DISCOVERY AND DIVERSITY STUDIES OF VIRUSES WITH NEXT GENERATION SEQUENCING

Denis Kutnjak

Doctoral Dissertation
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia

Supervisor: Prof. Maja Ravnikar, PhD, National Institute of Biology, Ljubljana, Slovenia and Jožef Stefan International Postgraduate School, Ljubljana, Slovenia

Evaluation Board:
Prof. Kristina Gruden, PhD, Chair, National Institute of Biology, Ljubljana, Slovenia and Jožef Stefan International Postgraduate School, Ljubljana, Slovenia
Prof. Jernej Jakše, PhD, Member, Biotechnical Faculty, University of Ljubljana, Slovenia
Prof. Neil Boonham, PhD, Member, Fera Science Ltd., York, United Kingdom

MEDNARODNA PODIPLOMSKA ŠOLA JOŽEFA STEFANA
JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL



Denis Kutnjak

DISCOVERY AND DIVERSITY STUDIES OF VIRUSES WITH NEXT GENERATION SEQUENCING

Doctoral Dissertation

ODKRIVANJE IN ŠTUDIJ RAZNOLIKOSTI VIRUSOV S SEKVENCIRANJEM NASLEDNJE GENERACIJE

Doktorska disertacija

Supervisor: Prof. Maja Ravnikar, PhD

Ljubljana, Slovenia, October 2016

Acknowledgments

This thesis would not exist without the help and collaborative work of many colleagues and friendly support of my friends and family as well as support of funding bodies who funded the research. The research presented in this thesis was supported by the funding of Slovenian Research Agency (ARRS) through the PhD grant and project grant L4-5525. My visit to Santiago Elena's group in Valencia was supported by an EU-financed COST action FA1407.

First, I would like to thank my supervisor, Maja Ravnikar, who supported and guided all of my work throughout the past three years, shared my research enthusiasm and made sure to keep a broad picture. Big thanks also to Ion Gutierrez-Aguirre and Matevž Rupar for their help with chromatography, and, even more, for many enthusiastic (scientific) discussions. Thanks to Tomaž Curk (FRI) for the help with bioinformatics. Thanks to Andrej Steyer (IMI), Nataša Toplak and Simon Koren (Omega) for a great and fruitful collaboration. I had the opportunity to spend a part of my PhD period at Jan Kreuze's laboratory in Lima, where I learned a lot about sRNA sequencing, but also enjoyed the very friendly company of Dina, Rocio, Joao, Marcos, Anna, Mirella and others. Thank you! Thanks also to Santiago Elena from the Polytechnic University of Valencia for hosting me in his lab. I am happy I have been able to learn from the great knowledge about viral evolution he possesses. Thanks also to the members of the dissertation committee, Kristina Gruden, Neil Boonham and Jernej Jakše.

Just as important are also all other current and past members of the FITO group. It is a pleasure and a privilege to work in such a friendly and stimulating environment. I realized I learned or received help from each and every one of you. Thank you all! Especially, I would like to thank Špela A, Maja K, Anja, Larisa and Zala for the help with plants inoculations, Lidija for the preparation of potato plantlets, Nataša and Tanja for great office company and many helpful insights, Ana L, Anna, Živa, Tjaša S, Klara and many others for discussions over coffee.

I would not have come to this point, if there were no inspiring teachers, be it in primary or secondary school, or later on during my studies and my first steps into research. Thank you all!

Thanks to my parents and grandparents for supporting me and helping me to become the person I am today. Thanks to Martin and all of my friends for the support and a lot of quality time off-duty.

Abstract

In a recent decade, the advance of high-throughput nucleic acid sequencing technologies, termed also next-generation sequencing (NGS), broadened our understanding of the diversity and abundance of viruses. Viruses were shown to be likely the most abundant biological entities in a biosphere, comprising an unparalleled genetic diversity. NGS represents truly generic method for massive detection and sequencing of viral nucleic acids and as such offers new opportunities for studies of viral biology on different levels. When focusing the research on viruses, removal of host background nucleic acids before NGS is often needed. This can be achieved by several approaches, differing in their efficiency. Within this doctoral thesis we aimed to establish new sample preparation approaches to enhance NGS-aided virus research, as well as to use and compare the newly developed and established technological solutions to dig into the diversity and evolution of viruses.

First, we coupled Convective Interaction Media (CIM) monolithical chromatography based purification of viral particles with NGS to sequence a novel human enteric virus (*Mammalian orthoreovirus*), as well as one of the most important plant pathogenic viruses (*Potato virus Y*). Such an approach enhanced the reconstruction of viral consensus sequences in both cases. To characterize the genome of a new strain of another important plant virus (*Potato virus X*), we have employed deep sequencing of virus-derived small RNAs (sRNA), which is currently the most used approach for discovery of new viruses in plants. sRNA deep sequencing of several isolates of potato virus X from South America revealed new phylogenetic lineage of this virus in the region and enabled discrimination between two strains of the virus in a mixed infection. Both of the previously used approaches (CIM purification of viral particles and sRNA deep sequencing) were then compared, using potato-*Potato virus Y* as a model system. We demonstrated that both sequence pools (virus derived sRNAs and RNA from viral particles) reflect a very similar structure of virus populations within a plant. Finally, we used sRNA deep sequencing to follow the temporal dynamics of PVY populations during an experimental evolution in potato genotypes, which differ in response to the virus. The results showed an important role of the interplay between selection and genetic drift during the evolution of the virus and indicated a diversifying selection acting on the virus in the less permissive genotypes of potato (genotypes efficient in accumulation of salicylic acid).

The results of research work presented in this thesis are relevant for broader virology community in the field of new virus discovery and detailed studies of within-host virus population diversity. Specifically, (I) monolithic chromatography based sample preparation approach will enhance the pace of new virus discovery using NGS; (II) first demonstration of the use of sRNA deep sequencing to discriminate between two strains of the same virus in a mixed infection broadens the utility of the method; (III) a comprehensive characterization of plant virus population structure and comparison of the diversity of two different viral sequence pools within a plant gives a firm baseline for further studies of within-host variability of plant viruses; (IV) time-sampled small RNA deep sequencing allows the use of innovative methodological solutions for viral haplotype reconstruction and provides new insights into understanding of viral microevolution.

Povzetek

Z napredkom tehnologij, ki omogočajo hitro sočasno določanje nukleotidnih zaporedij in jih označujemo tudi s pojmom sekvenciranje naslednje generacije (NGS), smo pridobili širši vpogled v raznolikost in pogostost virusov. Ugotovili smo, da so virusi verjetno najpogostejši in najbolj raznoliki biološki subjekti v biosferi. NGS omogoča sočasno zaznavo in določanje zaporedij vseh nukleinskih kislin v vzorcu, s tem pa odpira mnoge nove priložnosti za študij biologije virusov na različnih ravneh. Pri raziskavah, pri katerih se osredotočamo na virusna zaporedja, moramo pogosto pred izvedbo NGS odstraniti nukleinske kisline gostitelja, saj te predstavljajo nezaželeno ozadje. To lahko dosežemo z uporabo različno učinkovitih tehnik za pripravo vzorcev. V okviru tega doktorskega dela smo želeli razviti nove načine priprave vzorcev, ki bi izboljšali raziskave virusov, ki temeljijo na NGS. Prav tako smo želeli uporabiti in primerjati tako nove kot že obstoječe tehnike obogatitve virusov in z njihovo pomočjo dobiti vpogled v raznolikost ter evolucijo virusov.

Z uporabo kombinacije CIM monolitne kromatografije (za čiščenje virusnih delcev) in NGS smo določili zaporedje novega humanega enteričnega virusa (orthoreovirus sesalcev), nato pa tudi zaporedje pomembnega patogenega rastlinskega virusa (virus Y krompirja). Uporaba tega pristopa je izboljšala učinkovitost rekonstrukcije zaporedja genoma v primeru obeh virusov. Za določitev genoma novih različkov drugega pomembnega rastlinskega virusa (virus X krompirja) smo uporabili sekvenciranje malih RNA, ki je trenutno najbolj pogosto uporabljen pristop za odkrivanje novih virusov v rastlinah. Zaznali smo novo filogenetsko linijo virusa v Južni Ameriki in uspeli ločiti med dvema različkoma virusa v mešani okužbi. Oba uporabljena načina priprave vzorcev (čiščenje virusnih delcev in sekvenciranje malih RNA) smo nato primerjali in pri tem uporabili virus Y krompirja, ki okužuje krompir, kot modelni sistem. Pokazali smo, da obe skupini zaporedij (virusne male RNA in RNA iz virusnih delcev) odražata zelo podobno strukturo virusnih populacij znotraj rastline. Globoko sekvenciranje virusnih malih RNA smo nato uporabili za študij časovne dinamike populacij PVY med eksperimentalno evolucijo v genotipih krompirja, ki se razlikujejo v odzivu na virus. Rezultati so razkrili pomemben prispevek prepletanja naravne selekcije in genetskega drsa v evoluciji virusa in nakazali na ločujočo selekcijo, ki deluje na virus v genotipih krompirja, ki so manj občutljivi na virus (genotipi, ki kopičijo salicilno kislino).

Rezultati raziskovalnega dela, ki so predstavljeni v tej disertaciji, so pomembni za razumevanje raznolikosti in evolucije virusov, še zlasti pa predstavljajo pomemben prispevek na področju odkrivanja novih virusov in študija raznolikosti virusnih populacij znotraj gostiteljev. Natančneje, (I) nov način priprave vzorcev z monolitno kromatografijo bo pripomogel k bolj učinkovitemu odkrivanju virusov z uporabo NGS; (II) prikaz uporabe sekvenciranja malih RNA za ločevanje dveh različkov virusa v mešani okužbi razširja možnosti uporabe te metode; (III) izčrpen opis populacijske strukture in primerjava dveh skupin virusnih zaporedij znotraj rastline predstavlja pomembno izhodišče za študij raznolikosti rastlinskih virusov znotraj gostitelja; (IV) sekvenciranje malih RNA v zaporednih časovnih točkah omogoča uporabo inovativnih metodoloških pristopov za rekonstrukcijo virusnih haplotipov in nove vpoglede v mehanizme virusne mikroevolucije.

Contents

Abbreviations **xiii**

1 Introduction **1**
1.1 Viruses and Different Levels of Their Diversity ... 1
1.2 Detection and Discovery of Viruses ... 1
1.3 Next-Generation Sequencing in Virology ... 2
1.4 Scientific Problems and Aims of this Thesis ... 2
1.4.1 Host background and enrichment for viral nucleic acids ... 2
1.4.2 Studies of viral populations' diversity ... 3
1.4.3 Experimental evolution of virus populations ... 4
1.5 Research Hypotheses ... 5
1.6 Publications Included and Candidate's Contributions ... 5

2 Scientific Publications **7**
2.1 High Similarity of Novel Orthoreovirus Detected in a Child Hospitalized with Acute Gastroenteritis to Mammalian Orthoreoviruses Found in Bats in Europe ... 7
2.2 Complete Genome Sequences of New Divergent Potato Virus X Isolates and Discrimination Between Strains in a Mixed Infection Using Small RNAs Sequencing Approach ... 17
2.3 Deep Sequencing of Virus-Derived Small Interfering RNAs and RNA from Viral Particles Shows Highly Similar Mutational Landscapes of a Plant Virus Population ... 25
2.4 Time-Sampled Population Sequencing Reveals the Interplay of Selection and Genetic Drift in Experimental Virus Evolution (Submitted Manuscript) ... 37

3 Discussion **65**
3.1 Viral Sequence Enrichment Approaches ... 65
3.2 Comparison of Virus Derived sRNAs and RNA from Purified Virus Particles ... 66
3.3 Experimental Evolution of PVY ... 67

4 Conclusions **69**

Appendix A Supplementary Material of Included Publications **71**
A.1 Supplementary Material for Publication 2.1 ... 71
A.2 Supplementary Material for Publication 2.2 ... 79
A.3 Supplementary Material for Publication 2.3 ... 84
A.4 Supplementary Material for Publication 2.4 ... 84

Appendix B Permissions for Reproduction of Included Publications **91**
B.1 Permission for Reproduction of Publication 2.1 ... 91

B.2 Permission for Reproduction of Publication 2.2 92
B.3 Permission for Reproduction of Publication 2.3 97

References **99**

Bibliography **103**

Biography **105**

Abbreviations

CIM	. . . Convective Interaction Media
ddPCR	. . . digital droplet polymerase chain reaction
DNA	. . . deoxyribonucleic acid
ML	. . . maximum likelihood
MRV	. . . *Mammalian orthoreovirus*
N_a	. . . effective population size
NGS	. . . next-generation sequencing
PCR	. . . polymerase chain reaction
PVX	. . . *Potato virus X*
PVY	. . . *Potato virus Y*
qPCR	. . . quantitative polymerase chain reaction
RNA	. . . ribonucleic acid
RT-PCR	. . . reverse transcription, followed by polymerase chain reaction
s	. . . selection coefficient
sRNAs	. . . small ribonucleic acids
VP	. . . viral particles
WF-ABC	. . . Wright-Fisher Approximate Bayesian Calculations

Chapter 1

Introduction

1.1 Viruses and Different Levels of Their Diversity

Viruses and virus-like agents (*e.g.*, viroids) represent the most abundant and likely the most diverse biological entities in the biosphere [1]. Even though they were first detected as plant and animal pathogens [2], it is now known that they infect more or less all forms of living organisms and are found in most diverse environments [3]–[5].

Virus particles (virions) are comprised of nucleic acids protected by structural proteins. Additionally, some viral particles contain also non-structural proteins and lipid membranes [2]. Virus genetic information can be encoded by different forms of nucleic acids: single-stranded RNA (ssRNA), double-stranded RNA (dsRNA), single-stranded DNA (ssDNA) and double-stranded RNA (dsRNA) and different replication strategies are used even within each of these virus groups. Currently, there are close to 4000 viral species recognized by International Committee on Taxonomy of Viruses [6], however, this number likely represents a very small portion of the true viral diversity. For example, a recent study showed that only for mammalian hosts, the minimal predicted number of undiscovered viruses amounts to 320,000 [7].

Besides great variability of viruses on the species level, viruses also constitute highly diverse populations within their hosts. For some groups of viruses (particularly ssRNA and ssDNA viruses) error rates of virus-encoded polymerases are especially high [8], generating an unparalleled diversity of their populations. Thus, the viruses are present within a host as swarms of mutant sequences, often referred to as quasispecies. The composition of the mutant swarm is an important characteristic of the virus, since it represents a reservoir of genetic variants, which can be subjected to different evolutionary processes [9]. High mutation rate of (RNA) viruses and their quasispecies nature drive their quick adaptability and evolution rate. Thus, new viruses or new pathogenic viral strains can emerge as a result of different processes, such as host shifts [10].

1.2 Detection and Discovery of Viruses

Historically, viruses have been detected and discovered by observation of disease symptoms on different hosts (especially for plant viruses), observation of virus particles by electron microscopy or by specifically targeting their proteins (most commonly ELISA and ELISA-derived methods) or nucleic acids (PCR based methods) [11]. All of these methods require either high concentration of viruses in host samples (electron microscopy) or previous knowledge about biology or genome sequence of the virus. Due to high variability of viruses and consequentially a lack of possible highly conserved marker genes (as opposed to bacteria and eukaryotes) targeted methods, relying on nucleic or amino acid sequences, are

limited in detection to a specific viral species or at most specific viral genera. Consequently, until recently, discovery of new or unexpected viral species represented a complicated task, requiring employment of a combination of different methods.

1.3 Next-Generation Sequencing in Virology

With the advent of next-generation sequencing (NGS), a truly generic method for sensitive detection and discovery of viruses emerged. The term NGS covers a broad range of technologies, which enable high throughput sequencing of (most commonly short) nucleic acid sequences, not requiring any specific knowledge about the investigated (viral) sequences. Most of the currently often used NGS technologies employ similar basic workflow, using different approaches for amplification of signal and sequencing. First, adaptor sequences are ligated to the library of short DNA or RNA fragments. These adaptors contain specific sequences, which enable amplification and subsequently sequencing using technology-specific approaches [12]. Currently the most used NGS platform is Illumina's sequencing by synthesis, employing bridge amplification and reversible florescence dye sequencing. It is followed by Life Technologies Ion line of sequencers and Roche 454, which both employ droplet PCR for amplification, but Ion line sequencers use semiconductor technology for sequencing, whereas 454 use fluorescently labelled nucleotides. So-called third generation sequencing technologies are also establishing their position among the high-throughput sequencing platforms, notably PacBio and recently also Oxford Nanopore, both characterized by long read lengths and currently considerably high error rates.

The untargeted NGS approach coupled with its high-throughput capacities made a dramatic impact in the field of virology. Several levels of virus diversity can now be efficiently studied using NGS. New applications became possible and numerous fields of study increased their resolution, including viral metagenomics, untargeted discovery of new viruses, characterization of putative viral pathogens, virus epidemiology studies and studies of within-host virus populations [13].

1.4 Scientific Problems and Aims of this Thesis

1.4.1 Host background and enrichment for viral nucleic acids

When focusing on viral sequences, one has to consider that they are present in the vast background of the host ones. Sequencing such sample can result in mostly host sequences, and sometimes less than a percent of reads can be attributed to viral origin [14]. This can considerably affect the price and time of the analysis. To solve this problem, enrichment for viral nucleic acids is most often needed. There are several approaches to achieve this [15], [16], such as: depletion of ribosomal RNA [17], classical purification of virions [18], enrichment for double stranded RNAs [19], subtractive hybridization [20] and small RNA sequencing [21]. We have focused on two of the techniques, namely: (i) virus particles purification and (ii) small RNA deep sequencing.

Purification of virus particles (VP) from a complex host sample before NGS can provide high enrichment for viral sequences [18]. Classical virus purification protocols consist of several steps, including several centrifugation and ultracentrifugation steps and can thus be time-consuming. Even though this approach can be very efficient for removal of host sequences, speed, specificity or/and complexity of protocols used for purification of different viruses limit its usage. An alternative purification approach is Convective Interaction

Media (CIM) monolithic chromatography, which can considerably shorten the time needed for purification of viral particles [22], [23] and has not been coupled with NGS before.

Our **first aim** was to establish new sample preparation approaches to enhance NGS-aided research of viruses by coupling CIM monolithical chromatography and NGS and apply the technology to characterize new viruses (*Mammalian orthoreovirus*) and study within-host diversity of important pathogens (*Potato virus Y*).

Deep sequencing of small RNAs has been widely adopted as an approach to search for viruses in plants [16] and invertebrates [24]. Virus-derived small interfering RNAs (vsiRNAs) are produced from viral double-stranded RNA (dsRNA) or dsRNA-like parts (*e.g.*, hairpins in secondary structure of RNA) through a host RNA interference mechanism. In plants, it has been shown that small RNAs are generated from all the different types of viruses, including DNA viruses [25] and as such, likely represent a truly generic technique for detection of viruses in plants.

Our **second aim** was to use existing virus enrichment techniques and NGS (sRNA deep sequencing) to characterize the genomes of several isolates of widely distributed plant virus (*Potato virus X*) in an understudied region (Andean region of South America).

Virus enrichment methods differ in efficiency and difficulty of execution, each having its own strengths and weaknesses [16]. The choice of the method can depend on the type of the virus under study or the type of the study (e.g., screening for all viruses, searching for specific group of viruses, etc.). There is a lack of a direct comparison of the virus enrichment methods, even though such comparison would provide valuable information for many research and diagnostics labs, which are starting to implement NGS to their analysis workflows.

Our **third aim** was to compare two of the previously used virus enrichment techniques in terms of efficiency of enrichment and consensus virus genome sequence generation, using PVY-potato as a model system.

