but the group is widespread amongst bacteria, with all others in the group sharing very high (>75%) sequence identity to EcAlkB (Fedeles et al., 2015). Group 1A can notably repair both methylated and ethano lesions. Group 1B proteins share a similar repair activity to vertebrate AlkB homologues ALKBH2 and ALKBH3 and are characterized by their wide substrate preference (van den Born et al., 2009). Group 1B are found primarily in the β- and γ-subdivisions of proteobacteria (van den Born et al., 2009).

Found in the α-subdivision of proteobacteria (e.g. Agrobacterium, and Rickettsia), group 2A proteins share a strong homology to the vertebrate protein ALKH8: a protein implicated in the post-transcriptional modification of tRNAs (Van den Born et al., 2009). Group 2B AlkB proteins are most likely to show substrate specialization. It has been proposed that this specialization is the result of adaptation to specific environmental stressors. They are most commonly found in actinobacteria but are also found in some root-associated plant pathogens such as Xanthomonas (γ-proteobacteria) and Burkholderia (β-proteobacteria) (Van den Born et al., 2009).
Phylogenetic analysis of bacterial AlkB sequences demonstrated that bacterial AlkB clustering patterns are significantly different from phylogenies generated using the 16S rRNA sequences. This apparent disconnect indicates a high degree of horizontal gene transfer between bacterial species. The patchy distribution of AlkB sequences is mimicked in the distribution of viral AlkBs (Van den Born et al., 2009). When viral AlkB phylogenies are compared to more highly conserved genes in viruses [such as RNA-dependent RNA polymerase (RdRp)] significantly different clustering patterns are observed (Van den Born et al., 2009). Similar to horizontal gene transfer in bacteria, it is likely that AlkBs have undergone a high degree of movement between viral species.

1.3.2 Mammalian AlkB

Whereas bacteria tend to have one, or at most two, AlkB homologues, multicellular eukaryotes tend to have several, with the notable exception of *Saccharomyces cerevisiae* which has none (Bratlie and Drabløs, 2005). Mammals share nine homologs denoted ALKBH1-8 and FTO (Aravind and Koonin, 2001; Duncan et al., 2002; Yi and He, 2013) which differ in substrate specificity and subcellular localization (Bratlie and Drabløs, 2005).

Among the nine human homologues only two can complement the DNA repair activity of EcAlkB *in vivo* (Aas et al., 2003; Duncan et al., 2002); ALKBH2 and ALKBH3 have been shown to protect cells against alkylation damage both in cell culture and in animals (Aas et al., 2003; Jennifer A Calvo et al., 2012; Fu and Samson, 2012). The other seven homologues have adapted to fill different biological niches.
ALKBH1, with the highest sequence identity to EcAlkB, functions as a mitochondrial nucleic acid demethylase (Westbye et al., 2008b) with apyrimidinic/apurinic lyase activity. However it acts not only on DNA, but it can also demethylate the protein histone H2A (Ougland et al., 2012).

The remaining homologues primarily demethylate proteins or RNA. They have an extensive range of functions that are described briefly here.

Both FTO and ALKBH5 primarily repair N6-methyladenosine (m6A) in RNA (Church et al., 2010; Liu et al., 2013; Shen et al., 2015; Zhao et al., 2014; Zheng et al., 2013), however FTO has been shown to also repair 3-methylthymine (m3T) and 3-methyluracil (m3U) in ssDNA (Jia et al., 2008). ALKBH8 is involved in the maturation of tRNA by modifying its wobble position (Fu et al., 2010; Fu et al., 2010). The action of ALKBH4 appears to be actin demethylation (Li et al., 2013). It is believed that ALKBH7, which plays a role in alkylation-induced necrosis (Fu et al., 2013; Solberg et al., 2013; Wang et al., 2014a), acts on protein substrates, but the identity of these substrates is not currently known. Finally, neither the function or substrates of ALKBH6 have been identified (Fedeles et al., 2015).

It is believed that ALKBH5 and FTO are the most recently diverged AlkB proteins because they can be found only in vertebrates (Falnes et al., 2007; Robbens et al., 2008), whereas the other seven human AlkB proteins are conserved across all members of the kingdom Animalia (Drabløs et al., 2004).
1.4 Sequence conservation and structural features of cellular AlkB proteins

The crystal structures of EcAlkB (C. Yang et al., 2008; Yi et al., 2010; Yu et al., 2006a), ALKBH2 (C. Yang et al., 2008), ALKBH3 (Sundheim et al., 2006), ALKBH5 (Feng et al., 2014), ALKBH7 (Wang et al., 2014b), ALKBH8 (Pastore et al., 2012), and FTO (Han et al., 2010) have provided significant insight into the molecular mechanism of AlkB.

The EcAlkB protein is comprised of 216 amino acids with a molecular mass of 23.9 kDa. Its active site is contained within a β-helix consisting of eight β-strands arranged in pairs in a helical conformation (Fedeles et al., 2015). This helical structure is commonly referred to as a "jelly-roll" fold (Figure 1.3A). Using conserved residues His-131, Asp-133, and His-187 (Figure 1.3B) the jelly-roll fold binds Fe(II) and αKG cofactors (C.-G. Yang et al., 2008; Yu et al., 2006b).
**Figure 1.3**: Structural aspects of *E. coli* AlkB protein. A, crystal structure (Protein Data Bank (PDB) 3O1P) of *E. coli* AlkB, featuring the jelly-roll fold, with the m1A substrate flipped out of the DNA duplex into the active site of the enzyme. The orange sphere denotes the position of the central metal ion (here Mn(II)). B, active site of *E. coli* AlkB showing the octahedral coordination around the central metal ion and the relative position of the substrates m1A (Fedeles et al., 2015).

EcAlkB further contains a "nucleotide recognition lid" comprised of a unique 90 amino acid long N-terminal subdomain that interacts with the damaged base and covers the active site (Zdzalik et al., 2014). The recognition lid was hypothesized to be conformationally flexible, which can accommodate the broad substrate specificity of AlkB homologues. However, recent evidence suggests that the entire recognition lid is not solely responsible for substrate specificity, but rather only a few key residues (Holland and Hollis, 2010; Zdzalik et al., 2014).
The active site of AlkB contains a 3 Å-wide oxygen diffusion tunnel from the protein surface to the oxygen binding site which allows access to the conserved domains required for base repair (Yu et al., 2006b). The octahedral conformation of the active site is conserved even when Fe(II) is replaced with Co(II) or Mn(II). However, these substitutions result in the closure of the oxygen diffusion tunnel (Yu and Hunt, 2009).

It has been observed that the three-dimensional conformation of many of the oxygenases is highly conserved; however, only a few key amino acid residues are similarly conserved (Bratlie and Drabløs, 2005). It appears that only an HxD, and a RxxxxxxR motif are universally conserved in functional AlkBs (Bratlie and Drabløs, 2005). However, the rest of the amino acid sequence shows relatively low levels of conservation, with an average sequence similarity less than 40%, and only 10% that is universally conserved (Bratlie and Drabløs, 2005).

1.5 An overview of the current state of knowledge of viral AlkB domains

Grapevines are among the most virus infected crops today, with over 80 different viruses identified (Martelli, 2017). The majority of these viruses are associated with one of five major diseases in grapevines: infectious degeneration and decline, leafroll, rugose wood, fleck, and red blotch disease (Martelli, 2017). Surprisingly, the main etiological agents for many of these important diseases contain an AlkB domain within their replicase (poly)protein. It is widely believed that these diseases are caused by mixed infections. The variety of symptoms such as malformations of leaves and twigs,
foliar discolorations (reddening, yellowing, chlorotic or bright yellow mottling, ringspots, and line patterns), grooving and/or pitting of the woody cylinder, delayed bud break, stunting, and decline, in addition to the nature of mixed infections, has made direct causation between any given virus and a specific disease difficult; however several viruses have been associated with their respective disease (Martelli, 2017). These difficulties have made virus management historically difficult.

Viral infection can result in significant losses to a vineyard by affecting vine quality and the quantity of grapes produced by any given vine. Here I focus on two of the most economically devastating diseases in grapevines: Grapevine leafroll disease (GLD), and Rugose wood complex (RW), as they are relevant to my thesis research area.

Recently, it has been discovered, through a bioinformatics approach, that an AlkB domain is encoded as part of the replicase (poly)protein by a small subset of (+)ssRNA viruses. Remarkably, all these viruses happen to infect perennial plants, including such economically important fruit crops as grapevine and citrus (Bratlie and Drabløs, 2005; Dolja et al., 2017; Martelli et al., 2007; Van den Born et al., 2008), and are the main etiological agents for diseases such as GLD, and RW. These AlkB-containing viruses belong to five viral families: Closteroviridae, Betaflexiviridae, Alphaflexiviridae, Secoviridae, and Potyviridae with the latter each containing only a single AlkB-containing member at this point. Here I provide an overview of the three major AlkB-containing families.
1.5.1 Closteroviridae

The family *Closteroviridae* currently has four identified genera: *Closterovirus*, *Ampelovirus*, *Crinivirus*, and *Velarivirus* (Martelli, 2017), of which only members of the genus *Ampelovirus* contain an AlkB homolog. Members of this genus include Grapevine leafroll associated viruses, which are the main etiological agent of Grapevine leafroll disease (GLD). GLD is widely recognized as the most complex and economically important viral disease across grapevine-growing regions worldwide (Naidu, 2017; Naidu et al., 2014). This disease alone is responsible for an estimated $25,000 – $41,000 per hectare in economic losses in a vineyard over a 25-year lifespan (Atallah et al., 2012). Due to its economic and scientific significance it is difficult to overstate the importance of GLD and its associated viruses. Here I provide a brief overview of its symptoms, discovery, and characteristics.

GLD is a graft-transmissible disease common to grapevines worldwide. Its symptoms are best observed in late summer and fall where on red-skinned varieties of *Vitis vinifera*, leaf tissue between the veins turns deep red to purple, with downward curling or cupping of the leaf margins. On white varieties, the leaf tissue will turn yellow with curling or cupping of the leaf margins (Figure 1.4). The disease and its associated symptoms decrease the photosynthetic capabilities of the grapevine therefore decreasing vine, and any subsequently produced wine's, quality (Martelli, 2017).
GLD was first identified in Germany in 1936 (Scheu, 1936), however it would be nearly 40 years until an etiological agent was identified. A decade after its identification in Germany (Harmon and Snyder, 1946), identified a graft-transmissible disease in California, which after another decade, was identified as GLD (Goheen et al. 1958). Despite the similar symptoms to those caused by Citrus tristeza virus (CTV; genus Closterovirus, family Closteroviridae), the cause of GLD remained unclear. It was not until 1979 where a Japanese research group associated GLD with Closterovirus-like-particles that the symptoms were linked to the virus (Namba et al. 1979).