1.4.2 Studies of viral populations' diversity

The composition of within-host virus populations has been studied in the past decades using Sanger sequencing of molecular clones. Such an approach allowed to study diversity of viral populations with limited resolution, usually sequencing 10-100 clones per population, commonly focusing on amplified shorter parts of the viral genome. The high-throughput nature of NGS enabled a leap also in this area, now providing coverage of complete genomes and much higher resolutions, usually amounting 1000-100000x sequencing depth. It is now possible to discover variants present in population even at very low frequencies (around and below 1%), providing new insights into diversity and evolution of viruses. Even though NGS provides a much higher resolution in virus population studies compared to Sanger sequencing, its full potential in this field is still developing. Sequencing error rates are decreasing with a constant development of sequencing platforms, but are still significant enough to prevent reliable detection of mutations with lowest frequencies, *e.g.*, nonsense mutations. This is why improved approaches (which enable to further filter errors), such as individual molecular barcoding [26] or circular sequencing [27], were recently developed. Further, several viral sequence enrichment approaches, already mentioned above, provide a subsample of total viral sequence representation in a host. Even though different approaches have been used to study diversity of virus populations, it has not been tested if they describe the same picture of a viral population structure. For example, sRNA deep sequencing was used to characterize virus populations in few examples [28], [29], yet, it was not empirically shown that frequencies of variants obtained by sequencing sRNAs are similar to the ones present in e.g. pool of packed viral genomes.

Our **fourth aim** was to compare the structure of viral populations described by each of these two pools (sRNAs vs VP) using PVY-potato as a model system.

1.4.3 Experimental evolution of virus populations

RNA viruses are, due to their high mutation and recombination rates, one of the most quickly evolvable biological entities and as such represent a good model for studies of basic evolutionary mechanisms, such as natural selection and genetic drift [30]. Indeed, they have been commonly used in evolution experiments to answer important questions about microevolution (dynamics of variants in virus populations) [31] and also to detect major evolutionary transitions [32], trying to tackle the field of macroevolution. Again, the use of NGS for the description of virus population structures enabled to get a much more detailed picture of dynamics of viral populations within hosts during experimental evolution studies. Moreover, the studies of viral evolution in hosts that differ in response to the virus could provide us with a broader understanding of the mechanisms, which shape the diversity of viruses in wild and anthropogenic ecosystems, leading to better management strategies for prevention of viral epidemics.

Most of the experimental evolution studies rely on the snapshots of viral population structures, *e.g.*, at the end of the experiment. Such an approach offers limited ability to detect imprints of selection and/or genetic drift, mostly relying on the ratio between synonymous and nonsynonymous mutations in a fixed time point. Alternatively, tracking the dynamics of variants in time provides better options to detect variants under selection, using population genetics approaches [27],[33]–[35].

Further, short read lengths of most of the established NGS technologies limit the reconstruction of complete viral haplotypes, which are needed for reconstruction of evolutionary histories of populations by using *e.g.* coalescent phylogenetic methods. Currently available methods for reconstruction of full or partial viral haplotypes from NGS data require relatively long reads, they are difficult to optimize and produce a high number of false positive results [36]. A possible alternative to these approaches is to follow the changing frequencies of SNPs in time (SNPs trajectories) and search for the patterns of covariation between them, using this information to assign the variants into groups [37] and ultimately reconstruct full or partial haplotypes [35].

Experimental evolution studies can be employed to reveal the effect of different environments on the evolution of viruses, e.g., following the adaptation to different host species [38], different vectors [39] or a presence of a drug [34]. In a recent study, the experimental evolution of *Tobacco etch virus* (TEV, genus *Potyvirus*, family *Potyviridae*) in different ecotypes of the same species was performed, revealing some parallel mutations in the same ecotypes, indicating difference in evolution of the virus in hosts with very similar genetic background, yet different response to the virus [40]. In the experimental pathosystem, which we are focusing on (potato-PVY), salicylic acid represents one of the important components of the potato response to the virus. Even though the changes in plant physiology on different levels (transcriptomics, proteomics, metabolomics) have been studied in potato genotypes efficient and depleted in accumulation of salicylic acid [41]–[43], conversely, the influence of differently permissive potato genotypes on the evolution of the virus have not yet been studied.

Our **fifth aim** was to address the abovementioned problems by studying dynamics of PVY evolution in several potato genotypes, which differ in response to the virus. We aimed to use a time-sampled sRNA deep sequencing approach to track the temporal dynamics of within-host virus populations and use this information to reconstruct ancestral haplotypes and reveal the impact of selection and genetic drift on the evolution of the virus populations in differently permissive potato genotypes.

1.5 Research Hypotheses

1. CIM monolithic chromatography will enable efficient purification of viruses before NGS and will as such positively impact the results of sequencing and enhance de-novo reconstruction of viral genomes.

2. Small RNA sequencing will enable genome characterization of several unknown potato virus X isolates from South America. This will illustrate that diversity of potato virus X strains in this region is greater than currently known.

3. Comparison of two sample preparation approaches from point 1 and 2 will show that RNA isolated from viral particles and virus-derived small interfering RNAs reflect a similar within-plant population structure.

4. Time-sampled population sequencing approach will enable assignment of groups of linked variants in PVY populations during experimental evolution and reveal the differences in evolution of the virus in different potato genotypes.

1.6 Publications Included and Candidate's Contributions

In the first paper (*High Similarity of Novel Orthoreovirus Detected in a Child Hospitalized with Acute Gastroenteritis to Mammalian Orthoreoviruses Found in Bats in Europe*) the coupling of CIM monolithic chromatography for purification of viruses and NGS is described for the first time. In this research, we have used a developed CIM chromatographic method [44] with modifications to purify putative Orthoreovirus and then sequence the RNA isolated from purified virions using Ion Torrent technology. The candidate is the coauthor of this publication. He has performed bioinformatics analyses of the obtained sequencing reads, including reconstruction of complete viral genome and comparison of the data obtained using three different viral enrichment strategies, and drafted the corresponding part of the manuscript.

The second paper (*Complete genome sequences of new divergent potato virus X isolates and discrimination between strains in a mixed infection using small RNAs sequencing approach*) describes the use of a small RNA Illumina deep sequencing approach for the genome characterization of several potato virus X isolates (PVX) and their discrimination in a mixed infection. The candidate is the first author of the publication. He has performed bioinformatics analyses of the obtained sequencing reads, reconstructing the complete viral genomes and discriminating between two strains of the virus in a mixed infection. Further, he performed phylogenetic analyses to taxonomically classify the characterized viral isolates, designed diagnostic PCR primer pairs for quick discrimination between different PVX strains by RT-PCR, and drafted the manuscript.

The comparison between the two approaches (small RNA deep sequencing and sequencing of RNA from CIM purified viral particles) is described in the third paper (*Deep Sequencing of Virus-Derived Small Interfering RNAs and RNA from Viral Particles Shows Highly Similar Mutational Landscapes of a Plant Virus Population*). The two pools of viral sequences within a plant are compared in terms of efficiency of genome reconstruction and population diversity description, using *Potato virus Y*-potato as a model system. The candidate is the first author of this publication; he has designed and conducted the wet-lab part of the experiment, performed bioinformatics analyses, comparisons and drafted the manuscript.

Deep sequencing of small RNAs was then used to track the dynamics of a PVY population structure within different genotypes of potato during an evolution experiment. The results show the interplay between the selection and genetic drift during the PVY evolution and indicate a possible diversifying selection affecting the virus in the more permissive potato genotypes. The study is described in the fourth publication (*Time-sampled population sequencing reveals the interplay of selection and genetic drift in experimental virus evolution*, currently as a submitted manuscript). The candidate is the first author of this publication; he has designed and conducted the wet-lab part of the experiment, performed bioinformatics analyses, population genetics analyses and drafted the manuscript.

Chapter 2

Scientific Publications

2.1 High Similarity of Novel Orthoreovirus Detected in a Child Hospitalized with Acute Gastroenteritis to Mammalian Orthoreoviruses Found in Bats in Europe

Andrej Steyer, Ion Gutiérrez-Aguire, Marko Kolenc, Simon Koren, Denis Kutnjak, Marko Pokorn, Mateja Poljsak-Prijatelj, Nejc Racki, Maja Ravnikar, Martin Sagadin, Adela Fratnik Steyer and Natasa Toplak

Journal of Clinical Microbiology, 2013, 51(11): 3818-25. DOI: 10.1128/JCM.01531-13

In this publication, we described a discovery of a novel *Mammalian orthoreovirus*, which was found in a stool sample from a child with acute gastroenteritis. Phylogenetic analyses of a whole-genome sequence showed that the new isolate (SI-MRV01) is highly similar to MRV found in a bat in Germany, indicating a possible zoonotic transmission. In this research, we demonstrated, for the first time, that coupling CIM monolithic chromatography with Ion Torrent sequencing represents an efficient approach for virus sequence enrichment.

The PhD candidate analyzed the NGS reads and reconstructed the complete consensus sequences of all of the viral genome segments. Further, he performed the comparison of the three used viral enrichment strategies, showing CIM monolithic purification of viral particles as the most efficient one. He also helped prepare the paper by drafting the parts of the manuscript describing the NGS sequence analyses and the abovementioned comparisons.

JCM
Journals.ASM.org

High Similarity of Novel Orthoreovirus Detected in a Child Hospitalized with Acute Gastroenteritis to Mammalian Orthoreoviruses Found in Bats in Europe

Andrej Steyer,[a] Ion Gutiérrez-Aguire,[b,e] Marko Kolenc,[a] Simon Koren,[c] Denis Kutnjak,[b] Marko Pokorn,[d] Mateja Poljšak-Prijatelj,[a] Nejc Rački,[b] Maja Ravnikar,[b,e] Martin Sagadin,[a] Adela Fratnik Steyer,[a] Nataša Toplak[c]

Institute of Microbiology and Immunology, Faculty of Medicine, University of Ljubljana, Ljubljana, Slovenia[a]; National Institute of Biology, Ljubljana, Slovenia[b]; Omega d.o.o., Ljubljana, Slovenia[c]; Department of Infectious Diseases, University Medical Centre Ljubljana, Ljubljana, Slovenia[d]; The Centre of Excellence for Biosensors, Instrumentation and Process Control—COBIK, Solkan, Slovenia[e]

Mammalian orthoreoviruses (MRVs) are known to cause mild enteric and respiratory infections in humans. They are widespread and infect a broad spectrum of mammals. We report here the first case of an MRV detected in a child with acute gastroenteritis, which showed the highest similarity to an MRV reported recently in European bats. An examination of a stool sample from the child was negative for most common viral and bacterial pathogens. Reovirus particles were identified by electron microscopic examination of both the stool suspension and cell culture supernatant. The whole-genome sequence was obtained with the Ion Torrent next-generation sequencing platform. Prior to sequencing, the stool sample suspension and cell culture supernatant were pretreated with nucleases and/or the convective interaction medium (CIM) monolithic chromatographic method to purify and concentrate the target viral nucleic acid. Whole-genome sequence analysis revealed that the Slovenian SI-MRV01 isolate was most similar to an MRV found in a bat in Germany. High similarity was shared in all genome segments, with nucleotide and amino acid identities between 93.8 to 99.0% and 98.4 to 99.7%, respectively. It was shown that CIM monolithic chromatography alone is an efficient method for enriching the sample in viral particles before nucleic acid isolation and next-generation sequencing application.

Reoviridae is a highly diverse virus family, including viruses capable of infecting various host species (mammals, reptiles, fish, birds, protozoa, fungi, plants, and insects). These viruses consist of an icosahedric capsid and a segmented genome of 10 to 12 double-stranded RNA (dsRNA) segments (1). They are unenveloped and relatively stable in the environment (2). Two subfamilies have been described within the *Reoviridae* family: *Spinareovirinae* and *Sedoreovirinae*, including 9 and 6 genera, respectively (3). In human medicine, the most recognized are *Rotaviruses* within the *Sedoreovirinae* subfamily and *Orthoreoviruses* within *Spinoreovirinae*. *Orthoreoviruses*, with mammalian orthoreovirus (MRV) as the type species, were already recognized in the 1950s as respiratory and enteric orphan viruses (4). They were found in hosts with or without clinical manifestations (1). MRVs have been reported to date in various mammalian hosts, including human and animal species. In the last few years, they have often been described as the sole pathogen in various hosts presenting severe clinical manifestations, such as hemorrhagic enteritis, acute respiratory infections, central nerve system implications, and others (5–10). There is consequently increasing concern about the widespread nature and pathogenesis of these viruses. A German group of researchers recently reported infections of bat species with MRVs, which resulted in evident pathological signs in the organs of infected bats (11). This was the first report of reoviruses in bats that were not clustered into the species *Pteropine orthoreovirus* but in MRVs. Almost at the same time, an Italian group published data on reovirus detection in various bat species, reporting nucleotide sequences of partial L1 and complete S1 segments, which showed the highest similarity to the German bat isolate (12). Both research groups speculated on bat-to-human interspecies transmission, but there was no evidence to support this hypothesis. A closely related MRV found in a dog with hemorrhagic enteritis had been previously described by another Italian group. The authors proposed that these viruses might be important zoonotic pathogens (13).

The zoonotic potential of reoviruses has already been described and discussed elsewhere (6, 7, 11). The transmission of reoviruses from one host to another is not limited to close contacts but extends to indirect transmission. Infection through contaminated food, water, or other factors in the environment is highly possible, since infective reovirus particles have been found in environmental samples (14–17). Viral persistence outside the host is one of the advantageous features that enables them to spread efficiently.

Researchers throughout the world have recently focused very intensively on zoonotic or potentially zoonotic viruses. Bats are of special interest because of their diversity, the wide spectrum of virus populations found in different bat samples, and a special virus-host interaction (18, 19). Screening of animals for potentially zoonotic viruses has resulted in the discovery of novel viruses. In the last few years, the pathogen discovery field has made a major step forward. Application of new technologies, such as next-generation sequencing (NGS), enables researchers to take a

Received 13 June 2013 **Returned for modification** 22 July 2013
Accepted 5 September 2013
Published ahead of print 11 September 2013
Address correspondence to Andrej Steyer, andrej.steyer@mf.uni-lj.si.
Supplemental material for this article may be found at http://dx.doi.org/10.1128/JCM.01531-13.
Copyright © 2013, American Society for Microbiology. All Rights Reserved.
doi:10.1128/JCM.01531-13

different approach to pathogen discovery, gathering a huge amount of genomic information from a sample of interest in a short period of time (20). In the last decade, the price of sequencing technology has decreased drastically, while the capacity of NGS data obtained has increased sharply (20). This research tool also has major potential as part of microbiological diagnostics in combination with classical or standard microbiological methodologies (21). In the future, NGS could help resolve clinical cases of infections with undetermined etiology. However, when using NGS in pathogen discovery or diagnostics, the preparation of high-quality and high-quantity nucleic acid is of major concern (22). Improved and optimized sample preparation, purification, and concentration of the target are essential for achieving the best results in NGS. Isolation of as pure as possible viral nucleic acid enables easier sequencing of the target of interest and faster bioinformatics workflow. Convective interaction medium (CIM) chromatography is the method of choice for virus concentration and purification from different samples (23, 24).

In this work, we discovered an MRV strain with high similarity to orthoreoviruses detected in European bats. The MRV was found in a child with acute gastroenteritis requiring hospitalization. This is an indication of zoonotic transmission. The whole virus genome was sequenced using the Ion Torrent platform. It is highly possible that this group of MRVs detected in bats in Europe is widespread and could easily cross species barriers. However, the detailed pathogenesis has yet to be determined. To the best of our knowledge, this is the first report of a bat MRV detected in a human with a clinical manifestation. In addition, a new approach for sample preparation prior to NGS is presented. A chromatographic method using CIM monoliths was used, which proved to be effective in enriching and purifying the sample of MRV prior to RNA extraction and NGS application.

CASE REPORT

A 17-month-old boy presented at the Department of Infectious Diseases, University Medical Centre Ljubljana, Ljubljana, Slovenia, with a 5-day history of nonbloody diarrhea. Stools were frequent; the child refused to drink and after 4 days developed fever. His previous medical history was unremarkable, with only a fever-related thrombocytopenia at the age of 3 months. He had been vaccinated against rotavirus. He attended a day care center at which no gastroenteritis cases were observed.

On admission, the child was mildly dehydrated and afebrile, and the abdomen was not sensitive to palpation. The C-reactive protein (CRP) concentration was 20 mg/liter, the white blood cell (WBC) count was 8.4×10^9/liter (with normal differential), the red blood cell (RBC) count was 4.94×10^{12}/liter, the hemoglobin (Hb) level was 124 g/liter, the hematocrit (Hct) was 0.384, and the blood sugar, blood urea nitrogen, creatinine, and electrolytes were normal. The child was given 500 ml of intravenous fluids and discharged home. The next day he returned to the department because of ongoing fever, frequent nonbloody stools accompanied by colicky pains, and red and swollen gums with oral ulcers. Aphthous stomatitis was diagnosed, in addition to gastroenteritis, and the child was admitted for parenteral rehydration and discharged the next day. He made an uneventful recovery. The total duration of diarrhea was 8 days.

A stool sample was analyzed according to the standard diagnostic protocol, including classical bacterial examination with culturing techniques, checking for the most common bacterial pathogens (*Campylobacter* spp., *Salmonella* sp., pathogenic *Escherichia coli*, *Yersinia* sp., and *Shigella* sp.). A virological examination was carried out at the same time: negative-staining electron microscopy (EM) and enzyme-linked immunosorbent assays (ELISAs) for group A rotaviruses and adenoviruses 40 and 41 (Meridian Bioscience, Inc., Cincinnati, OH) and in-house real-time reverse transcription (RT)-PCR for noroviruses of genogroups I and II and astroviruses, as described previously (25, 26). Reoviruses were observed by electron microscopy in the stool sample taken at the first admission.

MATERIALS AND METHODS

Virus identification and cell culture propagation. Cell line LLC-MK2 (rhesus monkey kidney epithelial cells) grown in modified Eagle's medium (MEM) supplemented with 10% fetal calf serum was used for virus propagation. A 10% stool suspension was prepared and clarified by centrifugation at 1,600 $\times$ *g*. Clarified supernatant of a 10% stool suspension was filtered through a 0.22-μm-pore filter (Millipore, Billerica, MA) and inoculated on a 48-h cell culture monolayer in 25-cm^2 flasks (TPP, Trasadingen, Switzerland). After incubation for 1 h at 37°C, the inoculum was removed, and cells were supplemented with the original growth medium and incubated further at 37°C in an atmosphere with 5% CO_2 until the appearance of a cytopathic effect (CPE). A 10-fold serial dilution of reovirus isolate was prepared to determine the 50% tissue culture infective dose ($TCID_{50}$) value (27). A known concentration of reovirus isolate helped us to estimate the reovirus concentration in the child's stool sample using real-time RT-PCR, as described previously (11).

In addition, infected cells were prepared for thin sectioning. Briefly, the specimen was fixed with 2% (vol/vol) glutaraldehyde in 0.1 M phosphate-buffered saline (PBS; pH 7.4) for 2 h, rinsed with 3 changes of PBS, and postfixed with 1% (vol/vol) OsO_4 in PBS for 2 h. After being washed, the specimen was dehydrated in a graded series of ethanol and embedded in epoxy (agar low-viscosity) resin, following the standard protocol. Ultrathin sections were collected on carbon-coated 200 mesh copper grids and stained with 1% uranyl acetate and 1% lead citrate. In addition, clarified stool suspension was used for direct EM examination of the sample after negative staining with 2% phosphotungstic acid (pH 4.5). EM grids were screened at 80 kV in a JEM 1200 EXII transmission electron microscope (JEOL, Tokyo, Japan).

Sample preparation for NGS. The RNA for the sequencing procedure was obtained from two sources: clarified stool suspension and cell culture supernatant after virus propagation. Two samples (6 ml of stool and 6 ml of cell culture supernatant) were incubated with 1,000 U/ml of Benzonase (Novagen, San Diego, CA) for 1 h at room temperature (samples 1 and 2) to digest nonencapsidated nucleic acid. Samples were then diluted 6$\times$ in chromatography running buffer (50 mM HEPES; pH 7) and left overnight at 4°C. An additional 6-ml sample of cell culture supernatant was diluted in the same way but without Benzonase treatment (sample 3). Ten milliliters of all three samples was loaded into a 0.34-ml-volume CIM-QA disk (BIA Separations, Ajdovščina, Slovenia), using an AKTA purifier chromatographic system (GE Healthcare, Uppsala, Sweden). After loading, nonbound material was washed with running buffer, and the elution of bound viruses proceeded in two different ways. The Benzonase-treated stool suspension and cell culture samples (samples 1 and 2) were eluted in a single step, by including 1 M NaCl in the running buffer. The Benzonase-untreated cell culture sample (sample 3) was eluted by using a gradient of 0 to 500 mM NaCl in ≈60 column volumes in order to purify the viruses as much as possible. In all three cases, the elution was collected in 1-ml fractions. The presence and approximate number of putative reoviruses in the elution fractions were estimated by the virus particle counting method under an electron microscope using a negative-staining technique (28). To have an indication of the presence of eukaryotic nucleic acid, RNA was extracted from the fractions using a QIAamp virus RNA kit (QIAgen, Valencia, CA) and applied to a eukaryotic 18S assay (Life Technologies, Applied Biosystems Division, Foster City, CA), in a one-step

real-time RT-PCR format, using an Ag-Path master mix (Life Technologies) and an ABI 7900 HT real-time PCR system. To overcome the lower sensitivity of the electron microscopy technique and in order to confirm the presence of reovirus, the RNA from fractions 6 to 10 obtained in the stool sample run, in which no virus particles were observed, was applied to RT-PCR using a set of reovirus generic primers (5). Fractions that were selected for the sequencing procedure (fractions 6 to 10, 3 to 6, and 15 to 18 for runs involving samples 1, 2, and 3, respectively) were pooled, and RNA was extracted from the whole pooled volume using a TRIzol LS Plus RNA purification kit (Life Technologies, Invitrogen Division, Carlsbad, CA). The final purified RNA solution obtained from each of the three samples was further evaporated to a final volume of ≈100 μl in a GeneVac personal evaporator (Genevac, Ltd., Ipswitch, United Kingdom).

Ion Torrent library preparation and sequencing. The RNA library was prepared using an Ion Total RNA-Seq kit v2 (Life Technologies, Invitrogen Division, Darmstadt, Germany) according to the manufacturer's protocol (4476286, revision D) for a low input of starting material (between 10 and 100 ng). The amount and size distribution of library DNA fragments were determined with a Labchip GX instrument (Caliper, Life Sciences, MA). Emulsion PCR and enrichment steps were carried out using an Ion OneTouch200 template kit v2 DL as described in the protocol (MAN0006957, revision 5.0). Assessment of the Ion Sphere particle quality was undertaken between the emulsion PCR and enrichment steps with an Ion Sphere quality control kit using a Qubit 2.0 fluorometer (Life Technologies). Each library was sequenced on a separate Ion 314 Chip (Life Technologies). Signal processing and base calling were performed with Torrent Suite software version 3.4.2. Adapter sequences were trimmed using the same software.

Complete genome sequence generation. Trimmed reads were used for *de novo* assembly using CLC Genomic Workbench 6.0 (CLC Bio, Aarhus, Denmark), with default program parameters (see Table S2 in the supplemental material). The generated contigs were compared for similarity against all virus sequences deposited in the NCBI GenBank database using the Basic Local Alignment Search Tool (BLASTn and BLASTx). In all three samples, the majority of the contigs from all 10 viral genome segments showed the highest identity values with the MRV strain 342/08 genome, deposited in GenBank under accession no. JQ412755 to JQ412764 (11). This strain was therefore selected as the reference genome. Reads obtained from each sample were initially mapped separately to the reference sequence using default parameters (see Table S2). The generated consensus sequences were additionally compared, and no difference was observed between the consensus sequences of the three different samples. The reads of all three samples were subsequently mapped together to the reference. After removal of duplicate reads, all of the single nucleotide polymorphisms (SNPs) and indels (in comparison to the reference genome sequence) were explored visually. They were considered to be reliable if they were covered with at least 30 reads and were derived from both the positive and negative strands of RNA.

In order to compare the efficiencies of host (LLC-MK2 cell culture) nucleic acid removal between samples 2 and 3, reads were mapped to the *Macaca mulatta* genome (rheMac3; Beijing Genomics Institute, Shenzhen, China) (GenBank assembly ID GCA_000230795.1). To prevent unspecific mapping, reads smaller than 25 bp were removed, and highly stringent parameters were used for these mappings (given in Table S2 in the supplemental material).

Sequence analysis. Consensus genome segments were further analyzed for open reading frames (ORFs), which were determined using Geneious 6.0.5 version (Biomatters, Ltd., Auckland, New Zealand), and deduced amino acid sequences were obtained.

For phylogenetic analysis, whole-genome MRV sequences available in GenBank were downloaded, and multiple sequence alignments (ClustalW) were carried out. A neighbor-joining phylogenetic tree (Kimura 2-parameter system) was made to show the phylogenetic relationships, using the MEGA 4 software system (29). Branch support was assessed by bootstrap analysis of 1,000 replicates. For reassortment or recombination events, the concatenated whole-genome sequence of the Slovenian MRV was compared to other whole-genome MRVs from GenBank. A similarity plot was constructed using SimPlot software, version 3.5.1 (30).

Nucleotide sequence accession numbers. The nucleotide sequences obtained in this study have been deposited in GenBank under accession no. KF154724 to KF154733.

RESULTS

Bacterial laboratory examination of the patient's stool sample was negative at the first and second admissions. Virological examination was performed only on the sample at first admission and was negative for specific antigen detection tests for group A rotaviruses and adenoviruses 40 and 41, and molecular tests were negative for noroviruses in genogroups I and II and astroviruses, which are the most prevalent pathogens detected in the age group of the case. However, clear 75-nm reovirus particles were observed upon EM examination of the stool suspension (Fig. 1A). After inoculation of the filtered stool suspension on an LLC-MK2 cell culture, CPE appeared 48 h postinoculation, comprising cell rounding with a notable membrane and final cell lysis. Again, reovirus particles were observed under an electron microscope, on examination of the cell culture supernatant without a prior concentration step (Fig. 1B) and in a thin section of infected cell culture (Fig. 1C). In ultrathin sections, several cytoplasmic inclusions with densely packed virus particles were found. An electron-dense center and a clear outer rim were observed in all particles.

The cell culture supernatant was cleared, and the reovirus concentration was determined to be 2.43×10^8 $TCID_{50}$/ml. The regression equation was obtained from the real-time RT-PCR analysis of serial 10-fold dilutions of reovirus isolate ($y = 40.73 - 3.49x$; $R^2 = 0.999$). The theoretical $TCID_{50}$/ml was calculated for reoviruses in child's stool sample, using the quantification cycle (C_q) value of the reovirus real-time RT-PCR analysis. The concentration was estimated to 2.26×10^7 $TCID_{50}$/ml, considering the 10% stool sample dilution.

Sample preparation and sequencing on the Ion Torrent PGM platform. The high 18S real-time RT-PCR C_q values (≈34) observed in the load of the first two samples treated with nucleases (in comparison with nontreated sample 3; $C_q \approx 25$) indicates that Benzonase degraded the nucleic acids present in the sample (Fig. 2). Comparing the 18S C_q values of the load and elution fractions in these two samples, it can be concluded that one-step elution did not further clean the sample from eukaryotic NA, although it allowed a single order of magnitude concentration of the viruses in sample 2 (1 virus per field of 400 mesh EM grid in the load compared to 12 to 22 in the eluted fractions). In sample 1 (stool suspension), the amount of virus was below the limit of detection (LOD) of the electron microscopy already in the load, and the presence of viruses was only detected by conventional RT-PCR (Fig. 2), in which a certain concentration was again observed when comparing the intensities of the agarose gel bands from load and elution fractions. Sample 3 showed the largest amount of virus in the eluted fractions (up to 59 viruses per field of 400 mesh EM grid), and it also allowed concentration of the viruses from 4 viruses per field to up to 59. Moreover, the gradient allowed more efficient separation of the virus-rich fraction, as seen from the chromatogram (Fig. 2) and from the 18S assay C_q values, which increased from 25 up to 35 in fraction 15. In this case, nonviral NA was removed physically from the sample, so a lower risk for the

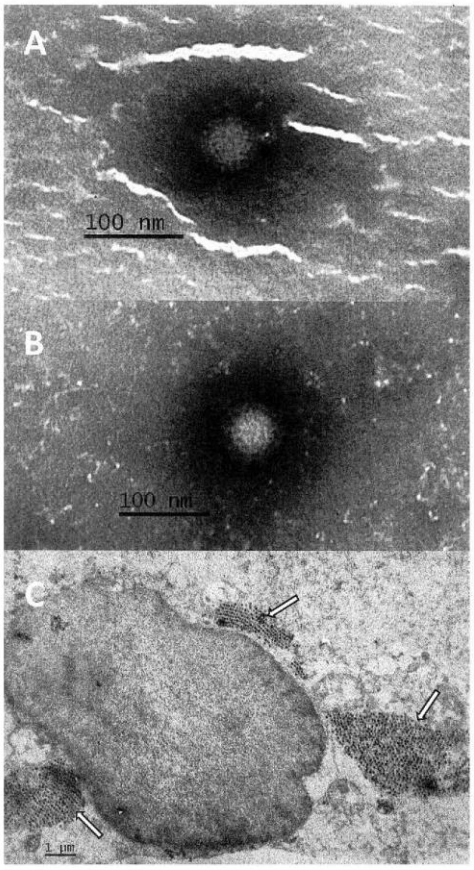


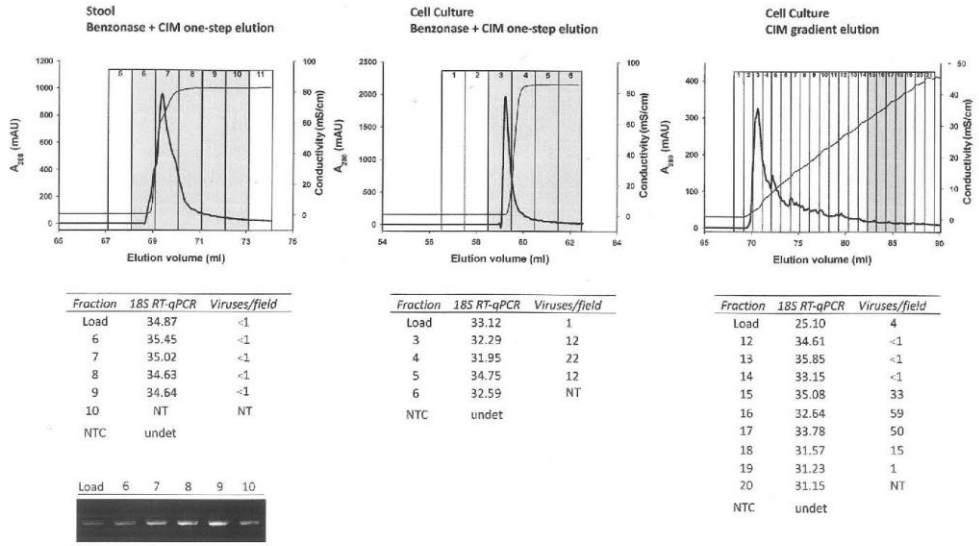
FIG 1 Electron micrographs of reoviruses in stool suspension (A) (magnification, 100,000×) cell culture supernatant (B) (magnification, 100,000×), and ultrathin section of LLC-MK2 cells (C) (magnification, 10,000×) infected with the SI-MRV01 orthoreovirus strain. Arrows in Fig. 1C indicate reovirus particles.

presence of small interfering fragments is expected in comparison to the Benzonase-treated ones.

The joint data set from all three sequencing runs consisted of 861,215 reads. The average depth of coverage across all segments of the virus genome was 257.41×, with a standard deviation (SD) of 117.29×. After removal of duplicate reads, the average depth of coverage amounted to 101.88× (SD, 25.13×). The comparison of the sequencing outputs of the three samples pretreated differently is summarized in Table 1. In the case of sample 1, only 4.91% of reads aligned to the final consensus sequence used as a reference for comparisons. Much higher enrichment for viral sequences was seen in the case of samples isolated from tissue culture: 33.4% of reads aligned to the reference in case of sample 2. In the case of sample 3 with gradient elution, an even higher proportion of sequences aligned to the reference (40.14% of reads). Of all three samples, sample 3 had the highest N_{50} value for the *de novo* assembly of reads, and the contigs produced covered almost the entire virus genome. In addition, the percentage of high-quality reads that map to the *M. mulatta* genome (source species for cell culture cells) is much lower in sample 3 (2.5%) than in sample 2 (30.8%) (Table 1), indicating that the elution gradient was more efficient at removing background NA than the Benzonase treatment combined with single-step elution. Since the total number of reads was a lot higher for sample 2 than for sample 3, more data and better depth of coverage were obtained from the former. However, increased depth in the case of sample 2 was shown to be a consequence of a high number of duplicate reads (Table 1). Due to the small size of the viral genome, a very good average coverage depth was achieved even on the smallest PGM sequencing chip.

Sequence analysis. Nucleotide and deduced amino acid sequences were obtained and further analyzed. Following BLAST of the nucleotide sequence, the newly sequenced Slovenian reovirus isolate SI-MRV01 showed the highest nucleotide sequence identity with reovirus strain 342/08, isolated from an insectivorous bat in Germany, in all genome segments except S2, in which a slightly lower nucleotide identity was observed (see Table S1 in the supplemental material). The nucleotide identity with the 342/08 strain varied from 93.8% for the S2 segment to 99.0% for the L2 segment. The amino acid sequence identity consequently also showed the highest identity with the 342/08 strain in all genome segments (98.4 to 99.7%) (see Table S1). In addition, phylogenetic analysis confirmed the highest relationship of Slovenian human reovirus isolate SI-MRV01 with German bat reovirus isolate 342/08. Both isolates clustered together in a separate branch in all genome segments except S1 and S2 (Fig. 3; see Fig. S1 to S7 in the supplemental material). For the S1 segment, the Slovenian SI-MRV01 strain was found in a specific bat cluster formed within the mammalian reovirus serotype 3 group. For this segment, the highest identity was shown with the bat reovirus characterized recently in Italy (12). In the S2 segment, SI-MRV01 did not show relatedness to any specific strain but was clustered in the serotype 3 group, together with porcine strain SC-A, MRV-HLJ/2007, civet strain MPC/04, and bat strain 342/08. According to the results of the SimPlot analysis, there are some indications of possible genome reassortments. Most indications of a reassortment event were found for the L2 segment, comparing the SI-MRV01 strain with the porcine GD-1, 729, and civet MPC/04 strains (Fig. 4).

The SI-MRV01 strain's deduced amino acid sequence was analyzed for indicative amino acid positions related to pathogenesis. In previous studies, two amino acid positions, 350D and 419E, in the σ1 protein (S1 gene) were shown to be indicative of neurotropism (31, 32). However, in the SI-MRV01 strain, those sites were 350I and 419E. The significance of such an amino acid composition is not known and needs to be examined. It was also shown that 249I is required for protease resistance of the σ1 protein, which enables efficient viral spread and replication (33) and was also found in the SI-MRV01 strain. The S1 segment is also important in fusogenic reoviruses, since it encodes the fusion protein FAST (fusion-associated transmembrane protein) (34). However, the Slovenian SI-MRV01 strain does not have the FAST coding region. In the cell culture, no syncytia were observed, which is in concordance with molecular findings of the absent FAST coding region.



Fraction	18S RT-qPCR	Viruses/field
Load	34.87	<1
6	35.45	<1
7	35.02	<1
8	34.63	<1
9	34.64	<1
10	NT	NT
NTC	undet	

Load 6 7 8 9 10

Fraction	18S RT-qPCR	Viruses/field
Load	33.12	1
3	32.29	12
4	31.95	22
5	34.75	12
6	32.59	NT
NTC	undet	

Fraction	18S RT-qPCR	Viruses/field
Load	25.10	4
12	34.61	<1
13	35.85	<1
14	33.15	<1
15	35.08	33
16	32.64	59
17	33.78	50
18	31.57	15
19	31.23	1
20	31.15	NT
NTC	undet	

FIG 2 Elution peaks obtained in the chromatographic runs involving samples 1 (stool, Benzonase, CIM one-step elution), 2 (cell culture, Benzonase, CIM one-step elution), and 3 (cell culture, CIM gradient elution). mAU, milliabsorbance units. Fraction analysis for 18S rRNA (as indicative of eukaryotic RNA presence) and reovirus EM detection is shown below the graphs. In the case of sample 1, the presence of reovirus had to be confirmed by classic quantitative RT-PCR (RT-qPCR) (shown to the left below). NTC, negative template control; NT, not tested; undet, undetermined.

DISCUSSION

The case described in this work was presented with acute gastroenteritis requiring hospitalization in a 17-month-old boy, and MRVs were found in a stool sample as the only possible causative agent. There is a small possibility, though, of some other pathogen being involved in the disease (i.e., some parasites that were not tested on this occasion). However, the high concentration of MRV shedding in the stool of this child indicates acute infection and efficient replication of MRV in the gut. No additional stool samples were available for analysis to check the excretion's dynamic and possible asymptomatic carriage of reoviruses. Also, no acute-phase serum sample was stored for additional tests to serologically confirm the infection. As MRV infections are common in humans and the seroprevalence in early childhood from 1 to 5 years of age was shown to be 8 to 50% (35), we believe that serum reactivity to other MRVs besides the isolate SI-MRV01 should also be tested to

TABLE 1 Quality and quantity of nucleotide sequences obtained from Ion Torrent sequencing for the three samples tested

Parameter	Result for[a]: Sample 1	Sample 2	Sample 3
De novo assembly			
N_{50}	155	88	236
No. of contigs	885	2,556	371
No. (total length) of regions not covered by contigs[b]	73 (4,787 nt)	96 (6,258 nt)	9 (89 nt)
Mapping to new *Orthoreovirus* consensus sequence			
% of reads mapped	4.91	33.4	40.14
Avg coverage	11.47× (SD, 5.49×)	151.77× (SD, 83.42×)	119.07× (SD, 46.06×)
Avg coverage excluding duplicates	10.08× (SD, 4.35×)	39.34× (SD, 12.11×)	76.74× (SD, 21.17×)
No. (total length) of zero coverage regions (all)	6 (17 nt)	4 (5 nt)	4 (5 nt)
Mapping to *Macaca mulatta* genome[c]			
% of reads mapped	0.03	30.83	2.54

[a] Sample 1, stool, Benzonase, CIM one-step elution; sample 2, cell culture, Benzonase, CIM one-step elution; sample 3, cell culture, no Benzonase, CIM gradient elution.
[b] After they had been mapped to the new *Orthoreovirus* consensus sequence.
[c] Only reads longer than 25 bp.