Initially, several serologically different leafroll-associated closteroviruses were identified, however it would not be until 1995 that the nomenclatural convention of Grapevine leafroll-associated virus (GLRaV) followed by an Arabic numeral (e.g., GLRaV-1, GLRaV-2, GLRaV-3) would be established (Boscia et al. 1995).
Following observations that Grapevine virus A (GVA) could be spread by the mealybug *Pseudococcus longispinus* (Rosciglione et al. 1983), a subsequent study identified that Grapevine leafroll-associated virus 3 (GLRaV-3) could be spread by *Planococcus ficus* (Rosciglione and Gugerli 1989). However, we now know that Grapevine leafroll-associated viruses can be spread by a multitude of mealybugs and scale insects (Krüger et al. 2006; Almeida et al. 2013). A full list of currently accepted vectors is published in Grapevine viruses, Molecular Biology, diagnostics and management, Chapter 2 (Martelli, 2017).

Members of the family *Closteroviridae* have genomes that range from 13,700 to 18,500 nucleotides, comprising between six and twelve genes (Martelli, 2017) (Figure 1.5). Open reading frames (ORF) 1A and 1B comprise the replicase polyprotein which universally contains three domains required for viral replication: the methyltransferase (MTR), helicase (HEL), and polymerase (POL). Within the family *Closteroviridae*, the AlkB domain is located within the replicase polyprotein, ORF1A, between the methyltransferase (MTR) and helicase (HEL) domains, a feature common to viral AlkBs.

**Grapevine leafroll-associated virus 3 (*Closteroviridae*)**

![Genome map of Grapevine leafroll associated virus 3](image)

**Figure 1.5:** Genome map of Grapevine leafroll associated virus 3. The replicase polyprotein is highlighted in blue, and the AlkB domain is highlighted in red. Other components of the replicase polyprotein include the papain-like protease (P-Pro),
methyltransferase (MTR), helicase (HEL), and RNA-dependent RNA polymerase (POL). Downstream, are two open reading frames encoding the coat protein (CP), and divergent coat protein (CPd).

1.5.2 Betaflexiviridae

The family Betaflexiviridae currently contains the majority of AlkB containing viruses. Nearly all of the genera within the family have AlkB-containing viruses (Martelli, 2017). The family has many disease causing viruses, including several believed to be the main etiological agents for Rugose wood (RW). RW is a complex disease which often presents itself in a mixed infection, which makes identifying its economic significance particularly difficult. Although it is impossible to estimate an exact figure in dollars for the damage caused by RW it is known that it is an economically and scientifically significant viral disease due to its global distribution patterns and its associated damage.

RW is another complex of graft-transmissible diseases spread via mealybugs (Martelli, 2017). The disease was first identified in Italy, however it was quickly identified in Hungary as well (Martelli, 1965, Martelli et al., 1967). RW is a complex disorder in which four separate symptoms have been identified: Rupestris stem pitting, Kober stem grooving, Corky bark, and LN-33 stem grooving (Martelli, 2017). The complex nature of this disease made the establishment of etiology very difficult. Its supposed viral nature was supported by mechanical isolation from a symptomatic vine of a virus with particles resembling those of closteroviruses (Conti et al. 1980). This virus would eventually be named Grapevine virus A (GVA) (Milne et al. 1984). Other similar viruses, all belonging
to the genus *Vitivirus* (family: *Betaflexiviridae*), were subsequently identified and associated with RW. An additional virus, Grapevine rupestris stem pitting-associated virus (GRSPaV) (Meng et al., 1998) was also associated with RW, but classified into a novel genus: *Foveavirus* (Martelli and Jelkmann 1998).

Vitiviruses and foveaviruses bear a similar morphology to that of closteroviruses, sharing a long, very flexuous filamentous particle. Their genomes, however, are considerably shorter with sizes of only approximately 7,300 – 8,700 nucleotides. Additionally, similar to *Closteroviridae*, members of the family *Betaflexiviridae* encode their AlkB as a portion of the replicase polyprotein between the MTR and HEL domains (Figure 1.6). There is however currently one identified exception, Grapevine virus E (GVE). Belonging to the genus *Vitivirus*, it has been identified that there is an encoded AlkB embedded within the HEL domain (Alabi et al., 2013) (Figure 1.7). This is surprising given that other closely related vitiviruses such as Grapevine virus A, Grapevine virus B, and Grapevine virus F each contain an AlkB, which is in the expected location (i.e. between the MTR and HEL domains). It is currently unclear as to why AlkB in GVE is embedded within the HEL domain, or if it is unique in this trait.

**Grapevine Rupestris stem pitting-associated virus** (*Betaflexiviridae*)

![Figure 1.6: Genome map of GRSPaV. Its replicase polyprotein is highlighted in blue, and the AlkB domain is highlighted in red. Other components of the replicase](image)
polyprotein include the papain-like protease (P-Pro), methyltransferase (MTR), helicase (HEL), RNA-dependent RNA polymerase (POL), hypervariable region (HVR), and ovarian tumor-like (OTU). Downstream is the open reading frame encoding the coat protein (CP).

The family Betaflexiviridae currently contains the majority of AlkB-containing viruses and contains members responsible for economically significant diseases in plants. It is undoubtable that understanding the vAlkB domain is important for understanding the viral family.

**Figure 1.7**: Genome map of Grapevine virus E. Note the inclusion of the AlkB domain within the helicase (Hel) domain of the replicase polyprotein. Other domains of the replicase polyprotein include the methyltransferase (MTR), helicase (HEL), and RNA-dependent RNA-polymerase (RdRp). Downstream the movement protein (MP), coat protein (CP), and a nucleic acid-binding protein (NB) are encoded as separate open reading frames.
1.5.3  *Alphaflexiviridae*

*Alphaflexiviridae* is a family of viruses in the order *Tymovirales*. Plants and fungi serve as natural hosts for members of this family. There are currently 51 species in this family, divided among 6 genera (Nemchinov et al., 2017). Diseases associated with this family include: mosaic and ringspot symptoms. The family *Alphaflexiviridae* are similar in shape and size to that of *Betaflexiviridae* with genomes between approximately 5,400 – 9,000 nucleotides in length encoding between one and six proteins (Nemchinov et al., 2017). Similar to the other viral families, the AlkB domain is encoded as part of the replicase protein, however their replicase does not consist of a polyprotein, but is a singular protein. The AlkB domain is found between the MTR, and HEL domains in all known cases for members of the family *Alphaflexiviridae*.

1.5.4  Proposed functions of vAlkB

Several research groups (Bratlie and Drabløs, 2005; Dolja et al., 2017; Martelli et al., 2007; Van den Born et al., 2008) have hypothesized potential functions for the viral domain, however the exact nature of the domain remains unknown despite its economic and scientific significance.

One such hypothesis (Bratlie and Drabløs, 2005) is that the vAlkB domain was included into these viruses to counteract a yet undiscovered host antiviral defense system, or to enhance viral propagation. This has been proposed due to the fact that AlkB containing viruses commonly infect woody perennial plants, and therefore there is potentially a conserved host antiviral defense system associating with these woody
perennials. However, this is merely a supposition as no detailed study of this has been published.

Viruses are in a constant arms race against their host, constantly struggling to outcompete the host just to replicate and multiply. Throughout history, plants have evolved post-transcriptional gene silencing (PTGS) to defend against viruses and transposable genetic elements (Waterhouse et al., 2001). In order to successfully infect plants, many viruses have developed strategies to counteract the PTGS (Roth et al., 2004). One research group favours the hypothesis that AlkB is one of the mechanisms evolved to counter the PTGS in plants (Brattie and Drabløs, 2005). This hypothesis is in some doubt mainly due to the recent integration of AlkB into viral genomes (van den Born et al., 2008). Given that the PTGS is an ancient system in plants it is likely that an AlkB domain was acquired on a similarly ancient timescale to defend against the new defense mechanism in plants. However, since we observe a more recent acquisition of vAlkB this is an unlikely scenario.

Given the recent integration of AlkB into viral genomes it has been suggested that AlkB was acquired as a response to pesticide-induced stress like in plants. Several common pesticides such as fonofos, parathion, and terbufo (Zhang et al., 2012) cause methylation in both DNA and RNA (Zayed and Mahdi, 1987; Starratt and Bond, 1988; Wiaderkiewicz et al., 1986; Braun et al., 1982). Thus, the treatment of any plant with these methylating agents would create a highly methylating environment. It is clear that an integrated methylation damage repair domain would offer a competitive advantage to a virus in such an environment. By creating such an environment, it is likely that bacteria infecting a plant would begin to overexpress AlkB which would provide easy
access of AlkB to viruses for integration (Bratlie and Drabløs, 2005). This hypothesis is also considered unlikely because most AlkB-containing viruses do not infect plants treated with such pesticides. There has also likely not been enough time since the introduction of such pesticides to have observed the phylogenetic divergence that is seen in viral AlkBs from their bacterial ancestors (van den Born et al., 2008).

Many of the viruses with an AlkB domain infect woody perennials where they can survive for years in the phloem tissue until their eventual transmission to new plants (van den Born et al., 2008). While residing in the phloem for such prolonged time periods viruses are often subjected to considerable methylation damage (van den Born et al., 2008). Once again, an integrated methylation damage repair domain would confer a considerable advantage to the virus in such an environment (Bratlie and Drabløs, 2005). This hypothesis assumes that viral AlkBs are under positive selection in certain woody perennial plants, and is currently the favoured hypothesis for the presence of AlkB in certain ssRNA plant infecting viruses.

1.5.5 Function of AlkB in viruses remains largely unknown and understudied despite a single paper

The following section addresses the analyses conducted by van den Born et al., (2008), the only study that was conducted on the biological function of viral AlkB domains.

Previously, a research group applied *in vitro* analyses of the viral AlkB domain to identify and demonstrate that AlkB is a functional protein domain, and not just an artifact identified through bioinformatic analyses and to further identify whether it shared a
similar function to cellular AlkB homologs. Specifically, these researchers targeted five viruses: Grapevine virus A (GVA; family Betaflexiviridae), Blueberry scorch virus (BlScV; family Betaflexiviridae), Citrus leaf blotch virus (CLBV; family Betaflexiviridae), Little cherry virus 2 (LChV-2; family Closteroviridae), and Blackberry virus Y (BVY, family Potyviridae). They observed that all of these viruses shared greater similarity to bacterial AlkB homologs than to mammalian AlkB homologs, which could be indicative of a bacterial origin. However, limited sequence identity makes it impossible to draw a direct origin from sequence data alone.

The researchers then expressed vAlkB proteins of three different lengths in order to study the potential sizes of the viral AlkB that exhibited full function. They established an AlkB “core-domain”, as defined by the shortest functional viral AlkB identified at the time, which they designated as AlkB with an extension. This “core-domain” is hereto referred to as the “core-domain” as defined by Van den Born (2008). The researchers then arbitrarily selected lengths of 36, and 94 amino acids to extend the “core-domain” as defined by Van den Born (2008) (Figure 1.8) to observe whether the viral AlkB domain required segments, or the entire, nucleotide recognition lid to function.
Figure 1.8: Viral AlkB proteins used in the van den Born et al. (2008) study. Arrows indicate the borders of the AlkB-coding regions included in the various expression constructs. The solid line indicates the ‘AlkB core’ as defined by the research group, i.e. the region which displays homology to other members of the AlkB family, and extensive sequence homology within the subfamily of viral AlkB proteins. Black shading indicates complete conservation, whereas grey shading indicates similarity (i.e. similar residues).

Their *in vitro* analyses were comprised primarily of a phage reactivation assay in which the researchers treated DNA bacteriophage M13mp18 and the RNA bacteriophage MS2 with methyl methanesulphonate (MMS) in order to significantly methylate the genome and subsequently attempted to recover phage infectivity by
repairing the phages with viral AlkB proteins of their three lengths defined above. They additionally attempted to restore the methylated phage genome using EcAlkB to act as a control as they knew its function and were certain it would act on DNA and RNA. Phage survival was scored by counting plaques.

Ultimately, they selected only three viruses for these phage reactivation trials: GVA, BVY, and BiScV. They observed that in cases an extension beyond their “core domain” alone resulted in a more active viral protein (Figure 1.9) and concluded that N-terminal to the AlkB core appeared to be required for viral AlkB activity.
Figure 1.9: Reactivation of MMS-treated bacteriophages by expression of AlkB proteins in *E. coli*. (A) Reactivation of methylated ssRNA phage MS2 by AlkB proteins from GVA, BVY, BIScV, LChV-2 or CLBV. Reactivation of methylated MS2 by different variants of AlkB from (B) GVA, (C) BIScV and (D) BVY. (E) Reactivation of methylated ssDNA phage M13 by GVA-36, BVY-36 and BIScV-94. (F) Reactivation of methylated MS2 by BVY-94 and mutants BVY-94-H59A and BVY-94-D61A. Expression plasmid pJB658 without insert was used as control. Error bars in B–D represent the standard deviation of triplicate measurements (Van den Born et al., 2008).

These results suggested that, similarly to their cellular homologs, viral AlkBs likely represent nucleic acid repair proteins. Since viral AlkBs are encoded by RNA viruses, RNA repair is the most obvious candidate function for these proteins. These results were corroborated by the fact that EcAlkB, a known DNA preferring AlkB, was able to
recover DNA phage M13 exposed to 40mM MMS by 6–7 orders of magnitude, whereas
the viral counterparts were only able to recover the same phage with increase of only 3–
4 orders of magnitude. This decreased affinity of viral AlkB towards DNA indicates that
DNA is not the preferred substrate of vAlkB.

The 36 amino acid extension plus the “core-domain” was observed to be the
most functional vAlkB in two of the three viruses tested: BVY, and GVA (Figure 1.9),
however in BIScV the 94 amino acid extension plus the “core domain” narrowly
outcompeted the 36 amino acid extension (Figure 1.9). So despite their best efforts with
these three viruses a functional boundary of the viral AlkB domain cannot be
established. However, it is evident that the domain requires more than just its “core” to
functional, and vAlkB likely contains, in some capacity, a nucleotide recognition lid.

1.5.6 Proposed origin and evolutionary history of the viral AlkB domain

The exact origins of the viral AlkB domain are still a mystery, however several
research groups have made attempts to position the acquisition of the viral AlkB domain
through sequence analyses.

Firstly, following the identification of AlkB in viruses the next step was to
determine whether this is an ancient viral domain, or whether it was acquired more
recently from a cellular organism. The relative alignment scores were calculated for the
domains of the replicase (poly)protein required for replication [i.e. the methyltransferase
(MT), RNA-dependent RNA-polymerase (RdRp), and Helicase (HEL) domains] and
AlkB (Bratlie and Drabløs, 2005). If these domains share a high degree of coevolution
then a trend line of ~1.0 is observed. This exactly was observed for the MT, RdRp, and
HEL domains all with high coefficients of correlation (Figure 1.10). However, the same trend is not observed when these three domains are compared to AlkB. It was observed that a coefficient of correlation between 0.10 and 0.16 exists between AlkB and the other three domains indicating little to no correlation and therefore no coevolution between AlkB and any of the three domains. These results suggest that the viral AlkB did not evolve alongside these other domains and is therefore likely a relatively recent acquisition (Bratlie and Drabløs, 2005).

**Figure 1.10:** Pairwise distances of a multiple sequence alignment between sequence regions corresponding to methyltransferase (MT), RdRp (A) and AlkB domains (B). Each data point corresponds to e.g. RP-RP and MT-MT distances for the same pair of sequences, and sequences showing similar evolutionary distance in these two regions will fall on the diagonal. The pairwise distances were estimated from multiple alignments using the Blosum50 score matrix. (C) Correlation coefficients calculated for each
pairwise alignment of the replicase (poly)protein domains required for viral replication plus AlkB. (Bratlie and Drabløs, 2005)

Following these observations, several research groups have attempted to develop an evolutionary timeline, or evolutionary scenario includes its origin which resulted in viruses acquiring an AlkB domain. Given that vAlkBs bear a greater sequence similarity to bacterial AlkB sequences than they do to vertebrate AlkBs (Bratlie and Drabløs, 2005; Van den Born et al., 2009) it is likely that the viral AlkB domain was initially acquired from a bacterial origin. Additionally, viral AlkB homologs are most significantly found in current day members of the family Betaflexiviridae, which has lead researchers to believe that AlkB was acquired by an ancestor of the family Betaflexiviridae simply based on its distribution patterns (Figure 1.11) (Martelli et al., 2007). However, the authors note that this suggestion is weak and requires additional verifications.
Figure 1.11: Evolutionary scenario for the families *Flexiviridae* and *Tymoviridae*. Boxes represent open reading frames. Replicase gene domains are as in Figure 3. M, methyltransferase; A, AlkB; O, OTu-like peptidase; P, papain-like protease; H, RNA helicase; R, RNA-dependent RNA polymerase. (Martelli et al., 2007). Family *Flexiviridae* is now extinct as its members are classified into *Alphaflexiviridae* and *Betaflexivirae*. 
1.5.7 Recent evidence suggests that plant infecting viruses can hijack their hosts' AlkB

Recently, it has been suggested that some RNA viruses that lack an AlkB domain are capable of hijacking host AlkB homologs to promote long-term genome stability. Such strategies have been observed in human viruses such as Human immunodeficiency virus-1 (HIV-1) (Lichinchi et al., 2016; Tirumuru et al., 2016), Hepatitis C virus (HCV) (Lichinchi et al., 2016) and Zika virus (Gokhale et al., 2016), and plant viruses such as Alfalfa mosaic virus (family: Bromoviridae) (Martínez-Pérez et al., 2017) and Tobacco mosaic virus (family: Virgaviridae) (Li et al., 2018).

Interestingly, the two human viruses currently associated with AlkB homologs, HIV-1 and HVC, are both causative agents of chronic long-term infections. As such, it stands to reason that some way to protect their genome from methylation damage – such as m6A methylation, the most common methylation in eukaryotes (Lichinchi et al., 2016) – would confer an obvious evolutionary advantage. It has in fact been demonstrated that in both HIV, and HPC, m6A negatively impacted viral activity in their respective cell lines (Lichinchi et al., 2016; Tirumuru et al., 2016). Their results suggested an important role of m6A in viral protein synthesis. m6A is notably reversed in humans by two RNA acting AlkB homologs: ALKBH5, and FTO.

Research conducted in Arabidopsis thaliana aimed to study whether the Arabidopsis AlkB homolog 9B (atALKBH9b) would aid in the survivability of two viruses: Alfalfa mosaic virus (AMV), and Cucumber mosaic virus (CMV). The researchers observed that deletion of the atalkbh9b gene nearly removed all activity of AMV,
observing faint banding both in Northern and Western blot; this indicates a decrease in both viral mRNA levels and virus capsid protein (Martínez-Pérez et al., 2017). This indicates that the virus itself could somehow be utilizing a cellular AlkB protein in order to promote its survivability through the removal of harmful m6A methylation damage.

The researchers observed no change in the activity of CMV (Martínez-Pérez et al., 2017). It is possible that this AlkB hijacking is not utilized by all viruses. However, it is more likely that, given the many AlkB homologs in Arabidopsis, CMV is simply targeting a different host AlkB to carry out demethylation. Additional research is required to observe whether viruses can hijack any AlkB homolog, or whether they are limited to a small subset of RNA preferring AlkB homologs.

1.5.8 Conclusion of biological studies of the vAlkB domain

AlkB is a ubiquitous protein family seemingly essential to virtually all forms of life. Its presence in a small subset of (+)ssRNA viruses was a surprise given the limited coding capacity of these viruses. Its unique inclusion within three economically and scientifically significant viral families merits justifies my studies.

1.6 Methodologies used to study the novel vAlkB

Given that the vAlkB domain is so poorly studied, I thought to begin the study with computational biology in order to develop a stronger hypothesis for downstream research. By applying numerous standard and novel in silico techniques, I hope to develop a clearer picture of this novel viral domain more quickly than it would be possible through molecular techniques alone. I do this by applying standard
bioinformatic approaches such as multiple sequence alignments (MSA), and homology modelling, and expand these findings with less common techniques such as principal component analysis and the Shannon Diversity Index.