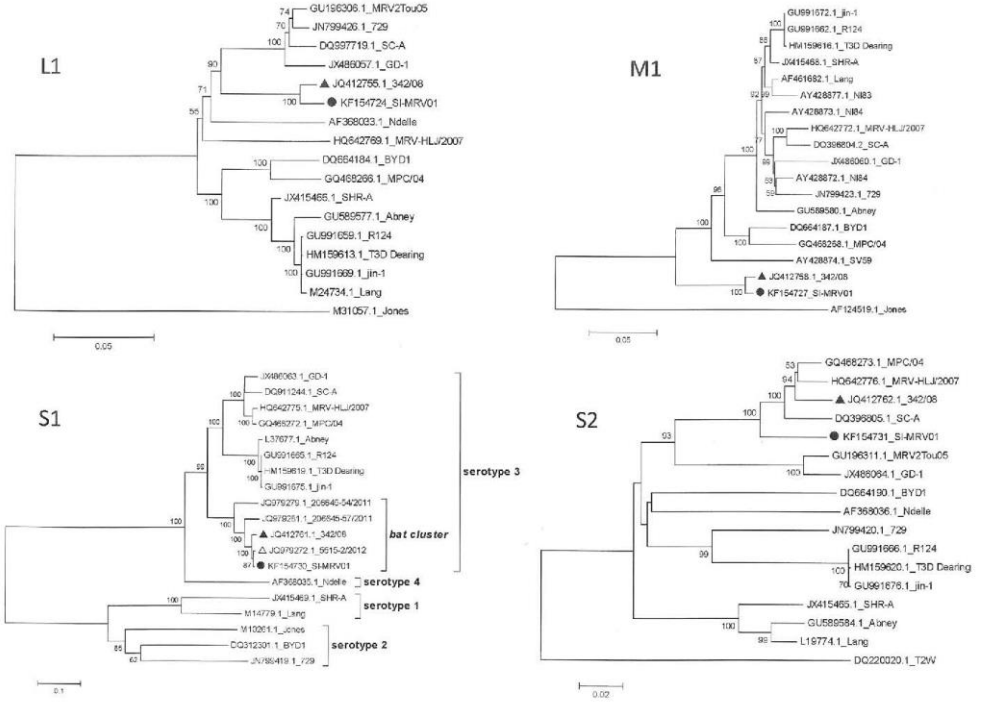


FIG 3 Phylogenetic analysis of the L1, M1, S1, and S2 genome segments for the Slovenian strain and most related whole-genome strains from GenBank. The neighbor-joining algorithm was used for the construction of the phylogenetic tree with bootstrap values of 1,000 replicates shown at the branches. The scale bar represents the *p*-distance. ●, Slovenian MRV isolate SI-MRV01; △, Italian MRV strain identified in a bat; ▲, German MRV isolate 342/08 identified in a bat.

exclude possible cross-reactivity. Although there was no serological proof of MRV infection causing gastroenteritis, we have presented data that demonstrate MRV as the possible causative agent of the disease.

It is well known that the disadvantage of electron microscopy is its low sensitivity. For successful detection of viruses with electron microscopy, the concentration of viral particles should be higher than 10^6 per milliliter of suspension (36). Bearing in mind the dilution factor of 10 (when preparing a 10% stool suspension), this means that the MRV concentration in the patient's stool sample was at least of 10^7 viruses per milliliter or even higher. Moreover, the viruses found in the stool sample were infective, since they were successfully propagated in a cell culture. The high concentration of reoviruses in stool sample was confirmed also by the theoretically calculated $TCID_{50}$/ml, which was as high as 2.26×10^7. All these facts suggest high replication activity of MRV in the patient's gut.

After the MRV isolate was cultured and genotyped, the parents were contacted with regard to the child's potential animal exposure. The family lived in a renovated house in a village close to the city, where no bats had been observed. The only animal that had close contact with the child was a dog at the grandparents' house, but the animal was predominantly kept indoors. At the time of the child's illness, none of the family members experienced similar symptoms. The parents stated that, at the time of the disease, the child was known to ingest nonfood items (e.g., mild form of pica). According to the reovirus inactivation studies, showing their persistent stability in environment, indirect infection through contaminated surfaces could also be possible (2, 17, 37). The source of infection therefore remains unknown. Unfortunately, no bat stool samples from the child's residence were available for analysis to investigate the possible source of infection. It would be interesting in the future to screen bats in Slovenia for reoviruses, in order to have a clear picture of reovirus molecular epidemiology and to link possible zoonotic transmissions, such as the one described in our case.

MRVs were traditionally believed to be causative agents of mild respiratory and enteric infections, without significant clinical impact (1). However, in the last decade or two, increasing reports on severe human infections with reoviruses have been published, including central nervous system involvement (10, 38, 39). The pathogenesis of reoviruses is not yet elucidated in detail, but some parameters influencing tissue tropism, efficient spread within the host, and severe outcomes of infection are already known (31–33).

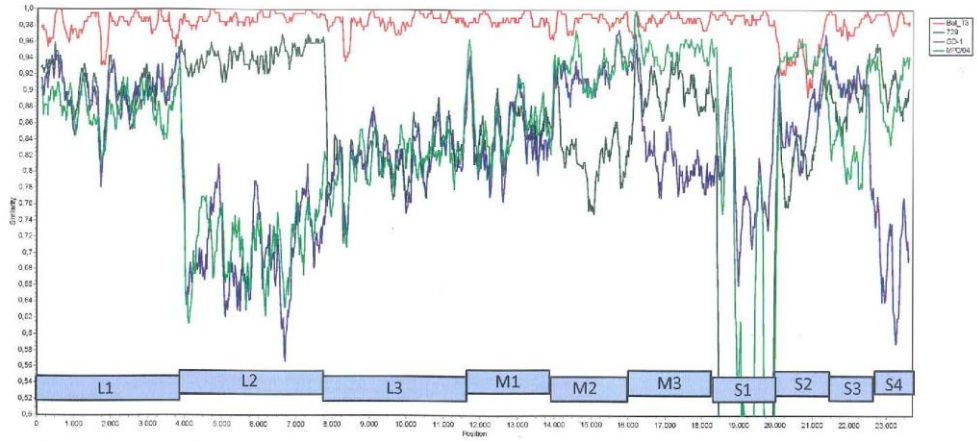


FIG 4 Similarity plot analysis of the whole-genome nucleotide sequence comparing the Slovenian orthoreovirus strain with some of the most related strains from GenBank.

It is thus important for diagnostic personnel to be aware of a possible reovirus etiology and to have tools to prove this infection. However, any reovirus-positive result should be interpreted carefully, with a broad spectrum of cotested pathogens and clinical picture of the patient. Whether or not zoonotic transmission of reoviruses is one of the possible factors associated with a severe clinical outcome is not yet clear. Nevertheless, there are reports in the literature of bat reoviruses in humans with severe clinical manifestations. Those reovirus isolates, though, all clustered to *Pteropine reovirus* species and not to MRV species (6–8). This study is the first description of a bat MRV found in humans. Probably MRVs are widely distributed among bat species, since they have been found in insectivorous bats in Germany and in Italy (11, 12).

Whole-genome sequence comparison of our strain to MRVs available in GenBank clearly shows that it is most closely related to MRVs found in bats in Germany. This was supported by high nucleotide and amino acid identities, phylogenetic analysis of separate genome segments, and whole-genome sequences, including SimPlot analysis. Genome reassortment is indicated by a similarity plot of reovirus genome sequences most closely related to the Slovenian SI-MRV01 strain (Fig. 4). A sharp decrease in nucleotide sequence identities for some of the genome segments was observed, like the L2 and S1 segments, comparing SI-MRV01 to the GD-1 and MPC/04 strains, and the S2 segment, comparing SI-MRV01 to the 342/8 strain. Unfortunately, no other strains with higher nucleotide identity were found in GenBank to explain this diversity or possible reassortment events.

Finding an MRV in the stool sample of a hospitalized child with diarrhea was surprising for us, since no such case had previously been described in our laboratory. In the cell culture unit, no previous work with reoviruses was performed, there was no reovirus nucleic acid isolated in the molecular laboratory, nor had a reovirus genome amplification project been performed. Thus, the possibility of sample contamination is negligible. In addition to antigen detection tests and real-time RT-PCR assays for the detection of the most common viral causes of gastroenteritis, electron microscopy is still an important diagnostic tool routinely used in our laboratory. It is possible with the electron microscope to detect various viruses associated with rare cases, not included in specific antigen and/or molecular testing procedures. It is a "catch-all" method, providing a broad-spectrum analysis of examined samples (40). For rapid and accurate diagnostics, combination of classical and new techniques in diagnostic virology is often needed. This study demonstrates a good case of such an interplay. Moreover, an improvement in sample preparation for successful downstream application was introduced with virus enrichment using CIM monolithic support. It was shown that gradient elution, used in sample 3 without Benzonase treatment, in principle allows more detailed separation between the eluted molecules, including viruses and nucleic acids. In addition, virus particles were concentrated in specific fractions after chromatographic separation. This was confirmed also by sequencing results. In sample 3, the total number of reads was the lowest, but the percentage of orthoreovirus reads was the highest.

Summarizing the results of our study, there are two major outputs of this work. First, the detection of an MRV with high identity to MRV isolates found in European bats is an important indicator of bat-to-human interspecies transmission, as was already speculated by German and Italian groups (11, 12). The real epidemiological situation regarding MRVs in bats and their transmission to other animal species should be further investigated in order to understand the full zoonotic potential and pathogenesis of these viruses. The second output is a tool for improving the sample pretreatment in the search for enrichment in viral particles and nucleic acid. The pretreatment step is essential in NGS application, in which the target nucleic acid is expected to be present with high background. The described CIM chromatographic approach demonstrates an excellent method combining concentration and purification of the target, providing a new potential shortcut for NGS application in diagnostic clinical virology.

ACKNOWLEDGMENTS

We thank Irena Šest for technical support with electron microscopy, diagnostics, and sample preparation and Snežana Kramar for assistance with cell culture propagation.

This work was financially supported by the Slovenian Research Agency (contracts no. L2-4314 and J3-4252).

REFERENCES

1. Schiff LA, Nibert ML, Tyler KL. 2007. Orthoreoviruses and their replication, p 1854–1915. *In* Knipe DM, Howley PM, Griffin DE, Lamb RA, Martin MA, Roizman B, Straus SE (ed), Fields virology, 5th ed. Lippincott Williams & Wilkins, Philadelphia, PA.
2. Katz BD, Margolin AB. 2007. Inactivation of hepatitis A HM-175/18f, reovirus T1 Lang and MS2 during alkaline stabilization of human biosolids. J. Appl. Microbiol. 103:2225–2233.
3. Day JM. 2009. The diversity of the orthoreoviruses: molecular taxonomy and phylogenetic divides. Infection, genetics and evolution. J. Mol. Epidemiol. Evol. Genet. Infect. Dis. 9:390–400.
4. Sabin AB. 1959. Reoviruses. Science 130:1387–1389.
5. Ouattara LA, Barin F, Barthez MA, Bonnaud B, Roingeard P, Goudeau A, Castelnau P, Vernet G, Paranhos-Baccala G, Komurian-Pradel F. 2011. Novel human reovirus isolated from children with acute necrotizing encephalopathy. Emerg. Infect. Dis. 17:1436–1444.
6. Chua KB, Voon K, Crameri G, Tan HS, Rosli J, McEachern JA, Suluraju S, Yu M, Wang LF. 2008. Identification and characterization of a new orthoreovirus from patients with acute respiratory infections. PLoS One 3:e3803. doi:10.1371/journal.pone.0003803.
7. Chua KB, Voon K, Yu M, Keniscope C, Abdul Rasid K, Wang LF. 2011. Investigation of a potential zoonotic transmission of orthoreovirus associated with acute influenza-like illness in an adult patient. PLoS One 6:e25434. doi:10.1371/journal.pone.0025434.
8. Cheng P, Lau CS, Lai A, Ho E, Leung P, Chan F, Wong A, Lim W. 2009. A novel reovirus isolated from a patient with acute respiratory disease. J. Clin. Virol. 45:79–80.
9. Hermann L, Embree J, Hazelton P, Wells B, Coombs RT. 2004. Reovirus type 2 isolated from cerebrospinal fluid. Pediatr. Infect. Dis. J. 23:373–375.
10. Tyler KL, Barton ES, Ibach ML, Robinson C, Campbell JA, O'Donnell SM, Valyi-Nagy T, Clarke P, Wetzel JD, Dermody TS. 2004. Isolation and molecular characterization of a novel type 3 reovirus from a child with meningitis. J. Infect. Dis. 189:1664–1675.
11. Kohl C, Lesnik R, Brinkmann A, Ebinger A, Radonic A, Nitsche A, Muhldorfer K, Wibbelt G, Kurth A. 2012. Isolation and characterization of three mammalian orthoreoviruses from European bats. PLoS One 7:e43106. doi:10.1371/journal.pone.0043106.
12. Lelli D, Moreno A, Lavazza A, Bresaola M, Canelli E, Boniotti MB, Cordioli P. 2013. Identification of mammalian orthoreovirus type 3 in Italian bats. Zoonoses Public Health 60:84–92.
13. Decaro N, Campolo M, Desario C, Ricci D, Camero M, Lorusso E, Elia G, Lavazza A, Martella V, Buonavoglia C. 2005. Virological and molecular characterization of a mammalian orthoreovirus type 3 strain isolated from a dog in Italy. Vet. Microbiol. 109:19–27.
14. Lodder WJ, de Roda Husman AM. 2005. Presence of noroviruses and other enteric viruses in sewage and surface waters in The Netherlands. Appl. Environ. Microbiol. 71:1453–1461.
15. Lodder WJ, van den Berg HH, Rutjes SA, de Roda Husman AM. 2010. Presence of enteric viruses in source waters for drinking water production in The Netherlands. Appl. Environ. Microbiol. 76:5965–5971.
16. Spinner ML, Di Giovanni GD. 2001. Detection and identification of mammalian reoviruses in surface water by combined cell culture and reverse transcription-PCR. Appl. Environ. Microbiol. 67:3016–3020.
17. Irving LG, Smith FA. 1981. One-year survey of enteroviruses, adenoviruses, and reoviruses isolated from effluent at an activated-sludge purification plant. Appl. Environ. Microbiol. 41:51–59.
18. Wong S, Lau S, Woo P, Yuen KY. 2007. Bats as a continuing source of emerging infections in humans. Rev. Med. Virol. 17:67–91.
19. Wang LF, Walker PJ, Poon LL. 2011. Mass extinctions, biodiversity and mitochondrial function: are bats 'special' as reservoirs for emerging viruses? Curr. Opin. Virol. 1:649–657.
20. Lipkin WI. 2013. The changing face of pathogen discovery and surveillance. Nat. Rev. Microbiol. 11:133–141.
21. Capobianchi MR, Giombini E, Rozera G. 2013. Next-generation sequencing technology in clinical virology. Clin. Microbiol. Infect. 19:15–22.
22. Beerenwinkel N, Gunthard HF, Roth V, Metzner KJ. 2012. Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. Front. Microbiol. 3:329. doi:10.3389/fmicb.2012.00329.
23. Gutierrez-Aguirre I, Steyer A, Banjac M, Kramberger P, Poljsak-Prijatelj M, Ravnikar M. 2011. On-site reverse transcription-quantitative polymerase chain reaction detection of rotaviruses concentrated from environmental water samples using methacrylate monolithic supports. J. Chromatogr. A 1218:2368–2373.
24. Kovac K, Gutierrez-Aguirre I, Banjac M, Peterka M, Poljsak-Prijatelj M, Ravnikar M, Mijovski JZ, Schultz AC, Raspor P. 2009. A novel method for concentrating hepatitis A virus and caliciviruses from bottled water. J. Virol. Methods 162:272–275.
25. Kageyama T, Kojima S, Shinohara M, Uchida K, Fukushi S, Hoshino FB, Takeda N, Katayama K. 2003. Broadly reactive and highly sensitive assay for Norwalk-like viruses based on real-time quantitative reverse transcription-PCR. J. Clin. Microbiol. 41:1548–1557.
26. Svraka S, van der Veer B, Duizer E, Dekkers J, Koopmans M, Vennema H. 2009. Novel approach for detection of enteric viruses to enable syndrome surveillance of acute viral gastroenteritis. J. Clin. Microbiol. 47:1674–1679.
27. Reed LJ, Muench H. 1938. A simple method of estimating fifty per cent endpoints. Am. J. Hyg. (Lond.) 27:493–497.
28. Zheng YZ, Webb R, Greenfield PF, Reid S. 1996. Improved method for counting virus and virus like particles. J. Virol. Methods 62:153–159.
29. Tamura K, Dudley J, Nei M, Kumar S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol. Biol. Evol. 24:1596–1599.
30. Lole KS, Bollinger RC, Paranjape RS, Gadkari D, Kulkarni SS, Novak NG, Ingersoll R, Sheppard HW, Ray SC. 1999. Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. J. Virol. 73:152–160.
31. Bassel-Duby R, Spriggs DR, Tyler KL, Fields BN. 1986. Identification of attenuating mutations on the reovirus type 3 S1 double-stranded RNA segment with a rapid sequencing technique. J. Virol. 60:64–67.
32. Kaye KM, Spriggs DR, Bassel-Duby R, Fields BN, Tyler KL. 1986. Genetic basis for altered pathogenesis of an immune-selected antigenic variant of reovirus type 3 (Dearing). J. Virol. 59:90–97.
33. Chappell JD, Barton ES, Smith TH, Baer GS, Duong DT, Nibert ML, Dermody TS. 1998. Cleavage susceptibility of reovirus attachment protein sigma1 during proteolytic disassembly of virions is determined by a sequence polymorphism in the sigma1 neck. J. Virol. 72:8205–8213.
34. Shmulevitz M, Duncan R. 2000. A new class of fusion-associated small transmembrane (FAST) proteins encoded by the non-enveloped fusogenic reoviruses. EMBO J. 19:902–912.
35. Tai JH, Williams JV, Edwards KM, Wright PF, Crowe JE, Jr, Dermody TS. 2005. Prevalence of reovirus-specific antibodies in young children in Nashville, Tennessee. J. Infect. Dis. 191:1221–1224.
36. Hazelton PR, Gelderblom HR. 2003. Electron microscopy for rapid diagnosis of infectious agents in emergent situations. Emerg. Infect. Dis. 9:294–303.
37. Ward RL, Ashley CS. 1978. Heat inactivation of enteric viruses in dewatered wastewater sludge. Appl. Environ. Microbiol. 36:898–905.
38. Tyler KL. 1998. Pathogenesis of reovirus infections of the central nervous system. Curr. Top. Microbiol. Immunol. 233:93–124.
39. Johansson PJ, Sveger T, Ahlfors K, Ekstrand J, Svensson L. 1996. Reovirus type 1 associated with meningitis. Scand. J. Infect. Dis. 28:117–120.
40. Biel SS, Gelderblom HR. 1999. Diagnostic electron microscopy is still a timely and rewarding method. J. Clin. Virol. 13:105–119.

2.2 Complete Genome Sequences of New Divergent Potato Virus X Isolates and Discrimination Between Strains in a Mixed Infection Using Small RNAs Sequencing Approach

Denis Kutnjak, Rocio Silvestre, Wilmer Cuellar, Wilmer Perez, Giovanna Müller, Maja Ravnikar, Jan Kreuze

Virus Research, 2014, 191: 45-50. DOI: 10.1016/j.virusres.2014.07.012

We used small RNA deep sequencing to characterise genome sequences of several isolates of potato virus X, collected in the Andean region of South America. Phylogenetic analyses of complete viral genomes showed a presence of a new strain in this region (SA-CIP) as well as isolates belonging to the Eurasian group. Thus, results confirmed a previously postulated hypothesis that diversity of PVX is underestimated for this region. Further, we were able to distinguish between two strains of the virus in a mixed infection (SA-CIP and Eurasian), which is the first report of the use of small RNA deep sequencing for such an application.

The PhD candidate analyzed Illumina sequencing reads and constructed complete or near complete consensus viral genomes of the isolates, detecting also a mixed infection of two strains of the virus in one of the samples. He performed phylogenetic analyses of complete and partial viral genomes and designed PCR primer pairs for quick discrimination of different PVX strains. He drafted the manuscript.

Virus Research 191 (2014) 45–50

Contents lists available at ScienceDirect

Virus Research

journal homepage: www.elsevier.com/locate/virusres

Short communication

Complete genome sequences of new divergent potato virus X isolates and discrimination between strains in a mixed infection using small RNAs sequencing approach

Denis Kutnjak[a], Rocio Silvestre[b], Wilmer Cuellar[b,1], Wilmer Perez[b], Giovanna Müller[b], Maja Ravnikar[a], Jan Kreuze[b,*]

[a] Department of Biotechnology and Systems Biology, National Institute of Biology, Ljubljana, Slovenia
[b] International Potato Center (CIP), Lima, Peru

ARTICLE INFO

Article history:
Received 25 April 2014
Received in revised form 9 July 2014
Accepted 12 July 2014
Available online 19 July 2014

Keywords:
Andes
Next generation sequencing
Phylogeny
Potato virus X
Small RNAs

ABSTRACT

Potato virus X (PVX; genus *Potexvirus*, family *Alphaflexiviridae*, order *Tymovirales*) is one of the most widespread and intensively studied viruses of potato. However, little is known about its diversity in its likely center of radiation, the Andean region of South America. To fill this gap, the strategy of Illumina deep sequencing of small RNAs was used to obtain complete or near complete genome sequence of PVX from 5 symptomatically infected greenhouse and 3 field samples (*Solanum tuberosum*) from Peru. PVX sequences determined in this study were assigned into three different phylogenetic groups of isolates. Notably, a complete genome sequence of a representative of a new PVX phylogenetic lineage was obtained, which shows a high level of sequence dissimilarity to other completely sequenced isolates (~17%). The new PVX genotype was detected in greenhouse and field samples. One of the field samples was infected with the mixture of two PVX strains, which were efficiently discriminated using small RNA sequencing approach. The study confirms the utility of small RNAs deep sequencing for successful viral strain differentiation and discovery of new viral strains and indicates a high diversity of PVX in the Andean region of South America, a pattern which may be expected also for other potato pathogens.