1.6.1 Multiple sequence analysis

Multiple sequence alignments (MSA) are a technique commonly used in biology in which three or more DNA, RNA, or amino acid sequences are arranged to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences (Thompson et al., 1994).

There are several publicly available software options which implement different strategies to create these MSAs, one such popular option is Clustal Omega (Ω). Clustal Ω implements five main steps to generate a MSA (Larkin et al., 2007). Firstly, a pairwise alignment is generated using the k-tuple method, an efficient method to find so called ‘words’. Following this, the mBed method calculates the pairwise distance between input sequences and clusters them accordingly. This step is followed by the k-means clustering method (Larkin et al., 2007). Next, a guide tree is constructed using the unweighted pair group method (UPGMA) (Larkin et al., 2007). This is done in an iterative fashion where, at each step, the nearest two clusters are combined until the final tree has been calculated. In the final step, the multiple sequence alignment is produced using HHAlign package from the HH-Suite, which uses two profile hidden Markov-models (Larkin et al., 2007).
1.6.2 Homology modeling

Homology modeling is a technique that refers to the construction of an atomic-resolution model of a “target” protein from its primary sequence and an experimentally derived three-dimensional structure of a related homologous protein (Baker and Sali, 2001). There are several publicly available bioinformatic software platforms which specialize in homology utilizing one or more templates such as I-TASSER (Yang and Zhang, 2015), and MODELLER (Šali and Blundell, 1993).

This technique is a relatively new and exceptionally powerful tool to molecular biologists that has been improving drastically since its inception. Recent evidence suggests that, when using sequences with greater than 50% identity, models tend to be reliable, with only minor errors in side chain packing and rotameric state, and an overall root mean-square deviation (RMSD), the measure of the average distance between the atoms, between the modeled and the experimental structure falling around 1 Å (Baker and Sali, 2001). An error margin comparable to the typical resolution of a structure solved by NMR (Baker and Sali, 2001). As the technology continues to develop predictions are only becoming more and more accurate. Although not a replacement for experimentally derived structural models, homology modelling can provide significant research potential in emerging areas, such as viral AlkB.

1.6.3 Principal component analysis

Principal component analysis (PCA) is a technique used to emphasize variation and bring out strong patterns in a dataset. It is often used to make data easy to explore and visualize. Fundamentally the process by converting a set of observations of
possibly correlated variables into a set of values of linearly uncorrelated variables called principal components (Wang and Kennedy, 2014). This transformation is defined in such a way that the first principal component has the largest possible variance, and each subsequent component after that has the highest variance possible under the constraint that it is orthogonal to lower components.

In large protein families, such as the Fe(II)-dioxygenase alphaketoglutarate family, it can be challenging to determine the relative importance of amino acids within the protein. PCA is suited to such a task since the problem is based in a large variable space, i.e. the number of amino acids that make up the protein sequence (Wang and Kennedy, 2014). Since PCA is powerful at reducing the dimensionality of complex problems by projecting the data into an eigen-space that represents the directions of greatest variation it is ideal to address such problems of amino acid importance and relatedness.

1.6.4 Shannon diversity index

Diversity indices are statistics used to summarize the diversity of a population in which each member belongs to a unique group. One such statistic is the Shannon Diversity Index. The index is most popular in the field of ecology, however it was initially proposed by Claude Shannon to quantify the entropy (uncertainty or information content) in strings of text (Shannon, 1948). As such, it is a highly useful tool in order to measure the entropy in any string of text, such as MSAs.

The fundamental principal is that the Shannon index can identify how many different characters there are in a string of text, and the more unique characters in a
string, the more difficult it will be to correctly predict which character will be the next (Shannon, 1948). The Shannon entropy quantifies the uncertainty associated with each prediction at each position. It is commonly represented as the following formula:

$$H' = \sum_{i=1}^{R} p_i \ln p_i$$

where $p_i$ is the proportion of characters belonging to the $i^{th}$ type of character in the string of interest and $R$ is the length of the string (Shannon, 1948). In using this formula, I can derive an entropy score in which if all characters for a given position are identical, the entropy is 0.00, and if all characters are unique the value is 1.00.

By implementing this string analysis statistic to MSAs it is possible to quantify regions of similarity in any MSA based on unique strings such as DNA, RNA, amino acid, or any unique MSAs.

1.7 Significance

Alkylation B (AlkB) proteins are ubiquitous among cellular life where they act to reverse methylation damage to nucleic acids. This is such an essential function to life that AlkB homologs are observed in virtually all cellular life except Archaea and anaerobic bacteria. Surprisingly, a novel AlkB domain was discovered recently, through bioinformatics, in a small number of positive-sense, single-stranded RNA viruses. Interestingly, many of these viruses are pathogens of economically important fruit crops such as grapevine, stone fruits and citrus. Given the limited coding capacity of RNA
viruses, it is most likely that these viral AlkB domains play an important role in the replication and infection cycle of these viruses.

Unfortunately, little is known about the biological relevance of these viral AlkB domains barring a single publication (van den Born et al., 2008). In this study, the researchers demonstrated the demethylation activity of AlkB from five viruses using reactivation assays of methylated phages. As such, many fundamental questions about viral AlkB-domain remain unanswered, including its molecular boundary (due to the fact that AlkB is part of the replicase (poly)protein), its function in viruses, and the origin and evolutionary history.

1.8 Hypotheses and objectives

Given the limited coding capacity of these (+)ssRNA viruses it is likely that the viral AlkB is a functional protein domain. However, given the inconsistent in vitro results for the determination of the size of the viral AlkB domain the current annotation of the molecular boundaries for the viral AlkB domain is lacking.

I therefore hypothesize that the viral AlkB domain performs a similar role to that of cellular AlkB homologs and therefore must adopt a similar conformational shape. From this conformational shape I hypothesize we can propose a more accurate demarcation for the size of vAlkB.

I additionally hypothesize that the vAlkB domain was acquired a single time to an ancestor of a viral family and subsequently passed on through horizontal gene transfer
Therefore my objects are as followed:

(1) Generate three dimensional structures of many viral AlkB domains representing all AlkB-containing virus families through homology modeling using DNA and RNA AlkB homologs as a template to minimize potential sources of error

(2) Perform statistical analyses on the 3D structures and the primary sequences which they represent to propose a generalized molecular boundary for the viral AlkB domain

(3) Analyze the literature and phylogenetics of the viral AlkB domain to propose potential functions

(4) Study the conservation of the viral AlkB domain in order to identify motifs unique to viral AlkB domains to aid in the identification of viral AlkB domains in the future.
2 MATERIALS AND METHODS

2.1 Sequence preparations

The viral AlkB amino acid sequences of were obtained from the UniProt online database (http://www.uniprot.org/). Viral AlkB sequences were identified through homology searches using the EcAlkB catalytic core domain as a query while ensuring the entirety of the HxD and RxxxxxR motifs were included. In many cases, the query results detect only the catalytic core domain, or the “core domain” as defined by Van den Born (2008) as there is no clear demarcation of the exact size of vAlkBs. As it remains unclear where the upstream highly variable region begins it is difficult to determine a size from the N-terminus of the peptide. Instead, I opted to utilize the much more conserved C-terminus as it forms part of the catalytic core.

The sequence length of viral AlkB domains within the replicase (poly)protein were therefore determined by identifying the RxxxxxR motif, and taking the 200 amino acid residues upstream from the C-terminal R in the motif. Multiple sequence alignments were performed using Clustal Ω (Goujon et al., 2010; Larkin et al., 2007) in the R environment. Some manual corrections were made in order to ensure the alignment of both the HxD and RxxxxxR motifs.
2.2 Protein modelling

Protein models were generated using MODELLER (version 9.18) (Šali and Blundell, 1993) for the AlkB domain of 37 AlkB containing viruses. Models were generated using DNA- and RNA-preferring AlkB protein structures resolved through X-ray diffraction as homologs to minimize any possible substrate bias. Specifically, EcAlkB (PDBid: 2FDJ) (Yu et al., 2006b) – a DNA preferring AlkB, and ALKBH5 (PDBid: 4NJ4) – an RNA preferring AlkB (Feng et al., 2014) – were used in tandem to develop the viral AlkB homology models. These templates were selected for their substrate preference, and overall similarity in order to develop a predicted model without substrate bias as evidence by Van den Born (2008) suggested that vAlkBs can act on both DNA and RNA.

MODELLER was run locally with the default settings for a multi-template run. Viral sequences were input as a MSA to DNA and RNA AlkB homologs, and the atomic coordinates of the DNA and RNA templates were submitted to a Python simple script file from which the models were developed. Quality of the models was assessed in terms of Z-scores using the WHAT_IF program (Vriend, 1990). Z-scores are standardized statistically-derived structure quality assessment scales that include packing quality, Ramachandran plot appearance, chi-1/chi-2 rotamer normality, and backbone conformation. The structures with the highest Z-scores were used for subsequent analysis. Finally, the discrete optimization of potential energy (DOPE) scores were calculated using a python script included in MODELLER (version 9.18)
(Šali and Blundell, 1993). The DOPE scores were aligned to a multiple sequence alignment through a script written in R (version 3.3.3).

### 2.3 Principal component analysis

PCA was performed using the bio3d package (version 2.3) (Grant et al., 2006; Skjærven et al., 2016, 2014) in R (version 3.3.3). The 37 previously generated models were selected as an input which underwent iterative rounds of structural superposition to determine the HxD, and RxxxxxR motifs of the invariant catalytic core of the protein, which coincided with the canonical core domain. This was performed in order to provide the closest possible superposition to minimize variance based on alignment superpositions alone.

The structures were subsequently superimposed onto this core and PCA was conducted. In this process, a covariance matrix from the coordinates of the superimposed structures was diagonalized. The eigenvectors - a scaling factor used in PCA including direction - of this matrix represent the principal components of the system, and the eigenvalues, the raw values of the eigenvectors, are a measure of the variance within the distribution along the respective eigenvectors. All generated structures were then projected onto a two-dimensional plot from which principal components one through four were visualized in R.

In this process, the diagonalized covariance matrix of the superimposed structures can be used to determine similar groups through their principal components. In other
words, the eigenvectors of this diagonalized matrix represent the principal components of the system: the parts of the structure in which there are the most variability. By establishing the much better conserved catalytic core domain in previous steps I provide a backbone upon which similarities are established allowing more prominent differences to be identified in the principal components.