© 2014 Elsevier B.V. All rights reserved.

Isolates of *Potato virus X* (PVX), the type member of the genus *Potexvirus* (family: *Alphaflexiviridae*), are present world-wide in potato growing areas, infecting a wide range of hosts, notably within the *Solanaceae* family (King et al., 2011). It has been investigated intensively and thus become one of the most important models for studies of plant–virus interactions (Scholthof et al., 2011). Alone it is usually moderately pathogenic but it can cause a significant economic loss in synergistic co-infection with some potyviruses, especially potato virus Y (Khurana and Singh, 1988; Vance, 1991). PVX virions are flexuous filaments, 470–580 nm long. They contain a single positive sense ssRNA molecule (~6400 bp) with five open reading frames (ORFs) coding for RNA dependent RNA polymerase (RdRp), three proteins involved in cell-to-cell movement – the triple gene block (TGB) – and a coat protein (CP) (King et al., 2011).

According to the hypersensitive reaction induced by N_b and N_x resistance genes on potato (*Solanum tuberosum* L.) cultivars carrying them, PVX strains can be divided into four strain groups (Cockerham, 1955; Santa Cruz and Baulcombe, 1993). Group 1 cannot overcome either gene, group 2 overcomes N_x, group 3 N_b, and group 4 overcomes both genes. Furthermore, some isolates (PVX-HB, PVX-MS) overcome both N_b and N_x and also the extreme resistance gene R_x (Moreira et al., 1980; Jones, 1985; Querci et al., 1995; Feigelstock et al., 1995). "Gene for gene" interactions of these three host resistance genes with specific determinants in virus genomes have been determined (reviewed in Malcuit et al., 2000) in coat protein (for N_x and R_x) and the 25 kDa movement protein region (for N_b). There is no evidence for recombination between different PVX strains in nature. Thus, the acquisition of virulence determinants (or loss of virulence ones) can only be explained as a result of independent evolution (Malcuit et al., 2000).

Within PVX there is a phylogeographic split separating majority Eurasian (clade I) and South American (clade II) groups of isolates (Yu et al., 2008, 2010; Cox and Jones, 2010). However, some

* Corresponding author at: International Potato Center (CIP), Avenida La Molina 1895, La Molina Apartado Postal 1558, Lima, Peru. Tel.: +51 1 349 6017x3054; fax: +51 1 317 5326.
E-mail address: j.kreuze@cgiar.org (J. Kreuze).
[1] Current affiliation: International Center for Tropical Agriculture (CIAT), Cali, Colombia.

http://dx.doi.org/10.1016/j.virusres.2014.07.012
0168-1702/© 2014 Elsevier B.V. All rights reserved.

Table 1
Information about the samples included in the study.

Sample number	Locality/source (year of collection)	Host plant	Disease symptoms	No. of pooled individual plants	PVX group	PVX isolate name	NCBI GenBank accession number
GAF214	Samples from CIP greenhouse infection, Lima, Peru (2011)	Potato	Chlorosis	3	II-3	SA-CIP	KJ534601
GAF215		Potato	Albinism	4	II-3	SA-CIP	KJ534601
GAF216		Datura	Chlorosis	3	II-3	SA-CIP	KJ534601
GAF217		Datura	Albinism chlorosis	3	II-3	SA-CIP	KJ534601
GAF218		Datura	Mottling	4	II-3	SA-CIP	KJ534601
GAF224	Field samples: Quichas, Cerro de Pasco, Peru (2011)	Potato	Leaf yellowing	1	II-2	GAF224	KJ534602
GAF225		Potato	Leaf yellowing	1	II-3 I	GAF225	KJ534603 KJ534604
GAF226		Potato	Leaf yellowing	1	II-2	GAF226	KJ534605

partially sequenced European isolates also cluster within clade II (South American group) and some American isolates within clade I (Eurasian group) (Cox and Jones, 2010). Clade II can be further split into two sub-clades: II-1 and II-2 (Cox and Jones, 2010). There is no direct link between resistance grouping and phylogenetic placement of the isolates, nevertheless, strain groups 1 and 3 have exclusively been identified in phylogenetic clade I, and strain group 2 in clade II, whereas group 4 occurs in both clades (Cox and Jones, 2010).

Current sequencing techniques (next generation sequencing – NGS) have dramatically broadened possibilities for virus diagnostics, enabling the rapid discovery of new viruses and virus strains. The use of small RNA deep sequencing for complete plant viral sequence assembly (sRSA: small RNA sequencing and assembly) and the discovery of novel plant viruses was first introduced by Kreuze et al. (2009) and since then successfully implemented in many studies to identify novel viruses (Untiveros et al., 2010; Cuellar et al., 2011; Zhang et al., 2011; Bi et al., 2012; Fuentes et al., 2012; Kashif et al., 2012; Li et al., 2012; Loconsole et al., 2012a, 2012b; Sela et al., 2012; Wu et al., 2010; De Souza et al., 2013; Hwang et al., 2013; Kreuze et al., 2013; Roy et al., 2013) and viroids (Li et al., 2012). Here we employ sRSA to obtain complete or near complete (99% or more) genome sequences of PVX from greenhouse and field infected plants from Peru, identify a new phylogenetic lineage and obtain accurate assembly of two genetic variants in a mixed infected plant. Diagnostic primers were designed enabling rapid identification of PVX lineages.

Five leaf samples from greenhouse (*S. tuberosum* and *Datura stramonium* L.) and 3 field samples (*S. tuberosum*) were included in the study. Each greenhouse sample consisted of a bulk of several plants exhibiting similar symptoms (details given in Table 1); field samples were collected in plastic bags and kept on ice until processing in the lab during the same day.

Total RNA was isolated from 1 g of fresh leaf tissue using Trizol reagent (Invitrogen), run in 4% agarose gel and small RNAs were recovered from the gel using Freeze 'n Squeeze DNA gel extraction spin columns (Bio-Rad, Hercules, CA, USA). Isolated small RNAs were sent to Fasteris Life Sciences SA (Plan-les-Ouates, Switzerland) for library preparation and sequencing on Illumina Hiseq2000 platform in 50 bp single end read mode (Kreuze et al., 2009).

The analysis of obtained data was performed using the following work-flow: (1) reads of length 21–24 bp were assembled de-novo using Velvet v0.6.04 (Zerbino and Birney, 2008) and AssemblyAssembler 1.4 script (Jacob Crawford, Cornell University); (2) assembled contigs were compared for similarity against all viral sequences deposited in NCBI GenBank database using BLASTn and BLASTx; (3) the complete genome sequence of PVX isolate, which showed the highest similarity with the longest matching contig, was used as a reference for mapping reads (21–24 bp) using MAQ (http://maq.sourceforge.net) easyrun command; (4) using Seqman (DNASTAR, Madison, WI, USA) or ContigExpress (Vector NTI, Invitrogen, Carlsbad, CA, USA) the MAQ output consensus sequence and the Velvet assembled contigs were further assembled together to obtain new consensus sequence with improved coverage of the viral genome. In order to refine the consensus sequence of the viral genome, steps 3 (with the new consensus sequence as a reference) and 4 were iterated several times.

The consensus sequences of PVX obtained by analysis of the 5 greenhouse samples were significantly different from sequences of any other PVX isolate available in NCBI GenBank. They were partial, but identical to each other in extensively overlapping regions, and therefore, in this case, we generated a new consensus sequence assembling all 5 samples into one consensus sequence covering the complete genome. The new consensus sequence was then used in the analysis from step 3 onward for each of the plants separately again.

For sample GAF225, *de-novo* assembly of short reads resulted in contigs corresponding either to cluster I or cluster II-3, which implied mixed strains infection. Mapping the reads to corresponding reference genome sequences (representative of cluster I and II-3) enabled the generation of near complete genome sequence (more than 99%) for both strains, confirming the mixed infection. Several primer pairs were designed to allow for a validation of genome sequences from all of the samples (primers used and their annealing positions in the genome are given in Table S1). For cDNA synthesis 1 μg of total RNA and 250 ng/μl of random hexamer primers were used in a total reaction mix of 20 μl using 200 U of M-MLV reverse transcriptase (Invitrogen). After incubation at 37 °C for 50 min, the reaction was diluted 1:10 with sterile water and 5 μl were used for PCR. The PCR was performed using GoTaq® DNA Polymerase (Promega, Madison, WI, USA) in a volume of 20 μl with a final primer concentration of 0.5 μM each. The PCR protocol consisted of 5 min initial denaturation at 94 °C, followed by 35 cycles of 94 °C for 30 s, 55 °C for 30 s and 72 °C between 30 s to 1 min (depending on the fragment size), with final extension at 72 °C for 10 min. PCR fragments were visualized in 1% agarose gels stained with GelRed (Biotium, Hayward, CA, USA). The PCR fragments amplified for each isolate were purified using the High Pure PCR product purification kit (Roche, Indianapolis, IN, USA), then cloned into pGEM-T easy (Promega) following standard procedures and transformed into *Escherichia coli* (DH 5α). Samples were sent for sequencing to Macrogen (Seoul, Korea).

The alignment of obtained viral genome sequences was performed using ClustalW algorithm in MEGA 6 package (http://www.megasoftware.com), including also 22 complete

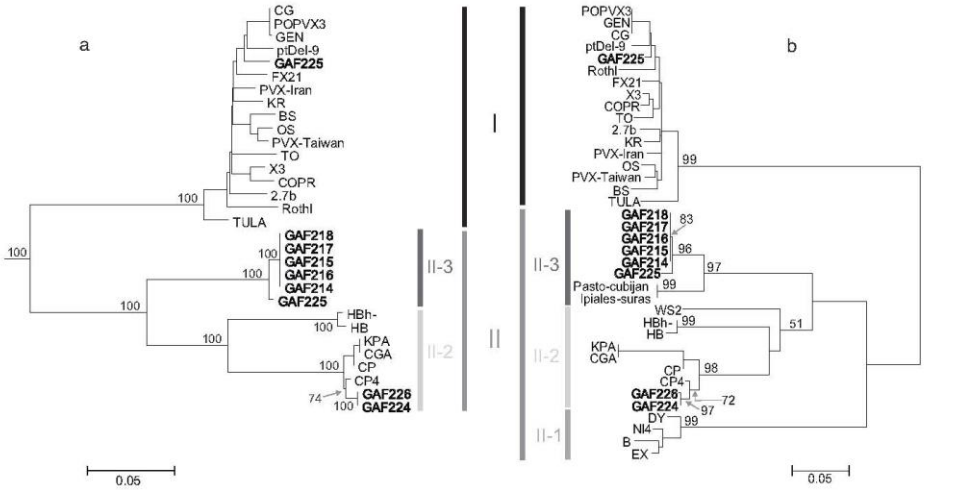


Fig. 1. Phylograms of PVX isolates constructed from (a) complete genome sequence alignment including outgroup and (b) partial coat protein sequence alignment (outgroup not included); numbers on branches represent the statistical (bootstrap) support of major branches. The phylogenetic trees were constructed using maximum-likelihood algorithm. Naming of the isolates follows the ones used in Tables 1 and 2. The samples from this study are bolded. The roman numbers designate recognized phylogenetic clusters within PVX.

genome sequences of PVX available from GenBank database and 20 other potexviruses (Table S2), used as an outgroup. Phylogenetic analyses were conducted using Mega 6 package. GTR + G + I substitution model was determined to fit data best and was thus used for construction of phylogenetic tree using maximum-likelihood algorithm. 1000 bootstrap pseudo-replicates were used for statistical testing. Separate alignment of partial coat protein sequence (~500 bp) was performed and analyzed in the same manner (without an out-group, K2 + G substitution model) to allow the inclusion of greater number of isolates from the South American cluster (7 more) in the comparison.

Nucleotide identities of PVX isolates included in the phylogenetic analyses were calculated using Bioedit (Hall, 1999). Recombination Detection Program (RDP3: Martin et al., 2010) was used to detect occurrence of any recombination events between PVX genotypes using the previously aligned complete genome sequences and the seven available algorithms within the program. The similarity between the new strain and other two phylogenetic groups of isolates with completely sequenced genomes (I, II-2) were plotted with Simplot (Lole et al., 1999) using Kimura (2-parameter) distance model (Window: 200 bp, Step: 20 bp, Transition/Transversion ratio: 2). The deduced coat protein sequence of the new PVX genotype reconstructed in this study was inspected to ascertain whether it possessed determinants which were suggested to be important for overcoming the resistance genes N_x (14 amino acid positions: Santa Cruz and Baulcombe, 1995), N_b (Malcuit et al., 1999) or R_x (Querci et al., 1995).

On the basis of the alignment of all available PVX complete genome sequences (31 isolates) specific RT-PCR primer pairs were designed to allow rapid discrimination between different phylogenetic groups of PVX isolates detected in this study. The specific annealing of all primers to PVX was confirmed *in silico* with the Primer-BLAST tool searching NCBI nt database. Primer pairs sequence information and annealing position in the corresponding genotype are given in Table S1 and Figure S3. The specificity of all the primer pairs were tested experimentally including all samples used for small RNA sequencing. RNA isolated from healthy potato and datura plants were used as negative controls. The PCR and visualization of amplified fragments were as described in previously for confirmation of siRNA assemblies.

Complete or near complete (99% or more) genome sequences of PVX were obtained from all of the analyzed samples using the small RNA high-throughput sequencing approach. Three to four percent of all the reads were mapped to PVX genome, accounting for an average coverage of around 100× in all of the samples. Capillary sequencing of selected parts confirmed the accuracy of the obtained sequences for all of the isolates (see Fig. 2 and Table S1 and Figure S3). The 5 isolates from the green house were identical in sequence except for a few nucleotides for which MAQ called polymorphism in some samples but not others. Sequences were submitted to GenBank and their accession numbers are provided in Table 1.

Phylogenetic analysis placed PVX isolates from this study into three different phylogenetic clusters. PVX from two samples (GAF224 and GAF226), collected from infected field plants near Quichas (Cerro de Pasco, Peru) clustered with other isolates reported from South America (Fig. 1: sub-cluster II-2). Five samples from greenhouse infection (GAF214–GAF218) contained a divergent sequence of PVX (designated SA-CIP from here on), which was found to differ by 17–22% from other completely sequenced PVX isolates (Table 2). They were identical to each-other and formed a divergent phylogenetic sub-lineage, which we designated sub-cluster II-3 (Fig. 1). The new PVX sub-lineage differed from most closely related completely sequenced isolates (from cluster II-2) in approximately 17% of the nucleotide positions (Table 2). There was no evidence for recombination events between the isolates of the different sub-groups as predicted by RDP3. The similarity plot (Fig. 2) indicates enhanced divergence between the groups in the central part of RdRp and is in accordance with the findings of Yu et al. (2010). They suggested this part of the PVX genome is under positive selection. The phylogenetic position of the strain and *in silico* inspection of virulence determinants suggest that SA-CIP is

Table 2
Information about other PVX isolates included in phylogenetic analyses and their nucleotide identity (%) to SA-CIP strain.

Isolate	NCBI GenBank accession number	Country of origin	PVX group	Nucleotide identity comparing to a SA-CIP strain (%)	
				Complete genome sequence	Partial coat protein sequence
Ipsales-suras	HQ396173.1	Colombia	II-3	x	93
Pasto-cubijan	HQ396166.1	Colombia	II-3	x	93
HBh	Z23256	Bolivia	II-2	83	83
WS2	X88786	USA	II-2	x	83
HB	X72214	Bolivia	II-2	83	82
CP4	AF172259	Peru	II-2	83	82
CP	X55802	Peru	II-2	83	82
CPKPA	M31541	Peru	II-2	83	79
CGA	M63141	Peru	II-2	83	79
X3	D00344	Netherlands	I	78	79
2.7b	HQ450388.1	USA	I	78	79
Gen	M38480	Russia	I	78	79
CG	M72416.1	Russia	I	78	79
COPR	M95516.1	UK	I	78	79
FX21	EF423572	China	I	78	79
KR	AF373782	Korea	I	78	79
Roth1	AF111193	England	I	78	79
Tula	EU021215	Russia	I	78	79
POPVX3	X05198.1	Russia	I	78	79
ptDel-9	JF430080.1	India	I	78	79
BS	AB056719	Japan	I	78	78
Iran	FJ461343	Iran	I	78	78
OS	AB056718	Japan	I	78	78
Taiwan	AF272736	Taiwan	I	78	78
TO	AB196001.1	Japan	I	78	78
B	GU384738	UK	II-1	x	76
EX	X88782	UK	II-1	x	76
DY	X88781	UK	II-1	x	76
NL4	X88785	Netherlands	II-1	x	76

x – complete genome sequence is not available.

representative of a strain group 2 (avirulent on cultivars carrying resistance gene N_b and R_x and virulent on cultivars carrying N_x). However, it has been shown, that a single amino acid residue change (in RdRp region of PVX) could determine the symptoms in tobacco plants (Kagiwada et al., 2005). Since SA-CIP sequence differs significantly from the sequence of closely related strains, with increased divergence in RdRp region, the *in silico* determination of strain group is only speculative and the biological characteristics should be further determined on test plants.

The phylogenetic analysis of a partial CP sequence (500 bp), which allowed inclusion of greater number of isolates into the comparison, showed that SA-CIP is similar (93% nucleic acids identity) to two isolates (Fig. 1b: group II-3) recently detected in a study of Colombian PVX isolates (Gil et al., 2012). This could indicate that representatives of sub-cluster II-3 are quite wide-spread in potato growing regions of the Andes.

One sample (GAF225) collected near Quichas (Cerro de Pasco, Peru) contained a mixture of two PVX strains: the first clustering

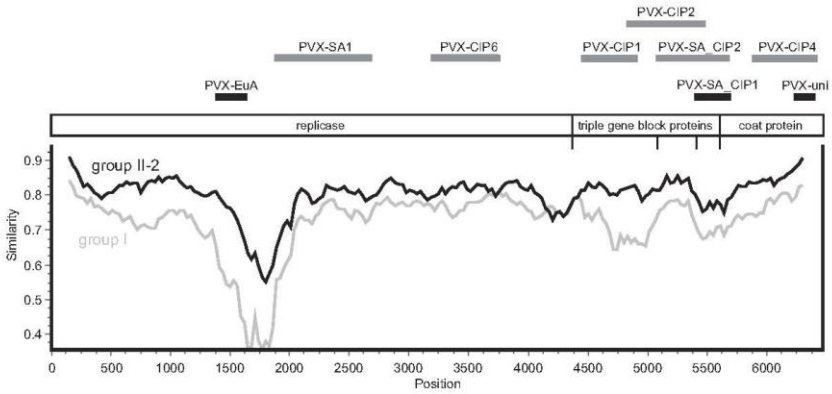


Fig. 2. Upper part: schematic representation of the PVX genome with plotted positions of amplicons produced with PCR primers used for discrimination between different groups of PVX isolates (black bars) and validation of genome sequences using Sanger sequencing (gray bars). The amplicon names follow the ones listed in Table S1. Below: diagram showing the similarity of SA-CIP (group II-3 as a query) to isolates from two other clusters with completely sequenced genome (I and II-2) over the complete genome sequence.