2.4 Shannon diversity index

Amino acid sequences of 200 residues of viral AlkB domains were characterized by their physiochemical properties and split into one of five groups: non-polar aliphatic R groups, polar uncharged R groups, positively charged R groups, negatively charged R groups, and non-polar aromatic R groups. The properties of this simplified amino acid alphabet were subsequently aligned in R (version 3.3.3). The Shannon diversity index (Ramazzotti et al., 2004) was calculated for each position using the simplified and standard amino acid alphabet and plotted in R (version 3.3.3). The average Shannon entropy score was calculated per 25 residues to observe the trend of physiochemical properties throughout the alignment of 200 residues.

By composing alphabets of 6 and 21 respectively (five physiochemical properties and a gap character; and 20 common amino acids and a gap character) I can implement the string comparison technique outlined by Shannon (1948). The entropy was calculated using the Shannon entropy equation, and inputting the count of the relative frequency of each amino acid, or amino acid physiochemical property as \( p_i \). From this, the relative diversity was plotted in using R.
Phylogenies were constructed using RAxML (version 8.2) (Stamatakis, 2014) with the AlkB sequences of all 37 viruses used in homology modeling. All sequences were ~150 residues and identified by counting 150 amino acids upstream of the C-terminal R in the RxxxxxR motif. A size of ~150 was selected based on the predicted size identified through PCA, and Shannon-Diversity indices. 1,000 iterations of maximum likelihood trees were developed, and 1,000 bootstrap replicates of the optimal (i.e. “most likely”) tree were calculated in RAxML, and corroborated with a neighbour-joining tree generated in PAUP* (version 4.0a161) (Swofford, 2011) with a further 1,000 bootstrap replicates. All trees were visualized using Dendroscope (version 3.5.8) (Huson and Scornavacca, 2012).

Multiple phylogenetic analyses were conducted in order to minimize bias of any given algorithm. Whereas neighbour-joining techniques are fast they lack robust iterative correction techniques, and should not be used alone for strong results. In order to make up for this, maximum likelihood was selected as a means of verification as its multi-step procedure allowed for a more statistically robust test of the viral AlkB domain relationships. The process does, however, take significantly longer to perform as it must perform iterative corrections/modifications to existing trees as many times as specified by the user. 1,000 iterations were selected for these trees. After these trees were generated the single tree with the greatest (i.e. maximum) likelihood was selected and an additional 1,000 rounds of bootstrapping were conducted on it to assess the quality of the tree.
3 RESULTS

3.1 AlkB-containing viruses are unique in that they are mostly pathogens of perennial plants and are phloem-limited

The recent identification of an AlkB domain in a small subset of single-stranded RNA viruses is intriguing. However, very little is understood about these viral AlkB domains. To search for clues on the biological relevance of vAlkBs in these RNA viruses, I first investigated the common features of AlkB containing viruses and their hosts in a hope to find clues on the biological function of the AlkB domains encoded by these viruses. By analyzing peer-reviewed publications, Descriptions of Plant Viruses (Adams and Antoniw, 2006), and the most up to date report of the International Committee on Taxonomy of Viruses (King et al., 2018), I aimed to identify commonalities among AlkB-containing viruses and the hosts they infect. Several commonalities were observed. First, for the majority of these AlkB-containing viruses, the AlkB domain is present as part of the replicase (poly)protein between the conserved MTR and the helicase domains. The only exceptions are three viruses of the genus *Vitivirus*, namely GVE, GVG, and GVI, in which the AlkB is inserted within the helicase domain (Blouin et al., 2018; Nakaune et al., 2008). Second, I found that the hosts of these viruses that contain intact (and thus predicted to be functional) AlkBs exclusively infect perennial plants (Table 1). The only viruses that do not infect a perennial plant were three viruses that infect potato: Potato virus M, Potato virus P, and Potato virus S; however these viruses notably contain a non-functional AlkB as suggested previously by van den Born et al. (2008). Third, I also found that a vast majority of these viruses
are limited to the phloem tissue. Of the 43 viruses with both predicted functional, and non-functional AlkBs, the majority (29) are known to be phloem-limited (Table 1). It is yet to be determined if the remaining 14 viruses are also phloem-limited because no data is available on the tissue tropism of these viruses.
Table 1: General overview of AlkB-containing viruses and some of their hosts’ characteristics.

<table>
<thead>
<tr>
<th>Family</th>
<th>Sub-family</th>
<th>Genus</th>
<th>Species</th>
<th>Perennial plant host</th>
<th>Woody plant host</th>
<th>Phloem limited</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alphaflexiviridae</td>
<td>Unassigned</td>
<td>Allexivirus</td>
<td>Shallot virus X</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mandarivirus</td>
<td>Indian citrus ringspot virus</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Potexivirus</td>
<td>Alstroemeria virus X</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Alphonanthera mosaic virus</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Cactus virus X</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Clover yellow mosaic virus</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Narcissus mosaic virus</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Opuntia virus X</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Papaya mosaic virus</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Zygocactus virus X</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Aconitum latent virus</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td>American hop latent virus</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Blueberry scorch virus</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Chrysanthemum virus B</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Coleus vein necrosis virus</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Daphne virus S</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hippeastrum latent virus</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lily symptomless virus</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Narcissus common latent virus</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Phlox Virus B</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Phlox virus S</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Poplar mosaic virus</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Potato virus M</td>
<td>Annual</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Potato Virus P</td>
<td>Annual</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Potato virus S</td>
<td>Annual</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sweet potato chlorotic fleck virus</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Betaflexiviridae</td>
<td>Quivirinae</td>
<td>Carlavirinae</td>
<td>Aconiton latent virus</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>American hop latent virus</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Blueberry scorch virus</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Chrysanthemum virus B</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Coleus vein necrosis virus</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Daphne virus S</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Hippeastrum latent virus</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Lily symptomless virus</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Narcissus common latent virus</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Phlox Virus B</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Phlox virus S</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Poplar mosaic virus</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Potato virus M</td>
<td>Annual</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Potato Virus P</td>
<td>Annual</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Potato virus S</td>
<td>Annual</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Sweet potato chlorotic fleck virus</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Citrivirus</td>
<td>Citrus leaf blight virus</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Foveavirus</td>
<td>Apple stem pitting virus</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rubivirus</td>
<td>Rupiyear stem pitting-associated virus</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Robivirus</td>
<td>Cherry green ring mottle virus</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Trichovirus</td>
<td>Cherry necrotic mottle virus</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Apricot pseudo-chlorotic leaf spot virus</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Cherry mottle leaf virus</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Vitivirus</td>
<td>Grapevine virus A</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Grapevine virus B</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Grapevine virus E</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Grapevine virus F</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Picornoviridae</td>
<td>Comovirinae</td>
<td>Unassigned</td>
<td>Black raspberry necrosis virus</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Closteroviridae</td>
<td>Unassigned</td>
<td>Ampelovirus</td>
<td>Little cherry virus 2</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Grapevine leafroll-associated virus 1</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Grapevine leafroll-associated virus 3</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Grapevine leafroll-associated virus 4</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Potyviridae</td>
<td>Unassigned</td>
<td>Brambyvirus</td>
<td>Blackberry virus Y</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
Previously, it had been predicted that an isolate of Grapevine virus B (GVB; GenBank accession number NP-619654) contained a non-functional AlkB which contains a histidine to asparagine mutation in the AlkB domain (Figure 3.1). This seems rather unusual because grapevine is a woody perennial, and I expect a functional AlkB in GVB. In order to explain why this GVB isolate has a non-functional AlkB, I examined its NCBI entry and discovered that this particular GVB isolate had been passaged through *Nicotiana occidentalis*, a herbaceous annual plant commonly used as a laboratory host plant in plant virology. Based on this information, I suspected that the altered AlkB may have been an isolated case unique only to this single isolate. To find out if such a mutation occurs only in this isolate and not in any other isolates, I retrieved the sequences of all GVB isolates for which the complete or near complete genomes have been sequenced. I found that this is indeed the only isolate that contains such a mutation (Figure 3.1). This, in conjunction with the observations in the three potato viruses, indicates that viruses negatively select against an included AlkB-domain when transitioned from a perennial to a non-perennial host.
**Figure 3.1:** Multiple sequence alignment of Grapevine virus B isolates. Conserved core HxD and RxxxxxR motifs are highlighted. A mutation in a single GVB isolate from Italy, NP_619654.1 in the HxD motif (green '!') has led GVB to have a predicted non-function AlkB. Both isolates were however passaged in *Nicotiana occidentalis*. All other isolates, derived from *Vitis* species, do not share the same mutation and rather contain both essential HxD, and RxxxxxR core motifs resulting in a predicted functional viral AlkB domain.

In summary, this association of vAlkBs with phloem-limited (+)ssRNA viruses that infect perennial plants is striking, and may suggest some critical function of AlkB domains in the replication and long term survival of these viruses.
3.2 Multiple bioinformatic analyses suggest a functional viral AlkB of between 150 and 170 amino acids in size

To expand on the very limited knowledge of viral AlkBs derived from the only published work by van den Born et al. (2008) on a small number of viruses, I aimed to further analyze viral AlkB domains through homology modelling, and expand the breadth of the viruses studied by applying *in silico* techniques to gather a more representative scale for all AlkB-containing viruses (Figure 3.2).
**Figure 3.2:** Multiple sequence alignment of viral AlkBs from members of each viral family with AlkB-containing members: *Alphaflexiviridae, Betaflexiviridae, Closteroviridae, Secoviridae,* and *Potyviridae,* with representative members of DNA-preferring EcAlkB, and RNA-preferring ALKBH5. The boundaries of the core domain are denoted using the previous research from EcAlkB (Sedgwick et al., 2007). The two universally conserved motifs, HxD, H and RxxxxxR, are highlighted with green boxes. Boundaries selected for previous *in vitro* analyses (Van den Born et al., 2008) are marked with arrows.
To this end, I first aligned viral AlkB sequences to identify conserved amino acid motifs. As expected, sequence alignments of viral AlkB yield relatively low sequence identities of only 31%, with the majority of the diversity observed in the N-terminal nucleotide recognition lid. This low level of sequence conservation has made previous analyses challenging. Many sequence-driven bioinformatic approaches provide little insight due to the poor conservation in primary structure i.e. sequence. However, through homology modelling I have found that, despite their limited sequence identity, the 3D structure of viral AlkBs from these five viruses were closely related to one another. Through the structural superposition of viral AlkB models with those of E. coli AlkB and human homolog AlkBH5, I calculated the root-mean-square deviation (RMSD) as a criterion to reflect how similar the models are to one another. I subsequently calculated the RMSD as a means of determining how similar my predicted models were to experimentally verified models in 3D space. As a result, I obtained a mean RMSD of 1.2 Å (data not shown), indicating a very close relationship among all viral AlkBs that were analyzed (Baxevanis and Ouellette, 2005). Additionally, the viral AlkB models were closely related to EcAlkB with a mean RMSD of 1.4 Å as well as to ALKBH5 with a mean RMSD of 1.6 Å (Figure 3.3A; and data not shown).
**Figure 3.3:** Homology modelling results for one of the viruses tested through phage reactivation assays by van den Born et al. (2008): Grapevine virus A (GVA). (A) Homology models were generated for the core domain alone (orange), the core domain with 36 amino acid extension (blue), and the core domain with a 94 amino acid extension (green) to be consistent with the *in vivo* analysis conducted by van den Born et al. (2008). (B) The discrete optimization of potential energy (DOPE) scores were calculated for each model, and compared to that of EcAlkB 2FDJ: a structure solved by X-ray diffraction.