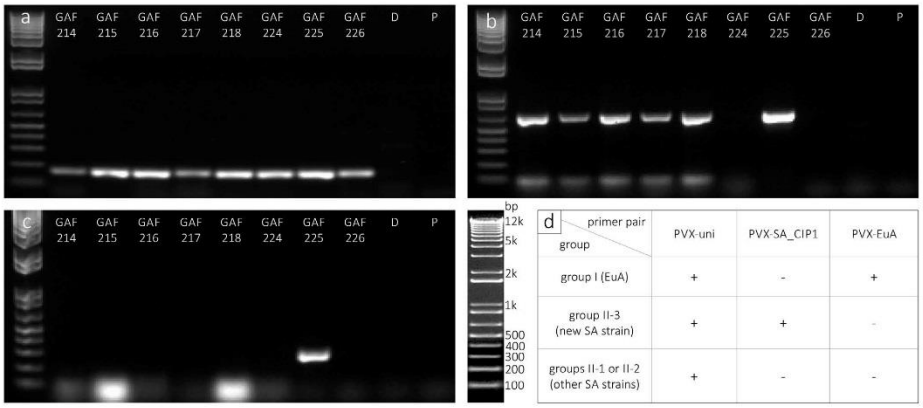


Fig. 3. Discrimination between PVX groups of isolates using specific RT-PCR primers. The figures (a)–(c) are showing pictures of RT-PCR amplified fragments on agarose gels using (a) generic primers for all PVX groups (PVX-uni), (b) primers specific for new (II-3) strain group (PVX-SA_CIP1) and (c) primers specific for Eurasian group (I) of isolates (PVX-EuA). All the samples analyzed for sRNAs were included (names following the ones in Table 1), healthy potato (P) and datura (D) plants were included as negative control. Part (d) shows decision table for the discrimination between groups of PVX isolates using designed primers.

within the Eurasian cluster I and the second with the new SA-CIP isolates detected in greenhouse samples (cluster II-3; Fig. 1). It was shown for several plant viruses that mixed infections could give a rise to new strains through recombination (reviewed in Sztuba-Solińska et al., 2011) or influence the disease rate in epidemics (Pérêfarres et al., 2014). For PVX, experimental evidence for appearance of new strains after mixed infection of different strains was reported (Thomson, 1961) and more recently, Draghici and Varrelmann (2009, 2010) characterized recombination parameters of PVX using an in vivo experimental assay. Nevertheless, possible recombinants of different strains of PVX have not been reported in nature so far and we were unable to detect any in this study either. On the other hand, it was shown that mixed infection of strains modulate evolutionary dynamics of closely related pepino mosaic virus (genus *Potexvirus*) through the expansion of host range and reduction of symptoms (Gómez et al., 2009), which could increase the spread of the pathogen. Similar effect may also occur in mixed infection of PVX strains and an efficient method for detection and tracing of such infections would thus be needed. Commonly used detection methods (ELISA, RT-PCR targeting conserved sites) could fail to discriminate between different strains due to the high sequence similarity. Therefore, NGS based approach represents an efficient method for a detection of such mixed strains infections.

For quick detection of PVX in more extensive screening studies, RT-PCR is a more rapid and cheap option. Thus, we developed a new PVX universal assay, which can be used for the detection of all known PVX phylogenetic clusters. Two specific RT-PCR assays were designed for discrimination between cluster I, sub-cluster II-3 and other, previously known South American isolates. Tests with the current isolates confirmed their specificity. The gel pictures of RT-PCR test and decision table are shown in Fig. 3. These assays could be used when screening for diversity of PVX isolates and especially for the detection of II-3 isolates in order to determine the extent of their distribution in potato growing regions of South America.

sRSA is being increasingly used for the discovery and diagnostics of viruses in plants and invertebrates. A specific challenge of this approach is the short length of obtained reads, which can complicate reconstruction of viral genomes. Nevertheless, sRSA has been successfully used for reliable whole genome reconstruction in cases of several different viruses and hosts. Here we confirmed this also for PVX (infecting potato and datura plants) and identify a novel phylogenetic lineage of this virus. We were also able to reliably discriminate and assemble the near complete genomes of two divergent variants of the same virus in a mixed infection within a single host plant, which has not previously been reported using sRSA. It should be noted however that, in our example, the difference in genome sequence between the two strains was quite high (20% dissimilarity in average, see Table 2 and Fig. 2) and differences were well distributed throughout the genome without stretches of high similarity, thus enabling reliable assembly of near complete genomes of both strains using very short (21–24 bp) RNA sequences. This might not be the case if the strains constituting the mixed infection are more similar one to each other, have long stretches with high similarity and/or if there are no available reference genomes. Nevertheless, the mixed infection can be identified by overlapping contigs with varying nucleotide sequences and if complete genomes are desired they can be determined by amplification using classical methods or by coupling NGS with other viral target enrichment approaches (e.g. viral particles purification) or sequencing total RNA (Kehoe et al., 2014), which would enable generation of sequencing libraries with longer fragment lengths.

The complete genome sequence of representatives of a new, divergent PVX phylogenetic lineage provides broader insight into the diversity of this pathogen in South America. PVX is considerably under-sampled in this region: Cox and Jones (2010) listed 9 complete coat protein sequences of isolates from this region (found in databases), compared to other 76, sampled mostly in Europe. However, as noted by Cox and Jones (2010), greatest diversity of PVX should be expected in Andean region of South America. Being in the center of diversity for cultivated potato, Peru is rich in wild and cultivated species and varieties of potato representing diverse environmental niches for the diversification of potato pathogens, such as PVX. In accordance with this notion, PVX isolates reported from the Andean region are found in all four recognized phylogenetic clusters, and in the current study three of them were found in a single region. An extensive study of this pathogen would be needed to assess the extent of a SA-CIP strain occurrence and to search for possible new strains. Similar pattern of diversity in this region may

also be expected for other potato pathogens and pathogens of other solanaceaous species.

Acknowledgments

We gratefully acknowledge financial contribution from ICGEB-TWAS-UNESCO/IBSP Joint Program on Capacity Buildingin Basic Molecular Biology (Contract No.: CRP/101005), EU FP7 through the QBOL project (Grant No. 226482). The work was also supported by Slovenian Research Agency through the grant L4-5525. Denis Kutnjak is a recipient of a PhD research grant from Slovenian Research Agency.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at http://dx.doi.org/10.1016/j.virusres.2014.07.012.

References

Bi, Y., Tugume, A.K., Valkonen, J.P., 2012. Small-RNA deep sequencing reveals *Arctium tomentosum* as a natural host of *Alstroemeria virus X* and a new putative Emaravirus. PLoS ONE 7, e42758.

Cockerham, G., 1955. Strains of potato virus X. In: Want, J.P.H. (Ed.), Proceeding of 2nd Conference on Potato Virus Diseases Lisse-Wageningen 1954. H. Veenman & Zonen, Wageningen, pp. 89–92.

Cox, B.A., Jones, R.A.C., 2010. Genetic variability of the coat protein gene of *Potato virus X* and the current relationship between phylogenetic placement and resistance groupings. Arch. Virol. 155, 1349–1356.

Cuellar, W.J., Cruzado, R.K., Fuentes, S., Untiveros, M., Soto, M., Kreuze, J.F., 2011. Sequence characterization of a Peruvian isolate of *Sweet potato chlorotic stunt virus*: further variability and a model for p22 acquisition. Virus Res. 157, 111–115.

De Souza, J., Fuentes, S., Savenkov, S., Cuellar, W., Kreuze, J., 2013. The complete nucleotide sequence of sweet potato C6 virus: a carlavirus lacking a cysteine-rich protein. Arch. Virol. 158, 1393–1396.

Draghici, H.K., Varrelmann, M., 2009. Evidence that the linker between the methyltransferase and helicase domains of potato virus X replicase is involved in homologous RNA recombination. J. Virol. 83, 7761–7769.

Draghici, H.K., Varrelmann, M., 2010. Evidence for similarity-assisted recombination and predicted stem-loop structure determinant in potato virus x RNA recombination. J. Gen. Virol. 91, 552–562.

Feigelstock, D.A., Tozzini, A.C., Hopp, H.E., 1995. Coat protein sequence of a resistance-breaking strain of potato virus X isolated in Argentina. Virus Genes 10, 289–292.

Fuentes, S., Heider, B., Tasso, R.C., Romero, E., Zum Felde, T., Kreuze, J.F., 2012. Complete genome sequence of a potyvirus infecting yam beans (*Pachyrhizus* spp.) in Peru. Arch. Virol. 157, 773–776.

Gil, J.F., Cotes, J.M., Marín, M., 2012. Detección y caracterización molecular del Virus X de la Papa (PVX) en regiones productoras de papa de Colombia. Rev. Protec. Vegetal 27, 69–76.

Gómez, P., Sempere, R.N., Elena, S.F., Aranda, M.A., 2009. Mixed infections of *Pepino Mosaic Virus* strains modulate the evolutionary dynamics of this emergent virus. J. Virol. 83, 12378–12387.

Hall, T.A., 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp. Ser. 41, 95–98.

Hwang, Y.T., Kalischuk, M., Fusaro, A.F., Waterhouse, P.M., Kawchuk, L., 2013. Small RNA sequencing of Potato leafroll virus-infected plants reveals an additional subgenomic RNA encoding a sequence-specific RNA-binding protein. Virology 438, 61–69.

Jones, R.A.C., 1985. Further studies on resistance breaking strains of potato virus X. Plant Pathol. 34, 182–189.

Kagiwada, S., Yamaji, Y., Komatsu, K., Takahashi, S., Mori, T., Hirata, H., Suzuki, M., Ugaki, M., Namba, S., 2005. A single amino acid residue of RNA-dependent RNA polymerase in the *Potato virus X* genome determines the symptoms in *Nicotiana* plants. Virus Res. 110, 117–182.

Kashif, M., Pietilä, S., Artola, K., Jones, R.A.C., Tugume, A.K., Mäkinen, V., Valkonen, J.P.T., 2012. Detection of viruses in sweetpotato from Honduras and Guatemala augmented by deep-sequencing of small-RNAs. Plant Dis. 96, 1430–1437.

Kehoe, M.a, Coutts, B.a, Buirchell, B.J., Jones, R.a C., 2014. Hardenbergia mosaic virus: crossing the barrier between native and introduced plant species. Virus Res. 184, 87–92, http://dx.doi.org/10.1016/j.virusres.2014.02.012.

Khurana, S.M.P., Singh, M.N., 1988. Yield loss potential of potato viruses X and Y in Indian potatoes. J. Indian Potato Assoc. 15, 27–29.

King, A.M.Q., Adams, M.J., Carstens, E.B., Lefkowitz, E.J. (Eds.), 2011. Virus Taxonomy—Ninth Report of the International Committee on Taxonomy of Viruses. Elsevier/Academic Press, London.

Kreuze, J., Koenig, R., De Souza, J., Vettenb, H.J., Muller, G., Flores, B., Ziebell, H., Cuellar, W., 2013. The complete genome sequences of a Peruvian and a Colombian isolate of Andean potato latent virus and partial sequences of further isolates suggest the existence of two distinct potato-infecting tymovirus species. Virus Res. 173, 431–435.

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

Li, R., Gao, S., Hernandez, A.G., Wechter, W.P., Fei, Z., Ling, K.S., 2012. Deep sequencing of small RNAs in tomato for virus and viroid identification and strain differentiation. PLoS ONE 7, e37127.

Loconsole, G., Onelge, N., Potere, O., Giampetruzzi, A., Bozan, O., Satar, S., De Stradis, A., Savino, V., Yokomi, R.K., Saponari, M., 2012a. Identification and characterization of *citrus yellow vein clearing virus*, a putative new member of the genus *Mandarivirus*. Phytopathology 102, 1168–1175.

Loconsole, G., Saldarelli, P., Doddapaneni, H., Savino, V., Martelli, G.P., Saponari, M., 2012b. Identification of a single-stranded DNA virus associated with citrus chlorotic dwarf disease, a new member in the family *Geminiviridae*. Virology 432, 162–172.

Lole, K.S., Bollinger, R.C., Paranjape, R.S., Gadkari, D., Kulkarni, S.S., Novak, N.G., Ingersoll, R., Sheppard, H.W., Ray, S.C., 1999. Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. J. Virol. 73, 152–160.

Malcuit, I., Jong, W., Baulcombe, D.C., Shields, D.C., Kavanagh, T.A., 2000. Acquisition of multiple virulence/avirulence determinants by Potato virus X (PVX) has occurred through convergent evolution rather than through recombination. Virus Genes 20, 165–172.

Malcuit, I., Marano, M.R., Kavanagh, T.A., De Jong, W., Forsyth, A., Baulcombe, D.C., 1999. The 25 kDa movement protein of PVX elicits Nb-mediated hypersensitive cell death in potato. Mol. Plant–Microbe Interact. 12, 536–543.

Martin, D.P., Lemey, P., Lott, M., Moulton, V., Posada, D., Lefeuvre, P., 2010. RDP3: a flexible and fast computer program for analyzing recombination. Bioinformatics 26, 2462–2463.

Moreira, A., Jones, R.A.C., Fribourg, C.E., 1980. Properties of a resistance-breaking strain of potato virus X. Ann. Appl. Biol. 95, 93–103.

Péréfarres, F., Thébaud, G., Lefeuvre, P., Chiroleu, F., Rimbaud, L., Hoareau, M., Reynaud, B., Lett, J.-M., 2014. Frequency-dependent assistance as a way out of competitive exclusion between two strains of an emerging virus. Proc. R. Soc. B: Biol. Sci. 281, 20133374.

Querci, M., Baulcombe, D.C., Goldbach, R.W., Salazar, L.F., 1995. Analysis of the resistance-breaking determinants of potato virus X (PVX) strain HB on different potato genotypes expressing extreme resistance to PVX. Phytopathology 85, 1003–1010.

Roy, A., Choudhary, N., Guillermo, L.M., Shao, J., Govindarajulu, A., Achor, D., Wei, G., Picton, D.D., Levy, L., Nakhla, M.K., Hartung, J.S., Brlansky, R.H., 2013. A novel virus of the genus *Cilevirus* causing symptoms similar to Citrus Leprosis. Phytopathology 103, 488–500.

Santa Cruz, S., Baulcombe, D.C., 1993. Molecular analysis of potato virus X isolate in relation to the potato hypersensitivity gene Nx. Mol. Plant–Microbe Interact. 6, 707–714.

Santa Cruz, S., Baulcombe, D.C., 1995. Analysis of potato virus X coat protein genes in relation to resistance conferred by the genes Nx, Nb and Rx1 of potato. J. Gen. Virol. 76, 2057–2061.

Scholthof, K.B.G., Adkins, S., Czosnek, H., Palukaitis, P., Jacquot, E., Hohn, T., Hohn, B., Saunders, K., Candresse, T., Ahlquist, P., Hemenway, C., Foster, G.D., 2011. Top 10 plant viruses in molecular plant pathology. Mol. Plant Pathol. 12, 938–954.

Sela, N., Luria, N., Dombrovsky, A., 2012. Genome assembly of bell pepper *Endornavirus* from small RNA. J. Virol. 86, 7721–7721.

Sztuba-Solińska, J., Urbanowicz, A., Figlerowicz, M., Bujarski, J.J., 2011. RNA–RNA recombination in plant virus replication and evolution. Annu. Rev. Phytopathol. 49, 415–443.

Thomson, A.D., 1961. Interactions between plant viruses. I. Appearance of new strains after mixed infection with potato virus X strains. Virology 13, 507–514.

Untiveros, M., Quispe, D., Kreuze, J.F., 2010. Analysis of complete genomic sequences of isolates of the Sweet potato feathery mottle virus strains C and EA: molecular evidence for two distinct potyvirus species and two P1 protein domains. Arch. Virol. 155, 2059–2063.

Vance, V.B., 1991. Replication of Potato virus X RNA is altered in coinfections with Potato virus Y. Virology 182, 486–494.

Wu, Q., Luo, Y., Lu, R., Lau, N., Lai, E.C., Li, W.X., Ding, S.W., 2010. Virus discovery by deep sequencing and assembly of virus derived small silencing RNAs. Proc. Natl. Acad. Sci. U. S. A. 107, 1606–1611.

Yu, X.Q., Jia, J.L., Zhang, C.L., Li, X.D., Wang, Y.J., 2010. Phylogenetic analyses of an isolate obtained from potato in 1985 revealed potato virus X was introduced to China via multiple events. Virus Genes 40, 447–451.

Yu, X.Q., Wang, H.Y., Lan, Y.F., Zhu, X.P., Li, X.-D., Fan, Z.F., Li, H.F., Wang, Y.Y., 2008. Complete genome sequence of a Chinese isolate of Potato virus X and analysis of genetic diversity. J. Phytopathol. 156, 346–351.

Zerbino, D.R., Birney, E., 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18, 821–829.

Zhang, Y., Singh, K., Kaur, R., Qiu, W., 2011. Association of a novel DNA virus with the grapevine vein-clearing and vine decline syndrome. Phytopathology 101, 1081–1090.

2.3 Deep Sequencing of Virus-Derived Small Interfering RNAs and RNA from Viral Particles Shows Highly Similar Mutational Landscapes of a Plant Virus Population

Denis Kutnjak, Matevž Rupar, Ion Gutierrez-Aguirre, Tomaž Curk, Jan F. Kreuze, Maja Ravnikar

Journal of Virology, 2015, 89(9): 4760-4769. DOI: 10.1128/JVI.03685-14

In this research we compared two previously used virus enrichment approaches: sRNA deep sequencing and sequencing of RNA from CIM monolithic chromatography purified viral particles, using *Potato virus Y* as a model. Sequencing RNA from purified viral particles outperformed sRNA sequencing in terms of the reconstruction of complete viral consensus genome sequence. It also allowed us to search for recombinations, which was not possible using short sRNA reads. Both of the sequence pools (virus derived sRNAs and RNAs from purified viral particles) reflected very similar diversity of virus population within a host.

The PhD candidate designed the experiment, assisted at the viral particles purification step, isolated the RNA and prepared sequencing libraries for Illumina sequencing. He performed all the sequence data analyses and comparisons and visualized the results. He drafted the manuscript.

JVI
Journals.ASM.org

Deep Sequencing of Virus-Derived Small Interfering RNAs and RNA from Viral Particles Shows Highly Similar Mutational Landscapes of a Plant Virus Population

Denis Kutnjak,[a,b] Matevž Rupar,[a] Ion Gutierrez-Aguirre,[a] Tomaž Curk,[c] Jan F. Kreuze,[d] Maja Ravnikar[a]

Department of Biotechnology and Systems Biology, National Institute of Biology, Ljubljana, Slovenia[a]; Jožef Stefan International Postgraduate School, Ljubljana, Slovenia[b]; University of Ljubljana, Faculty of Computer and Information Science, Ljubljana, Slovenia[c]; International Potato Center (CIP), Lima, Peru[d]

ABSTRACT

RNA viruses exist within a host as a population of mutant sequences, often referred to as quasispecies. Within a host, sequences of RNA viruses constitute several distinct but interconnected pools, such as RNA packed in viral particles, double-stranded RNA, and virus-derived small interfering RNAs. We aimed to test if the same representation of within-host viral population structure could be obtained by sequencing different viral sequence pools. Using ultradeep Illumina sequencing, the diversity of two coexisting *Potato virus Y* sequence pools present within a plant was investigated: RNA isolated from viral particles and virus-derived small interfering RNAs (the derivatives of a plant RNA silencing mechanism). The mutational landscape of the within-host virus population was highly similar between both pools, with no notable hotspots across the viral genome. Notably, all of the single-nucleotide polymorphisms with a frequency of higher than 1.6% were found in both pools. Some unique single-nucleotide polymorphisms (SNPs) with very low frequencies were found in each of the pools, with more of them occurring in the small RNA (sRNA) pool, possibly arising through genetic drift in localized virus populations within a plant and the errors introduced during the amplification of silencing signal. Sequencing of the viral particle pool enhanced the efficiency of consensus viral genome sequence reconstruction. Nonhomologous recombinations were commonly detected in the viral particle pool, with a hot spot in the 3′ untranslated and coat protein regions of the genome. We stress that they present an important but often overlooked aspect of virus population diversity.

IMPORTANCE

This study is the most comprehensive whole-genome characterization of a within-plant virus population to date and the first study comparing diversity of different pools of viral sequences within a host. We show that both virus-derived small RNAs and RNA from viral particles could be used for diversity assessment of within-plant virus population, since they show a highly congruent portrayal of the virus mutational landscape within a plant. The study is an important baseline for future studies of virus population dynamics, for example, during the adaptation to a new host. The comparison of the two virus sequence enrichment techniques, sequencing of virus-derived small interfering RNAs and RNA from purified viral particles, shows the strength of the latter for the detection of recombinant viral genomes and reconstruction of complete consensus viral genome sequence.

RNA viruses are one of the fastest-evolving biological entities known. Due to their high mutation and recombination rates, viral populations exist within hosts as a cloud of nonidentical but similar sequences, often referred to as viral quasispecies (1). The generated variability, coupled with natural selection, population bottlenecks, and stochasticity, shape the structure of virus populations, which was shown to have important implications in virus fitness and pathogenicity (1). With the advent of next-generation sequencing (NGS), in-depth studies of viral populations within a host became possible. In the past few years, several in-depth population studies have been conducted on human (2) or animal pathogenic viruses using NGS (3–5). Less attention has been given to plant-infecting viruses; several studies have been carried out using Sanger amplicon sequencing of a limited number of molecular clones (6–10). All NGS in-depth within-plant virus population studies reported to date employed amplicon sequencing, focusing on only a particular part of the viral genome (11–14). However, different parts of the viral genome can be subjected to different selection pressures (15, 16); thus, whole-genome characterization of virus populations would give a more complete picture.

High background levels of host nucleic acids and the high diversity of viral populations complicate the reconstruction of a complete consensus viral genome sequence from NGS data. Moreover, within-host viral population studies demand high sequencing coverage (10,000× and more). Most of such studies

Received 23 December 2014 **Accepted** 4 February 2015
Accepted manuscript posted online 11 February 2015
Citation Kutnjak D, Rupar M, Gutierrez-Aguirre I, Curk T, Kreuze JF, Ravnikar M. 2015. Deep sequencing of virus-derived small interfering RNAs and RNA from viral particles shows highly similar mutational landscapes of a plant virus population. J Virol 89:4760–4769. doi:10.1128/JVI.03685-14.
Editor: A. E. Simon
Address correspondence to Denis Kutnjak, denis.kutnjak@nib.si, and Maja Ravnikar, maja.ravnikar@nib.si.
Supplemental material for this article may be found at http://dx.doi.org/10.1128/JVI.03685-14.
Copyright © 2015, American Society for Microbiology. All Rights Reserved.
doi:10.1128/JVI.03685-14
The authors have paid a fee to allow immediate free access to this article.

Downloaded from http://jvi.asm.org/ on May 7, 2015 by guest

include a preamplification step for the amplification of complete or partial viral genomes (2). Such an approach allows us to achieve great sequencing depth easily; however, it can distort the variation occurring in primer-annealing regions (17) and affect the detection of other variants, which are epistatically connected with those regions. Moreover, this approach requires specific knowledge about viral genome sequence, since the sequences of PCR primers used for the amplification of the viral genome have to be known in advance.

A more generic solution is the enrichment for viral nucleic acids before sequencing library preparation, employing the characteristics of different viral sequence pools within the host. In plants, the sequences of RNA virus origin constitute several different pools: RNA molecules packed within the virions, double-stranded RNA (dsRNA) molecules, and virus-derived small interfering RNAs (vsiRNAs). dsRNA molecules can arise through replication or the action of plant RNA-dependent RNA polymerases. Shorter stretches of dsRNA-like structures also can be produced by intramolecular base pairing between different parts of the viral genome. In plants and invertebrates, viral dsRNA molecules represent a substrate for the generation of 21- to 24-nucleotides (nt)-long virus-derived small interfering RNAs (vsiRNAs), which result from the activity of dicer-like proteins (18).

Deep sequencing of vsiRNAs has been used efficiently for the reconstruction of consensus viral genome sequences from plants (19) and invertebrates (20). It is a relatively time-efficient, affordable, and generic method, and as such, it is widely applicable. It was observed that different variants could be detected in the pool of vsiRNAs (21). However, no comparisons have been made yet to test if the variation observed in vsiRNAs reflects the full diversity of viral populations. To fill this gap, we have used a sequence-independent, preamplification-free deep sequencing approach to investigate and compare two different pools of RNA virus sequences from the infected plants: single-stranded RNA (ssRNA) isolated from purified viral particles (VP pool) and vsiRNAs (sRNA pool). Deep sequencing of viral RNA directly from purified viral particles was performed only in a few cases and was shown to be highly efficient for enrichment for viral nucleic acids (4, 22, 23). Here, we have used the recently developed CIM monolithic chromatography-based approach for fast purification of viral particles (24).

Potato virus Y (PVY)-potato was used as a model system. PVY, a single-stranded positive-sense RNA virus and a member of the genus *Potyvirus* (family *Potyviridae*), is one of the most important potato pathogens and is distributed in potato-growing regions worldwide (25). In this study, the highly pathogenic recombinant strain NTN was investigated.

Our two main aims were to (i) compare the efficiency of the approaches for the consensus virus genome sequence generation and (ii) compare the within-host population structure inferred from the two viral sequence pools. Single-nucleotide polymorphisms (SNPs) were compared between the samples using a stringent filtering strategy; a low-frequency SNP detection algorithm and technical replicates were used to discriminate real SNPs from reverse transcription (RT), PCR, and sequencing artifacts. The VP pool was investigated additionally for recombination events. We hypothesized that the diversity of vsiRNAs and RNA isolated from viral particles show a highly similar mutational landscape.

This study presents the most comprehensive whole-genome characterization of within-plant virus population to date and allows independent validation and comparison of variants by sequencing two different pools of virus sequences within the same plants. The study provides a firm baseline for future studies of plant virus population diversity and dynamics after bottleneck (transmission by vectors, systemic movement within host) or radiation (spread to a new host) events.

MATERIALS AND METHODS

Preparation of infected plant samples. PVY (strain NTN; isolate NIB V 151)-infected *Solanum tuberosum* cv. Pentland Squire plants were propagated in stem node culture (*in vitro*) and transferred to soil. After 3 weeks, whole green parts (leafs and stem) of 60 plantlets were harvested. The leaves and stem of each plant were divided in two pools, one used for the purification of viral particles and the other for small RNA isolation. First, for the isolation of small RNAs, equal amounts of young leaf, old leaf, and stem were sampled from each plant and pooled (amounting in a total of 1 g of plant material). Subsequently, all of the remaining harvested material (16.74 g) from the same plants was combined and used for purification of viral particles. For a detailed scheme of the experiment, see Fig. 1.

Purification of PVY particles and isolation and fragmentation of viral RNA. Viral particles were purified using convective interactive media (CIM) monolithic chromatographic supports as previously described (24). Chromatographic fractions containing purified viruses were pooled and subjected to total RNA isolation using TRIzol LS reagent (Life Technologies, Invitrogen, Carlsbad, CA, USA) by following the manufacturer's instructions. Prior to sequence library preparation, isolated PVY RNA was fragmented using a NEBNext magnesium RNA fragmentation module (NEB, Ipswich, MA, USA). The fragmentation and end repair procedure followed the Illumina directional mRNA-Seq sample preparation guide (15018460, Rev. A, 2010). The size and quantity of purified fragmented viral RNA was assessed with the Agilent 2100 Bioanalyzer using an RNA Pico chip (Agilent Technologies, CA, USA).

Isolation of small RNAs. Total RNA was isolated from pooled plant material (1 g) using TRIzol reagent (Life Technologies, Invitrogen, Carlsbad, CA, USA) by following the manufacturer's instructions. RNA was separated in 15% urea-polyacrylamide gel. The band corresponding to the size of small RNAs (cloud of the smallest size, ~10 to 50 nt) was cut from the gel and crushed. Crushed gel pieces were soaked in 700 μl of 2.5 M NaCl with 0.1% β-mercaptoethanol at 4°C overnight. Samples then were spun down, and 700 μl supernatant was transferred to Freeze 'N Squeeze columns (Bio-Rad, Hercules, CA, USA). The remaining gel pieces were washed with 300 μl of 2.5 M NaCl with 0.1% β-mercaptoethanol, which was added to Freeze 'N Squeeze columns. The columns were spun at 5,000 rpm for 1 min. Resuspended RNA was cleaned using a 0.5 volume of chloroform, precipitated with 1 volume of isopropanol, and washed with 75% ethanol. Finally, the pellet was air dried and resuspended in 10 μl of RNase-free water.

Library preparation and sequencing. Illumina TruSeq sequencing libraries were prepared for both fragmented RNA isolated from purified PVY particles (VP) and small RNAs (sRNA) according to the published protocol (26), with the following modifications: (i) reverse transcription reactions were made in a reduced (50%) volume; (ii) in the final PCR enrichment step, a 4× concentration of forward and reverse primers was used. The libraries were size selected in 10% native polyacrylamide gel, and the DNA was purified from the gel using chloroform-isopropanol extraction. The size distribution of purified libraries was inspected on a Caliper LabChip GX using an HS DNA chip (PerkinElmer, Hopkinton, MA, USA). The libraries were sent for sequencing on an Illumina Hiseq2000 platform in single-end 50-nt mode to Fasteris Life Sciences SA (Plan-les-Ouates, Switzerland). In order to confidently discriminate real SNPs in the population from RT, PCR, and sequencing artifacts, we used technical replicates. Two library technical replicates were prepared (from the RT step onward) for each sequence pool, amounting to a total of 4 samples (sRNA1, sRNA2, VP1, and VP2).

Downloaded from http://jvi.asm.org/ on May 7, 2015 by guest

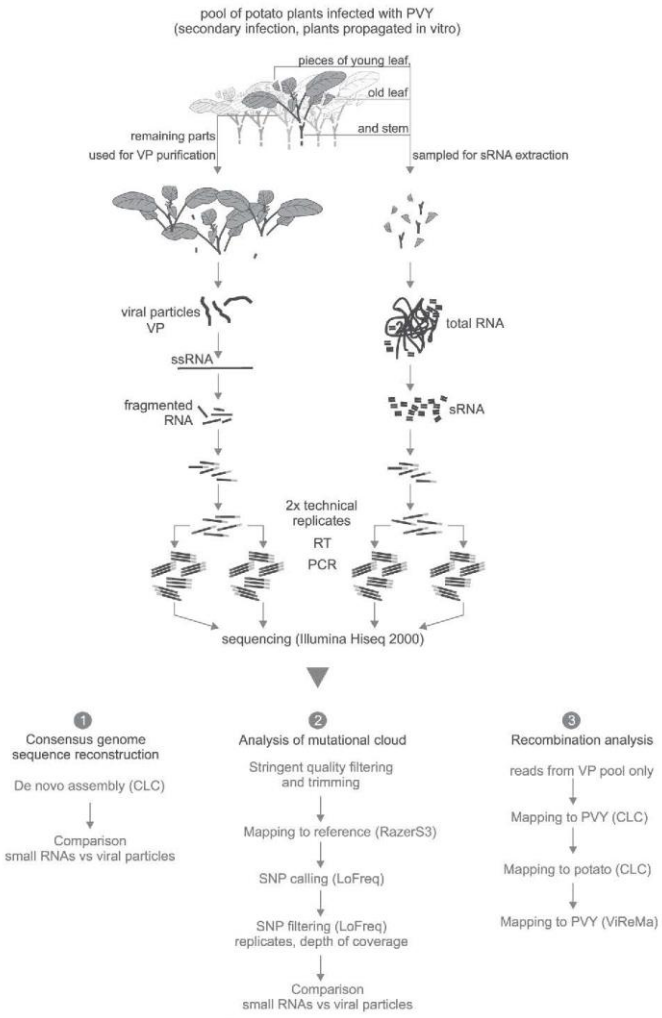


FIG 1 Scheme of the experiment and analysis pipeline of the study.

Reconstruction of consensus genome sequence. Subsamples of 500,000 reads were used for the reconstruction of the consensus genome sequence to achieve optimal performance of the algorithms (the performance of *de novo* assembly algorithms is not optimized for very high sequencing depths; up to ~100,000× in our case). Velvet 1.2.10 (27) and the CLC Genomics Workbench 6.0 (CLC bio) *de novo* assembler were compared initially. The latter showed better performance (significantly longer contigs); thus, it was used for all subsequent analyses (using default parameters, with the minimum contig length set to 50 nt). Parameters of assemblies were compared between sRNA and VP pools, and contigs were mapped to reconstructed consensus complete viral genome sequences to define the proportion of genome they cover for each of the samples. The analysis was repeated on 10 subsamples, and the results were plotted using R 3.0.2 (28). The reconstructed PVY complete genome sequence was deposited in GenBank under accession number KM396648.

To define the ratio between viral and host genome sequences, reads from each set first were mapped to the reconstructed PVY genome (KM396648). We removed reads shorter than 15 nt because they could map nonuniquely. Reads that did not map to PVY then were mapped to a published potato genome sequence (International Nucleotide Sequence

Downloaded from http://jvi.asm.org/ on May 7, 2015 by guest

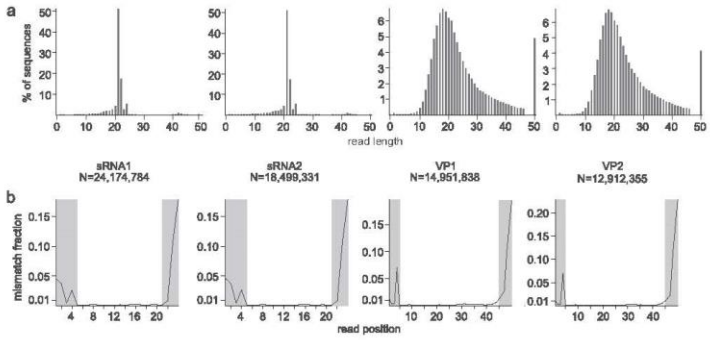


FIG 2 Read length profiles, mismatch fractions per read position, and read trimming. (a) Total number of reads and read length profiles for each of the samples. (b) Relative frequencies of mismatches per position in the read after the test mapping. The higher mismatch frequencies observed at the 5′ and 3′ ends of the reads were trimmed (the positions shaded in gray) and were not used for subsequent SNP calling.

Database Collaboration [INSDC] assembly no. GCA_000226075.1). CLC Genomics was used for the mapping. The depth of coverage for each position of PVY genome was exported from CLC Genomics and normalized to the sample with the lowest total read count (~11.67 million reads). The data were visualized as a line plot using R 3.0.2 (28), with lines being smoothed using the smooth.spline function (spar = 0.3).

Analysis of PVY within-plant population structure: SNP calling. For SNP analysis, a stringent data filtering and trimming strategy was used before mapping to avoid false-negative results: (i) the reads containing ambiguous bases (Ns) were discarded; (ii) for sRNA samples, 21- to 24-nt-long reads were selected, and for VP samples, reads longer than 24 nt (25 to 50 nt) were selected; (iii) test alignments were made and mismatch profiles per read position were inspected (Fig. 2), and then positions <5 and positions >45 were trimmed from VP reads and positions <5 and >21 were trimmed from sRNA reads; and (iv) reads containing nucleotides with a Phred Q score lower than 20 were discarded. Filtered and trimmed reads were aligned to the reference consensus genome sequence obtained by *de novo* assembly as described above (accession number KM396648). Several mapping algorithms were tested (CLC Genomics Workbench, bowtie, bowtie2, and RazerS3). RazerS3 was selected as the most suitable for our particular case (very short reads of variable length) and was used for the construction of final alignments, allowing no gaps and a minimum 90% identity match. SAMtools view (29) was used to randomly sample the alignments to comparable sequencing depths (the average depth of coverage was then ~8,000× for each of the samples). The resulting alignments were used for SNP calling using LoFreq (30). LoFreq allows the detection of variants with frequency lower than the average sequencing error rate. It uses Poisson-binomial distribution-based modeling of sequencing errors for each site in the genome separately, considering position-specific Phred quality scores as well as sequencing depth of the position to distinguish legitimate SNPs from sequencing artifacts (30). Thus, error probabilities are computed for each SNP separately without depending on a universal arbitrary frequency cutoff value. However, the algorithmic filtering implemented in LoFreq can account only for the mistakes arising during the last sequencing step of the process. To further eliminate false positives, which could arise during RT or PCR, we used two replicates (from the RT step on) for each of the samples. Thus, SNPs called by LoFreq were filtered, and only the ones occurring in both replicates were considered reliable. This filtering approach allowed us to confidently discriminate real SNPs from the RT, PCR, and sequencing noise, discovering SNPs in the range of 0.06 to 50% (see Fig. 4b) without the use of an arbitrary frequency cutoff value. Moreover, for comparisons between the pools, we used the lofreq.unique script implemented in the program to further filter out the SNPs which did not have sufficient sequencing depth in all of the samples (30).

SNPs were annotated for possible amino acid changes (synonymous/nonsynonymous mutation, consensus amino acid, and variant amino acid) using an in-house-made script. Additionally, the corresponding BLOSUM62 score for each of the SNPs was obtained from the BLOSUM62 matrix (31).

Correlation of SNP frequencies between replicates and between pools was calculated and plotted (see Fig. 4c to f). First, frequencies of SNPs present in both sRNA replicates were compared. Second, the same analysis was made for VP pool replicates. Third, frequencies of SNPs found in both pools (sRNA and VP) were compared. Finally, those SNPs also were sorted by frequency and each SNP was ranked. The ranks of the SNPs in one pool were compared to the ranks of the SNPs in the other pool.

Analysis of PVY within-plant population structure: recombination detection. For the detection of nonhomologous recombination events, only reads from the VP pool were used. Reads longer than 25 nt first were mapped to the consensus PVY genome and then to the host (potato) reference genome. Reads which did not map to either of them were extracted, and ViReMa (32) was used to detect possible virus-virus nonhomologous recombination events (seed length set to 20 nt, removing PCR duplicates). To further assess the robustness of the results, the same analysis was repeated with longer seed lengths (22 nt and 24 nt). Seed length represents the number of nucleotides used by mapping algorithm for the beginning alignment to the reference. Thus, increasing seed length should increase the specificity but decrease sensitivity of the method. To confirm the reliability of the ViReMa approach for the detection of recombination events, 10 of the most frequent recombination events and 5 other randomly chosen ones were further evaluated. We constructed simulated recombinant sequences, spanning the region of 25 nt before and 25 nt after recombination events detected by ViReMa. The ViReMa input reads then were mapped (using CLC Genomics Workbench) to this simulated sequences. We then visually inspected the mapping for each of the simulated recombinant sequences to ensure there were recombinant reads covering each of the tested recombinant points. Since the sRNAs are too short (21 to 24 nt) to be analyzed for recombination signal using ViReMa or other existing bioinformatics tools, we tried to detect recombination signals by mapping the sRNA sample reads to the same simulated recombinant sequences and then visually inspecting the results.