I then attempted to identify any inaccuracies in my homology models relative to a model resolved through X-ray diffraction (PDBid: 2FDJ) by calculating the discrete optimization of potential energy (DOPE) score per residue of the three viral AlkB amino acid extensions initially studied by van den Born et al. (2008): specifically, the viral AlkB “core domain” as defined by Van den Born et al. (2008) with 0, 36, 94 amino acid extensions. The mean sum of squared errors (MSSE) was calculated to assess how similar the DOPE scores per residue of the predicted structure of viral AlkBs were when compared to the DOPE score per residue of the *E. coli* AlkB experimental structure. This metric was selected to quantify how similar the energy profiles are between the *in silico* derived viral AlkB structural models and the X-ray diffraction derived EcAlkB. I performed this analysis for all five viruses previously tested *in vitro* in the study conducted by van den Born et al. (2008), and similar results were obtained. As an example, the results for the three versions of GVA AlkB protein with varying lengths of N-terminal extensions are shown in Figure 3.3B. I observed an MSSE of 0.321 for the
viral AlkB viral “core domain” as defined by Van den Born et al. (2008) alone (Figure 3.3B, top panel), 0.0153 for the AlkB “core domain” with the 36 amino acid extension (Figure 3.3B, middle panel), and 0.0421 for the AlkB “core domain” as defined by Van den Born et al. (2008) with the 94 amino acid extension (Figure 3.3B, bottom panel). Given that the viral AlkB core plus 36 amino acids had fairly similar length to EcAlkB when aligned, and had the lowest MSSE when comparing energy profiles (Figure 3.3) of the three sizes tested in vitro, the 36 amino-acid extension plus the “core domain” appeared to be most similar to EcAlkB. This was in line with the observation that in vitro, the AlkB-core with a 36 amino acid extension was consistently the most functional viral AlkB when tested in two of the five viruses: GVA and BVY. I observed that the majority of the differences in DOPE scores were in the N-terminal region (i.e. the nucleotide recognition lid), which was consistent with the expected results given the limited sequence identity in this region.

This approach of modelling vAlkBs of variable extensions to the viral “core domain” is time consuming and limited by current techniques. In order to expand beyond a small subset of AlkB-containing viruses I aimed to develop a generalized model for viruses by modelling a large representation of AlkB-containing viruses and applying statistical analyses to predict a functional size, rather than extending the viral AlkB “core domain” by an arbitrary value as used in the study by Van den Born et al. (2008). For this purpose, I first applied PCA to a deliberately oversized viral AlkB domain in an attempt to predict a molecular boundary. A size of 200 amino acids was selected for vAlkB because it was observed to be considerably larger than EcAlkB when homology modelling was applied. The results are shown in Figure 3.4A. It was
observed that in these models constructed based on 200 residues that the first 50 amino acids consistently exhibited high variations (Figure 3.4A). Results from the first principal component showed several local maxima which decreased in higher principal components. However, the first 50-60 residues consistently showed very high variability, suggesting low sequence conservation when comparing at the level of both sequence and structural information.

I further aimed to study the N-terminal extension to the “core domain” of viral AlkB based on the physiochemical characteristics of its amino acid residues, and generated a characteristics-derived multiple sequence alignment (MSA). This allows for the comparison of physiochemical and hence functional features rather than sequence alone. As similar amino acids will not significantly alter structure or function, by analyzing this MSA based on amino acid characteristics we can look for commonalities in a more broad sense. The Shannon-diversity index was calculated for each position along the MSA. I observed many gaps in the N-terminus in the MSA, which in turn produced a total aligned size of 273 positions. As shown in Figure 3.4B, the first 126 positions in this MSA averaged a Shannon entropy score of 2.81, whereas the last 147 positions had a significantly lower Shannon entropy score of 1.56. This indicates that the first 126 positions in this multiple sequence alignment share very few physiochemical properties, however the remaining 147 positions are much more similar.
Figure 3.4: Molecular boundary prediction of viral AlkBs. (A) First three principal components for 37 AlkB containing viruses. PCA analysis utilized the R package bio3d to combine structural and amino acid sequence data to identify which regions are the most variable in the sequence and structure. (B) Shannon entropy score per residue of a multiple sequence alignment of the AlkBs from these 37 viruses based on physiochemical and functional properties of the amino acid residues. The individual values are represented by the blue line, and the average value, calculated using a sliding window of 25 residues is represented in red.
In summary, findings from structure homology modeling, energy profiles as well as analysis of the physiochemical properties of amino acids agree with one another. I conclude that the functional vAlkB domains are between 150 and 170 amino acids in size, comprising a nucleotide recognition lid domain of 50-70 amino acids at its N-terminal region and a 100 amino acid core at the C-terminus.

3.3 The first viral AlkB was likely acquired by an ancestral virus of the family Alphaflexiviridae

The inclusion of the viral AlkB domain into a small subset of (+)ssRNA viruses presents a puzzling question on their evolutionary history, which has been difficult to elucidate due to the limited sequence identity within the entire AlkB family. Given the prevalence of AlkB within the viral family Betaflexiviridae, it has previously been suggested that the viral AlkB domain was likely acquired, through a single acquisition event, by an ancestor of the family Betaflexiviridae (Martelli et al., 2007). However, this hypothesis suffers from the lack of supporting evidence. Here I aimed to use my homology models and PCA to test this hypothesis and in hope to provide evidence via bioinformatic analyses.

When I first generated a phylogeny of the viral AlkB-domains alone I observed an indistinct clustering pattern of viral families where members of a viral family were distributed across different clusters (Figure 3.5A). Whereas with more conserved domains involved in genome replication and transcription, such as the RNA-dependent
RNA-polymerase (RdRp), I observed fairly discrete clustering based on viral families (Figure 3.5B). This non-discrete clustering patterns of viral AlkBs are consistent with high degrees of horizontal gene transfer, similar to what is observed in bacterial AlkBs (Van den Born et al., 2009). The supposition of horizontal gene transfer is further observable in atypical genomic positions of vAlkBs in several viruses including Grapevine virus E (Alabi et al., 2013), Grapevine virus G (Blouin et al., 2018), and Grapevine virus I (Blouin et al., 2018).
**Figure 3.5:** Comparison of the evolutionary relationship of viral AlkB domains. (A) Phylogeny comparing the ~150 amino acid viral AlkB domains of viruses of different viral families. (B) Phylogeny of the viral RNA-dependent RNA-polymerase domain of the same AlkB-containing viruses. (C) Principal component analysis of 37 AlkB viruses comparing structural and sequence relationships of viral AlkB domains to one another to assess their similarity and predict evolutionary origins.

In order to minimize the effect of poor sequence conservation, I grouped the vAlkB-domains through PCA (Figure 3.5C). I observed weak groups forming in the first principal component which disappeared in higher principal components in favour of a single homogenous group (Figure 3.5C). Three groups were identified via K-means clustering of which two are the result of the first principal component. One group was considerably larger and composed primarily of members of the family *Alphaflexiviridae*. These results seem to be contrary to the earlier proposal that the first vAlkB was acquired by an ancestor of current day family *Betaflexiviridae*. Rather, these results communicate a single acquisition event as was similarly suggested by Martelli et al. (2007). But the relative abundance of members of the family *Alphaflexiviridae* within the larger ancestral cluster has prompted us to suggest a different scenario on the origin and subsequent evolution of vAlkBs: this larger and more closely grouped cluster (family *Alphaflexiviridae*) contains the ancestral virus which initially acquired AlkB due to its closer grouping pattern and size.
3.4 Identification of additional conserved amino acid residues in viral AlkB proteins

I further aimed to identify additional conserved residues that are unique to viruses. These virus-specific motifs may aid in the identification of future vAlkB domains. As new viruses are continuously being discovered, the need for a reliable identification of functional AlkBs will remain significant for an exceptionally long time. Firstly, I observed that the universally conserved HxD motif could be expanded to HxDDE, as I consistently observe two negatively charged residues immediately following this established HxD motif in vAlkBs (Figure 3.6). I additionally observe greater conservation in the RxxxxxR motif in viral AlkBs and observe a universally conserved RxSxTxR motif in which each variable amino acid (i.e. 'x') is a hydrophobic residue. Although the function of these additional residues in viruses remains unclear, and it is unlikely that they are required for function; their universal conservation could aid in the rapid identification of vAlkB domains in the future.
Figure 3.6: Multiple sequence alignment of the functional viral AlkBs as predicted in this study with viral families identified. Conserved residues are denoted in a red highlight, whereas similar residues are highlighted in yellow. Proposed conserved viral HxDDE and RxSxTxR motifs are highlighted in green boxes. Nucleotide recognition lid, and the catalytic core domain are indicated with boxes, and previously identified essential residues are highlighted with red stars all from previous analyses conducted in EcAlkB (Sedgwick et al., 2007).
4 DISCUSSION

4.1 Homology modeling and principal component analysis suggest functional vAlkBs to comprise 150-170 amino acids.

Recently, a novel AlkB domain was identified through a bioinformatic approach in a small subset of single-stranded RNA viruses (Aravind and Koonin, 2001; Dolja et al., 2017; Drabløs et al., 2004; Sabanadzovic and Martelli, 2017). Given the limited coding capacity of RNA viruses, it is unlikely that such a significant portion of its genome would be functionally useless. It is therefore likely that the vAlkB is functional, and potentially has the same demethylating activity as its cellular homologs. Increasing evidence has suggested that m6A modifications serve as an important mechanism to regulate RNA-based functions through the use of demethylating enzymes such as AlkB homologs. However, little is known of the structure and function of viral AlkBs. In this work, I apply homology modelling and principal component analysis to characterize and predict the viral AlkB domain's functional size and propose a model for its acquisition and subsequent evolution.

A major challenge with the characterization of viral AlkBs is that they are encoded as part of the replicase (poly)protein. Previous work by van den Born et al. has attempted to address the molecular size of vAlkBs in vitro (Van den Born et al., 2008). However, while the C-terminal boundary is readily determined due to the identification of the highly conserved RxxxxR domain, they failed to define the N-terminal boundary of vAlkB. They identified what they define as the “core domain” of viral AlkB alone, of approximately 140 residues, as defined by the shortest discovered vAlkB at the time.
(Van den Born et al., 2008), hereafter I refer to this as the “core-domain”. It was observed that the “core domain” by itself was outcompeted on phage reactivation assays by N-terminal amino-acid extensions of 36 and 94 amino acids to the “core domain”, depending on the source virus. However, all viral AlkB sequences outperformed wild-type EcAlkB in reactivating RNA phage MS2 that was rendered non-infectious by methylation treatments (Van den Born et al., 2008). I applied homology modelling techniques to the same viruses that were treated by Van den Born et al. (2008) to predict and observe the effects of variable length N-terminal extensions. I observed very high structural conservation when the “core-domain”, as defined by Van den Born (2008), was used in the analysis, despite their limited sequence identities. This is likely due to an important function shared by cellular AlkB proteins and those in viruses: the reversal of methylation damage to ensure genome integrity (Fedele et al., 2015). Interestingly, the protein model of the “core domain”, plus a 36 amino acid extension, visually formed a similar conformational shape and followed a similar energy profile to that of an EcAlkB protein derived from X-ray diffraction (PDBid: 2FDJ) (Yu et al., 2006b). This was surprising because the AlkB “core domain” with a 36 amino acid extension has a considerably shorter sequence than that of EcAlkB.

As there is an apparent disconnect between levels of conservation in the sequence and structural features within the AlkB family, I applied principal component analysis (PCA) to capture both structural features and sequence motifs to identify an approximate size of functional vAlkBs. I deliberately overestimated the size of the viral AlkB domain beyond what was predicted based on homology models in an attempt to capture the highly variable region just upstream of the vAlkB domain. I observed that
the first 50 amino acid residues were highly variable in all principal components until they normalize to a relatively conserved region in the last 150-170 residues. These residues demarcate the boundaries of a functional vAlkB domain. My results are in general agreement with those from the phage reactivation assays by van den Born et al. (2008) To this end, I had chosen 200 residues for PCA for all viruses that contain an AlkB. I observed several local maxima within the catalytic-core domain, particularly in lower components. These are artifacts of the homology modelling approach, and our lab hopes to corroborate this predicted AlkB size through direct experimental tests such as phage reactivation and in vivo infectivity assays of varying AlkBs sizes to identify the most functional vAlkB length.

By utilizing the upstream highly variable region (HVR), I can use principal component analysis to identify which residues contribute the greatest dissimilarities when comparing all viral models. Internally the local maxima in the first principal component indicate regions of variability as expected by the inherent low sequence conservation of the vAlkB domain. However, in higher principal components, the effects of low conservation levels become less prominent because they contain less of the total variance. As such, I can identify the variability contributed by the HVR, rather than just observing the low sequence conservation.

I observed that extensions of the nucleotide recognition lid to include additional residues N-terminal to the identified vAlkB generated significant differences in homology modelling. The “core domain” plus a 94 amino acid extension adopted a significant tertiary structure that is absent in EcAlkB (Figure 3.3C). It is possible that this tertiary structure could explain the in vitro analysis results which identified a “core domain” plus
36 amino acid extension that exhibited the highest activity in two of the three assays performed (Van den Born et al., 2008). I suggest that extensions to the vAlkB “core domain” adopt a tertiary structure which interferes with nucleotide binding. As such, the domain is unable to efficiently bind ssRNA of phage MS2 which would correspond to the decreased activity in vitro (Van den Born et al., 2008). However, given that BiScV had its greatest activity with an extension of 94 amino acids to the “core domain” (Van den Born et al., 2008), it is reasonable to assume that the true size lies somewhere between these two arbitrarily selected values, i.e. 94 and 36 residue extensions plus the viral AlkB core domain. However, given the highly variable nature of this N-terminal region, it is unsuitable to model these interactions in silico to make concrete assertions to this hypothesis through in silico analysis alone.

Furthermore, it is unlikely that a clear N-terminal motif can be identified for vAlkBs without a better understanding of the proteolytic cleavage of the replicase polyproteins encoded by AlkB containing viruses. I propose that a functional molecular boundary can be established by identifying an amino acid position that is around 150-170 amino acids upstream of the final R in the RxxxxxR motif identified at the C-terminal boundary of AlkBs including viral AlkBs. In order to better understand this novel viral domain, it is important to first clearly establish the molecular boundaries of viral AlkB that is part of a polyprotein, for downstream analysis.
4.2 Results of *in silico* analysis point to the acquisition of AlkB by an ancestral virus of the family *Alphaflexiviridae*.

The majority of AlkB containing viruses identified to date belong to the family *Betaflexiviridae*, and based on this very reason, it has previously been suggested that an ancestor of the family *Betaflexiviridae* acquired the AlkB domain from a yet unknown cellular organism, and subsequently disseminated it to other viral families through horizontal gene transfer (Bratlie and Drabløs, 2005; Martelli et al., 2007). Since this observation was made on distribution patterns alone, I thought it is justified to provide a statistical analysis to group the viral AlkB domain in the hopes of predicting 1) either a single or multiple acquisition event(s) might have taken place for these viruses to acquire AlkB from cellular organisms and 2) if it was acquired only once, of which of the AlkB-containing viral families this ancestral virus was. To accomplish this, I utilized the grouping potential of PCA. The loss of distinct groupings in higher principal components (Figure 3.5C) indicates that vAlkBs are closely related when both structure and sequence are used in tandem. Had there been multiple acquisition events from distinct cellular organisms, I would expect multiple groups in all principal components segregated based on viral family. Due to this grouping pattern that was observed, our data supports the notion that vAlkBs were likely acquired through a single acquisition event and subsequently spread to other viral families through horizontal gene transfer (e.g. recombination) as proposed by others (Bratlie and Drabløs, 2005; Martelli et al., 2007; Van den Born et al., 2008).
At this point, we lack enough evidence to suggest an original family which acquired the first vAlkB. I did, however, observe some clustering in the first principal component. Since the first principal component explains the majority of the variation between vAlkBs (Figure 3.4C), this larger cluster contains the original ancestral vAlkB. Since all vAlkB families are represented within this large cluster, it suggests that all viral families with an AlkB domain are similar to this larger ‘ancestral’ cluster with minimal divergence resulting in the other groups observed in the first principal component. Because this larger cluster is composed primarily of members of the family *Alphaflexiviridae*, I marginally favour the new hypothesis that AlkB was first acquired by an ancestor of the family *Alphaflexiviridae*, followed by subsequent propagation to other viral families through horizontal gene transfer - likely during a mixed infection of the host. In fact, mixed infection by multiple and taxonomically distinct viruses are commonplace for perennial plants such as grapevines (Al Rwahnih et al., 2009; Coetzee et al., 2010; Martelli, 2017; Pantaleo et al., 2010). However, due to lack of corroborating evidence at this point we cannot conclusively state that the family *Alphaflexiviridae* acquired the first vAlkB but rather that the original vAlkB was likely acquired by an ancestor of *Alphaflexiviridae, Betaflexiviridae*, or even their common ancestor.

To my surprise, and contrary to my expectation, it was observed that the family *Betaflexiviridae* showed the greatest distribution in all principal components, which is indicative of greater diversity within the family. This increased diversity could be caused by a number of factors such as a switch in host leading to an evolutionary pressure resulting in many mutations to allow better adaptation to the new host. Alternatively, this
greater diversity could be explained by horizontal gene transfer. As viral AlkBs were transferred from the [hypothesized] ancestral Alphaflexiviridae via horizontal gene transfer there would likely be numerous mutations after incorporation into its new viral family.

4.3 Identification of expended conserved functional motifs in viral AlkBs.

Additionally, the greater conservation observed in viral AlkBs (31%) (Figure 3.6) relative to that of mammalian homologs (18%) (data not shown), is consistent the notion that viral AlkBs were acquired more recently and from a single acquisition event from a cellular organism. As such, it is possible that I could identify additional conserved motifs, or expand previously established motifs in cellular AlkBs, in viruses. As an example, I observed that the universally conserved HxD motif could be expanded to an HxDDE, as I consistently observe two negatively charged residues immediately following this established HxD motif (Figure 3.6). I also observed greater conservation in the RxxxxxR motif in viral AlkBs and a universally conserved RxSxTxR motif in which each variable amino acid (i.e. ‘x’) is a hydrophobic residue. I observed several relatively well-conserved residues in viral AlkBs which is involved in the coordination of the substrate, similar to observations made in E. coli that noted only a few residues were required to dictate substrate specificity (Holland and Hollis, 2010).
4.4 Viral AlkB is unique to RNA viruses that are pathogens of perennial plants with a tissue tropism to the phloem.

The acquisition of AlkB domain by members of several families of plant infecting RNA viruses is surprising because of the genomic nature of these viruses. All AlkB containing viruses have a single-stranded positive-sense RNA genome, which limits their coding capacity. As such, they must utilize as much of their genome as possible. The mere presence of this domain implies some kind of evolutionarily advantageous function. When I studied the nature of AlkB containing viruses, some striking trends arose. I observed that all AlkB containing viruses with canonical HxD, and RxxxxxR motifs exclusively infect perennial plant hosts. A predicted non-functional AlkB has been discovered in three viruses that infect potato, the only nonperennial plant which can be infected by an AlkB containing virus, albeit these AlkB domains appear to be inactive due to mutations at one of the critical residues such as mutations in the HxD or RxxxxxR domains. Due to its presence in perennial plant-infecting viruses, the AlkB domain may exert an important function in safeguarding the perpetual fitness of these viruses via genome demethylation to survive year after year in these perennial crops. Since cultivated potatoes came from wild potato plants that are perennial, I predict that the related viruses in wild potatoes will contain functional AlkB. However, since commercial potato cultivars are grown as annual plant crops, viruses infecting them no longer require a functional AlkB domain, hence the mutation of critical amino acid residues in these viruses. Similarly, in Grapevine virus B that naturally infects grapevine, a perennial plant, a non-functional vAlkB has previously been reported (Van den Born et al., 2008). I found this puzzling and studied it further. It was observed that
the source isolate for this seemingly non-functional vAlkB was passaged multiple times in *Nicotiana occidentalis*, an annual, herbaceous experimental host that supports the replication of GVB. This passaging through an annual herbaceous plant likely caused the observed mutations in its AlkB. Further investigation of AlkB sequence in eight GVB isolates for which complete genome sequences are available revealed intact AlkB domain with the conserved amino acid residues in all but this one isolate (Fig. 3.2).

Very recently, it has been suggested that a number of RNA viruses are capable of hijacking host AlkB homologs to promote long-term genome stability. Such strategies have been observed in human viruses such as Human immunodeficiency virus-1 (Lichinchi et al., 2016; Tirumuru et al., 2016), Hepatitis C virus (Lichinchi et al., 2016) and Zika virus (Gokhale et al., 2016), and plant viruses such as Alfalfa mosaic virus (family: *Bromoviridae*) (Martínez-Pérez et al., 2017) and Tobacco mosaic virus (family: *Virgaviridae*) (Li et al., 2018). However, I observe that the majority of AlkB containing viruses infect perennials specifically, and are mostly phloem-limited viruses. This curious finding has prompted me to propose that there is something unique of perennials. Whereas other viruses can hijack host AlkB homologs, levels of AlkB proteins encoded by the host may be too low or even absent within the phloem tissue of these perennial plants. As such, the virus is not able to efficiently hijack a host AlkB protein and acquired its own copy to promote survivability in these perennials.

In summary, I present an *in silico* based analysis using principal component analysis and homology modelling to better understand the novel viral AlkB domain recently identified in a small subset of (+)ssRNA viruses. I provide several lines of evidence to support the predicted functional boundary of approximately 150-170 amino
acid residues (Figure 3.6). I further propose that the original viral AlkB was likely acquired by an ancestor of the family *Alphaflexiviridae*, contrary to the earlier proposal by Martelli et al. (2007). The function of the vAlkB domain is likely to promote longevity of these viruses in perennial plants - perhaps to counteract a host defence mechanism to accommodate a special need that is unique to perennial plants. The AlkB domain is unique to only viruses of perennials; I hypothesize that this may have been due to an absence of host AlkB within the phloem tissue of these plants. Naturally, this rather bold supposition needs to be tested through experimentation. As a recently discovered novel domain in viruses, many questions remain to be answered. I hope that my *in silico* analyses would aid in the development of new research systems and approaches in order to advance our understanding of functions of vAlkB and related molecular mechanisms.

### 4.5 Conclusion

Here, I applied several bioinformatic approaches including principal component analysis, homology modeling, and traditional sequence-based phylogenetic analysis in an attempt to develop a clearer picture of this novel viral domain among nearly all RNA viruses known to contain AlkB. I reveal that all these viruses containing AlkB have perennial plants as hosts and that most of these viruses are restricted to the phloem tissue. I propose that this AlkB domain may exert a function in promoting long-term survival of these viruses in their perennial hosts. Furthermore, I predicted a size of 150-170 amino acid residues for a functional viral AlkB. I propose a new hypothesis that viral AlkB was originally acquired by an ancestor of the family *Alphaflexiviridae* through a
single acquisition event from a cellular organism. Finally, I have identified several amino acids within the viral domain conserved in viral AlkBs which will aid in the identification of AlkB-containing in the future as additional novel viruses are discovered.

I feel that the results reported in this thesis can serve as a foundation based on which further research can be conducted to better understand this novel domain in economically important RNA viruses. These findings will make significant contributions towards the advancement of these heavily understudied viruses that infect perennial plants in general and of viral AlkB domains in particular.
5 Works Cited


Alabi, O.J., Sudarsana, @bullet, @bullet, P., Sarver, K., Martin, R.R., Naidu, R.A., n.d. Complete genome sequence analysis of an American isolate of Grapevine virus E. doi:10.1007/s11262-012-0872-0


analysis of genes and proteins, 3rd Ed. ed. Wiley, Hoboken, N.J.


Cetica, V., Genitori, L., Giunti, L., Sanzo, M., Bernini, G., Massimino, M., Sardi, I., 2009. Pediatric brain tumors: Mutations of two dioxygenases (hABH2 and hABH3) that


by oxidative demethylation. Neuroscience 145, 1222–1232.

doi:10.1016/j.neuroscience.2006.11.018


Hill, P.W.S., Amouroux, R., Hajkova, P., 2014. DNA demethylation, Tet proteins and 5-


King, A.M.Q., Lefkowitz, E.J., Mushegian, A.R., Adams, M.J., Dutilh, B.E., Gorbalenya,


Lichinchi, G., Gao, S., Saletore, Y., Gonzalez, G.M., Bansal, V., Wang, Y., Mason, C.E.,
Rana, T.M., 2016. Dynamics of the human and viral m6A RNA methylomes during
HIV-1 infection of T cells. Nat. Microbiol. 1, 16011. doi:10.1038/nmicrobiol.2016.11

Liu, C., Mou, S., Cai, Y., Dixon, J., Ogden, C., Carroll, M., Kit, B., Flegal, K., Parsons,
T., Power, C., Logan, S., Summerbell, C., Whitaker, R., Wright, J., Pepe, M.,
Seidel, K., Dietz, W., Deckelbaum, R., Williams, C., Lakshman, R., Elks, C., Ong,
K., Zhao, J., Grant, S., Frayling, T., Timpson, N., Weedon, M., Zeggini, E., Freathy,
Altman, D., DerSimonian, R., Laird, N., Mantel, N., Haenszel, W., Beg, C.,
Mazumdar, M., Egger, M., Smith, G., Schneider, M., Minder, C., Cecil, J.,
Tavendale, R., Watt, P., Hetherington, M., Palmer, C., Hassanein, M., Lyon, H.,
Nguyen, T., Akylbekova, E., Waters, K., Scott, R., Bailey, M., Moran, C., Wilson, R.,
Fuku, N., Liu, G., Zhu, H., Dong, Y., Podolsky, R., Treiber, F., Elks, C., Loos, R.,
Sharp, S., Langenberg, C., Ring, S., Hinney, A., Nguyen, T., Scherag, A., Friedel,
S., Brönner, G., Müller, T., Hinney, A., Scherag, A., Nguyen, T., Schreiner, F.,
Zhang, Q., González, J., González-Carpio, M., Hernández-Sáez, R., Vargas, V.,
Hidalgo, G., Almén, M., Rask-Andersen, M., Jacobsson, J., Ameur, A., Kalnina, I.,
Ntalla, I., Panoutsopoulou, K., Vlachou, P., Southam, L., Rayner, N.W., Wardle, J.,
Carnell, S., Haworth, C., Farooqi, I., O’Rahilly, S., Jacobsson, J., Danielsson, P.,
J., Xi, B., Shen, Y., Zhang, M., Liu, X., Zhao, X., Mangge, H., Renner, W., Almer,
G., Weghuber, D., Möller, R., Dwivedi, O., Tabassum, R., Chauhan, G., Ghosh, S.,
Marwaha, R., Luczynski, W., Zalewski, G., Bossowski, A., Riffo, B., Asenjo, S.,
Sáez, K., Aguayo, C., Muñoz, I., Lauria, F., Siani, A., Bammann, K., Foraita, R.,
Huybrechts, I., Moleres, A., Ochoa, M., Rendo-Urteaga, T., Martínez-González, M.,
Julián, M.A.S., Albuquerque, D., Nóbrega, C., Manco, L., Dina, C., Meyre, D.,
Gallina, S., Durand, E., Körner, A., Meyre, D., Delplanque, J., Chèvre, J., Lecoeur,
C., Lobbens, S., Cauchi, S., Stutzmann, F., Cavalcanti-Proença, C., Durand, E.,
Pouta, A., Mejía-Benítez, A., Klünder-Klünder, M., Yengo, L., Meyre, D., Aradillas,
C., Okuda, M., Hinoda, Y., Okayama, N., Suehiro, Y., Shirabe, K., Scherag, A.,
Dina, C., Hinney, A., Vatin, V., Scherag, S., Grant, S., Li, M., Bradfield, J., Kim, C.,
Annaiah, K., Munafò, M., Flint, J., Church, C., Lee, S., Bagg, E., McTaggart, J.,
Deacon, R., Fischer, J., Koch, L., Emmerling, C., Vierkotten, J., Peters, T.,
Tanofsky-Kraff, M., Han, J., Anandalingam, K., Shomaker, L., Columbo, K.,
Timpson, N., Emmett, P., Frayling, T., Rogers, I., Hattersley, A., Tschritter, O.,
Preissl, H., Yokoyama, Y., Machicao, F., Häring, H., Biondi, B., Hakanen, M.,
Raitakari, O., Lehtimäki, T., Peltonen, N., Pahkala, K., Xi, B., Chandak, G., Shen,
Y., Wang, Q., Zhou, D., Goran, M., 2013. FTO Gene Variant and Risk of
Overweight and Obesity among Children and Adolescents: a Systematic Review

Maliogka, V.I., Olmos, A., Pappi, P.G., Lotos, L., Efthimiou, K., Grammatikaki, G.,
Candresse, T., Katis, N.I., Avgelis, A.D., 2015. A novel grapevine badnavirus is
doi:10.1016/j.virusres.2015.03.003

Martelli, G.P., 2017. Grapevine Viruses: Molecular Biology, Diagnostics and


Pan, Z., Sikandar, S., Witherspoon, M., Dizon, D., Nguyen, T., Benirschke, K., Wiley,


Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting,


Crystal structures of DNA/RNA repair enzymes AlkB and ABH2 bound to dsDNA.

Nature 452, 961–5. doi:10.1038/nature06889


Crystal