To search for possible recombination hot spots in the PVY genome, only recombination events supported by two or more nonidentical reads in both replicates were considered. Such a conservative filtering approach significantly reduced the number of the detected recombination points;

Downloaded from http://jvi.asm.org/ on May 7, 2015 by guest

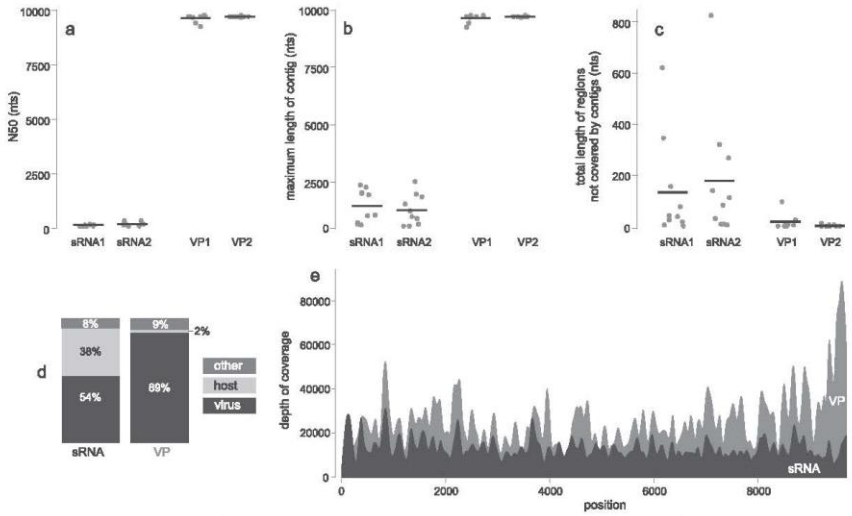


FIG 3 Generation of complete consensus viral genome sequence. *De novo* assembly quality parameters for 10 subsamples (500,000 reads per subsample). (a) N50; (b) maximum length of contig; (c) total length of viral genome not covered by contigs. Black lines represent average values. (d) The percentage of reads mapped to the viral and host reference genome for the small RNA (sRNA) pool and viral particle (VP) pool. (e) Distribution of sequencing depth across the viral genome for VP and sRNA pools.

however, we were able to reliably exclude possible artifacts of the wet-laboratory and analysis procedure. The number of filtered recombination points was counted in each 10-nt-wide window across the PVY genome and visualized as a histogram.

Circos 0.65 (33) was used for the visualization of the results. The square root of relative frequencies of filtered SNPs, their BLOSUM62 scores, and detected recombination events were visualized in a circular plot (see Fig. 5).

Nucleotide sequence accession numbers. Raw sequencing reads from this experiment have been deposited in the NCBI Sequence Read Archive (SRA) under accession numbers SRR1556759, SRR1556760, SRR1556761, and SRR1556762. The consensus viral genome sequence has been deposited in NCBI GenBank under accession number KM396648.

RESULTS

Illumina deep sequencing resulted in 13 to 24 million reads for four prepared libraries (Fig. 2). The size distribution of sRNA libraries implied excellent enrichment for sRNAs (prominent 21- to 24-nt peak) (Fig. 2a). It was previously shown that the small RNA library preparation could introduce a bias in the frequency of represented sRNAs, likely due to the strong adaptor-sRNA base pairing (34). To compensate for the bias, RNA isolated from VP was fragmented in similarly short fragments (reads of VP libraries range in length from 10 to 50 nt) (Fig. 2). Thus, the similar size of fragments in both types of libraries allowed us an unbiased comparison of the two pools (VP versus sRNA).

Sequencing RNA isolated from purified viral particles enhances reconstruction of consensus viral genome sequence. *De novo* assembly of reads obtained by sequencing fragmented RNA from purified viral particles repeatedly (two replicates with 10 500,000-read subsamples) produced one contig encompassing the complete or nearly complete viral genome sequence (9,278 to 9,761 nt) (Fig. 3b). *De novo* assembly of small RNAs resulted in several shorter contigs (with a maximum length of 611 to 2,543 nt in different subsamples) (Fig. 3b), which did not overlap; the N50 (the parameter used to describe the quality of assembly) was considerably lower for sRNA assemblies (Fig. 3b). The sRNA contigs covered the majority of, but not the complete, viral genome (up to 826 nt not covered by contigs) (Fig. 3c).

In both cases (VP and sRNA), the entire genome was covered when mapping the reads to the reconstructed PVY complete genome sequence, with no zero-coverage regions (Fig. 3e). In VP samples, viral sequences amounted to 89% of the sample, whereas in sRNA samples they amounted to around 54% (Fig. 3d). There was a low level of contamination with background host sequences in VP samples (only 2% of the reads). In sRNA samples, a higher proportion of reads mapped to the host genome (38%), probably mostly representing potato endogenous small RNAs. The remaining reads (8 to 9%; denoted as "other" in Fig. 3d) did not map to either of the genomes and may represent host sequences not sufficiently similar to the published potato genome sequence, which corresponds to *Solanum phureja* (a closely related diploid potato species), but it also may include sequences of other taxa present in plants and recombinant viral sequences (as confirmed by recombination analysis). The depth of sequencing coverage across the PVY genome was variable but high in both cases. The genome was more uniformly covered in the case of sRNAs, whereas in the case of the VP pool, a notable peak in coverage was observed in the 3′ region of the genome (Fig. 3e).

The mutational landscape is highly similar between VP and sRNA pools. SNP analysis, following rigorous filtering using rep-

Downloaded from http://jvi.asm.org/ on May 7, 2015 by guest

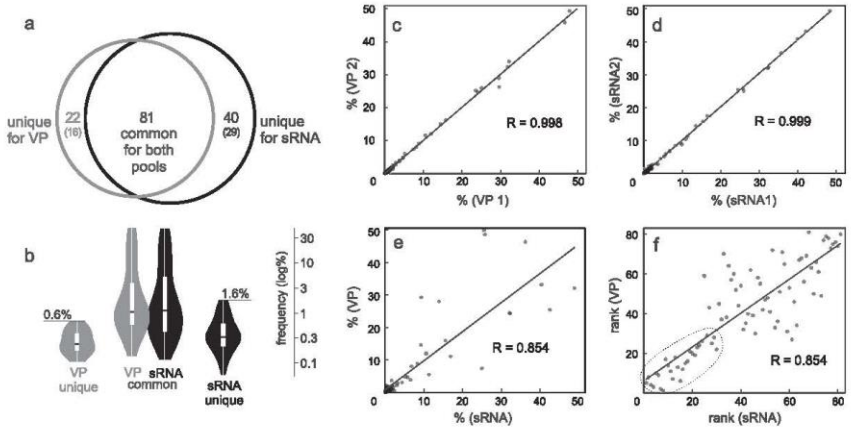


FIG 4 SNP analysis for VP and sRNA pool. (a) Number of predicted SNPs in two different pools, where the black circle represent SNPs present in the sRNA pool and the gray circle represents SNPs present in the VP pool. The overlap represents SNPs present in both of the pools. SNPs were confirmed in both technical replicates (sRNA1/sRNA2 and VP1/VP2). The numbers in parentheses represent further filtered unique SNPs, which had similar depths in all 4 samples. The SNPs detected in only one of the technical replicates were treated as artifacts and are not depicted here. (b) Violin plots showing the distribution of SNP frequencies in the corresponding pool. The black line indicates median frequency. The white box is the first quartile and the white line is the second quartile of the data. The shaded outer area is a kernel density plot, which describes the distribution of the data. The scale is $\log_{10}$ transformed. The percentages on the lines above violin plots for sRNA and VP unique SNPs denote the upper limits of their distribution. (c to f) Correlation (R is the correlation coefficient) between frequencies of common SNPs for two VP replicates (c), for two sRNA replicates (d), and for common SNPs between the sRNA and VP pool (e). (f) Correlation between ranks of common SNPs between sRNA and the VP pool. The outlined area represents roughly the first 30 ranks (frequencies of higher than 1%), which show higher correlation between the pools.

licate data, revealed 103 polymorphic sites for the VP pool and 121 polymorphic sites for the sRNA pool (details, including LoFreq-derived P values, are given in Table SA1 in the supplemental material). The frequencies of variants ranged from 0.06% to 49.86% (Fig. 4b). Eighty-one SNPs were present in both pools. Twenty-two SNPs were found only in the VP pool, and 40 were found only in the sRNA pool. After excluding the sites with uneven coverage across the samples (using the lofreq.unique script), 16 SNPs were confirmed to be unique for the VP pool and 29 for the sRNA pool (Fig. 4a).

The correlation of common SNP frequencies was nearly perfect ($R = 0.99$) (Fig. 4c and d) between the replicates. The correlation of SNP frequencies for the 81 SNPs present in both pools (VP and sRNA) was strong ($R = 0.85$) (Fig. 4e). Moreover, when comparing ranks of the SNPs instead of their frequencies, there was even higher interpool correlation for approximately the first 30 ranks, which largely correspond to the SNPs with frequencies higher than ~1% (Fig. 4f, outlined area).

All of the SNPs which have a frequency of 0.6% or more in the VP pool also were discovered in the sRNA pool. From the opposite perspective, there was a 100% rate of SNP discovery in the VP pool only if this SNP had a frequency of 1.6% or more in the sRNA pool (Fig. 4b).

The SNPs were distributed across the entire genome of PVY, with a possible cold spot in the VPg coding region (Fig. 5b and c). The annotation of SNPs shows larger amounts of nonsimilar amino acid substitutions (lower BLOSUM62 scores) in the P1 and NIb regions of PVY polyprotein (Fig. 5a and d).

Recombinant viral sequences are common in VP pool. The recombination analysis of the VP pool with the ViReMa algorithm suggested that 0.5 to 3% of reads are recombinant viral sequences (result dependent on the seed length: 20, 22, and 24 nt). However, this represents an extreme underestimation. The short length of reads constituting our libraries (15 to 50 nt) did not allow us to detect the majority of recombination events. According to Routh and Johnson (32), a theoretical maximum efficiency of recombination detection with the ViReMa algorithm can be estimated using the following calculation: there are 49 possible cutting sites in a 50-nt read at which a recombination may occur. With a search seed of 20 nt, recombination events occurring in the first or last 19 cutting sites of the reads will not be detected, leaving 11 possible sites. Therefore, a theoretical maximum efficiency of recombination can be calculated as $11/49 \times 100 = 22.4\%$. Using the same formula, the theoretical maximum efficiency of recombination detection using a seed length of 22 nt is 14%, and using a seed length of 24 nt it is 6%. Correction for theoretical algorithm efficiency was made by dividing the detected number of recombinant viral sequences by the approximate theoretical efficiency of the algorithm. Because our data set contained different read lengths, we could perform only a rough estimation using the most represented read length (50 bp) for the corrections (also being the most conservative one). Thus, considering the theoretical efficiency of recombination detection, we can roughly estimate that the "real" proportion of recombinant sequences amounts to 7 to 15% (Table 1). This is in accordance with the study of Routh and Johnson (32), where, with a similar approach, 8 to 13% of sequences encapsidated in flock house virus (FHV) particles could be annotated as recombinant viral sequences. Detailed results are presented in Table 1.

For further analysis, we filtered the detected recombinant

Downloaded from http://jvi.asm.org/ on May 7, 2015 by guest

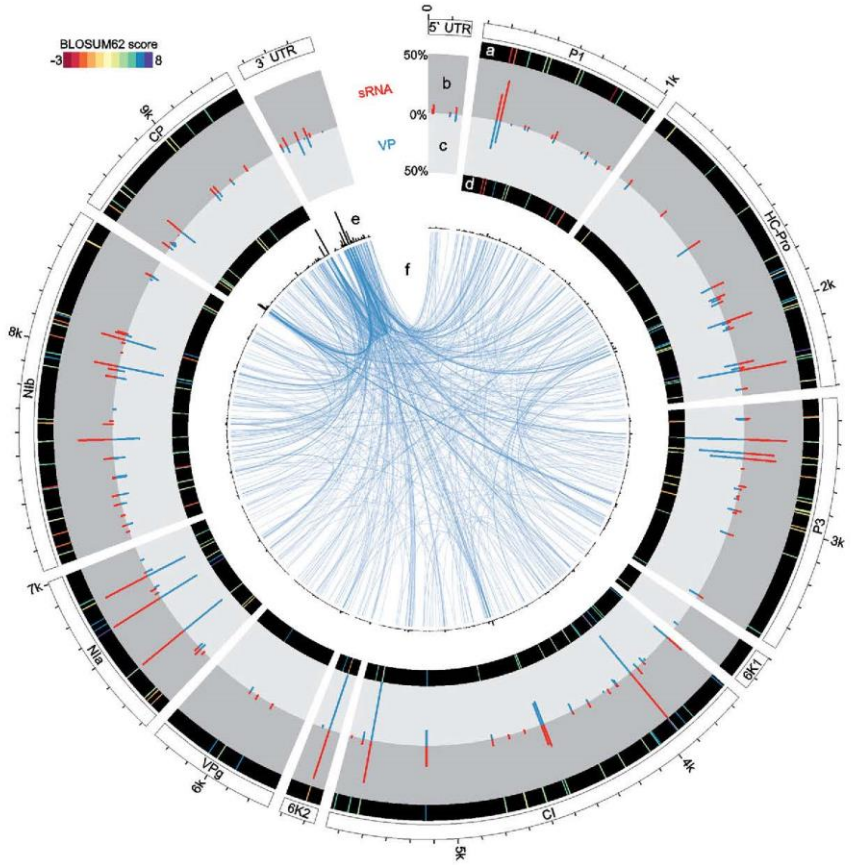


FIG 5 Circular representation of PVY genome with plotted frequencies of SNPs in the sRNA and VP pools, corresponding BLOSUM62 amino acid substitution scores, and recombination points detected by ViReMa (only for the VP pool). (a) BLOSUM62 scores for amino acid changes corresponding to SNPs in the sRNA pool. (b) Square-root-transformed relative frequencies of SNPs detected in the sRNA pool. (c) Square-root-transformed relative frequencies of SNPs detected in the VP pool. (d) BLOSUM62 scores for amino acid changes corresponding to SNPs in the VP pool. (e) Histogram of frequencies of recombination points in 10-nt windows across the PVY genome (scale maximum is 380 events). (f) Visualization of recombination breakpoints detected in the VP pool. The link connects the 3′ and 5′ recombination breakpoint, and its thickness corresponds to the frequency of the event.

events conservatively: only the recombination events confirmed by at least two nonidentical reads in both replicates were considered reliable. This way, we excluded potential experimental and analysis artifacts. With the seed length set to 20 nt, 2,295 (706 unique) filtered nonhomologous recombination events were revealed. Plotting the observed recombination points on the PVY genome revealed a notable hot spot of such events present in the 3′-untranslated region (UTR) and coat protein (CP) region of the genome (Fig. 5e and f). When the specificity of recombination detection was increased by setting seed lengths to 22 nt and 24 nt, the number of detected recombination points decreased to 668 (217 unique) and 38 (16 unique), respectively. Nevertheless, identical hotspots were identified in the viral genome (Fig. 6a). The predicted hotspots (Fig. 5e and f) coincide with the notable peak in the depth of sequencing coverage (Fig. 3e). We tested if the observed pattern could be explained only as a correlation between the number of detected recombination events and the depth of coverage. No notable correlation could be observed (Fig. 6b). Manual inspection of recombination events confirmed the reliability of the ViReMa approach, since all 15 tested simulated recombinant sequences were covered by recombinant reads. No recombinant events could be detected (using the same simulated sequences) within the sRNA pool (see Materials and Methods for details about the analysis).

Downloaded from http://jvi.asm.org/ on May 7, 2015 by guest

TABLE 1 Recombinant reads detected by ViReMa in VP1 and VP2 samples before and after correction for theoretical algorithm efficiency using different analysis settings (seed lengths)

Parameter	Value for: VP1	VP2
Total no. of reads (>25 nt)	5,058,459	4,413,342
No. of reads mapping to PVY	4,198,138	3,704,865
No. of reads mapping to potato	9,843	8,843
No. of reads for recombination analysis	850,478	699,634
Seed length set to 20 nt		
No. of recombinant reads	171,191	132,950
Proportion (%) of total no. of reads	3.38	3.01
Proportion (%) of total no. of reads, corrected for theoretical algorithm efficiency (/0.22)[a]	15.38	13.69
No. of reads after filtering[b]	2,295 (706 unique)	
Seed length set to 22 nt		
No. of recombinant reads	83,840	63,508
Proportion (%) of total no. of reads	1.66	1.44
Proportion (%) of total no. of reads, corrected for theoretical algorithm efficiency (/0.14)[a]	11.84	10.28
No. of reads after filtering[a]	668 (217 unique)	
Seed length set to 24 nt		
No. of recombinant reads	27,757	20,458
Proportion (%) of total no. of reads	0.55	0.46
Proportion (%) of total no. of reads, corrected for theoretical algorithm efficiency (/0.06)[a]	9.15	7.73
No. of reads after filtering[b]	38 (16 unique)	

[a] The correction was made by dividing the number of recombinant reads by theoretical algorithm efficiency, which is given in parentheses.
[b] Recombination events have to be detected in two nonidentical reads in both replicates.

DISCUSSION

Through a carefully planned experimental and analysis approach, we compared the diversity of two separate but coexisting viral sequence pools within a plant: virus-derived small RNAs and RNA isolated from purified viral particles. Both of the approaches tested here proved to be efficient for the enrichment of viral nucleic acids from infected plants. However, the isolation of viral particles was shown to be superior when considering the purity and relative amount of viral nucleic acids. It also performed better when the reads were used for *de novo* reconstruction of the consensus viral genome sequence. Thus, sequencing RNA isolated from purified viral particles is highly efficient when characterizing unknown viruses. The main drawback of this approach could be the length of the purification process. This was, however, greatly reduced in our experimental setup by using a novel purification approach based on CIM monolithic chromatography (24). On the other hand, the sRNA sequencing technique is a more time-efficient and generic approach. It allows us to detect many different viruses without a special protocol modification, and as such it represents a first choice for virus screening or diagnostic usage.

To explore and compare the mutational landscape of within-plant virus populations, we predicted SNPs present in both of the investigated pools. Using a stringent filtering approach, we were able to reliably detect variants with frequency as low as 0.06%, which is the highest resolution, obtained in whole-genome studies of plant virus populations to date. Confirmation of identified variants in two independent pools of sequences provides additional reliability for investigating the diversity of viral populations. We showed that the SNPs of two distinct viral sequence pools overlap greatly. All of the mutations with frequency higher than 1.6% were detected in both pools, implicating a highly conserved mutational landscape between VP and sRNA pools. Nevertheless, both of the pools contained some unique SNPs, which in all of the cases had low frequencies (up to 1.6% in the sRNA pool, up to 0.6% in the VP pool).

The occurrence of these pool-unique SNPs can be explained by several scenarios. First, it could be a result of a genetic drift in localized virus populations within a plant. Even though special care was taken to ensure representative sampling for both of the pools (Fig. 1), it is likely that mutations from a very recent and/or localized replication event would not occur in both of the pools. Moreover, in the case of sRNAs, unique SNPs could arise as random mutations through viral genome replication (dsRNA) but may later be negatively selected during viral genome encapsidation. Second, they also could arise through errors introduced by host RNA-dependent RNA polymerase during the generation of secondary vsiRNAs (18). The possibility of technical artifacts should not be completely excluded, because the extremely low input concentrations of the current protocols do not allow us to exactly equilibrate the amount of molecules between the pools before the ligation reaction. This could result in a dropout of low-frequency variants in some of the libraries. Nevertheless, since we observed unique SNPs in both of the pools and the frequencies of SNPs between technical replicates and between the pools have a good correlation, it is not likely that observed unique SNPs would arise solely as a technical artifact of unequilibrated molecule counts.

Several studies reported prominent hotspots and cold spots in depths of vsiRNAs coverage across the viral genome (35, 36). Different explanations have been suggested for these observations, including biological explanations, e.g., connecting the hotspots with predicted secondary structures in viral genomes (23, 36), as well as library preparation-connected ligation biases (34). In other cases (37, 38) such patterns were not observed. In our study, vsiRNAs were uniformly distributed across the PVY genome without prominent coverage hotspots or cold spots (Fig. 3e). In addition, the evenness of coverage was greater for sRNA samples than for chemically fragmented viral RNA (Fig. 3e). The pattern of vsiRNA coverage is specific for different plant-virus systems and should be evaluated in each specific case. However, one can speculate that even if notable cold/hotspots in coverage are observed, they do not greatly affect the relative frequency of detected variants but only reduce/enhance the resolution in such regions of the viral genome. Moreover, such a variation could be taken into account, when comparing the samples, filtering the positions with uneven coverage across the samples.

Recently, some studies have shown that recombination events should be approximately as common as mutations in the evolution of RNA viruses (32, 39, 40). A recent study (41) illustrated this phenomenon in the case of tobacco etch virus, a member of the same (*Potyvirus*) genus as PVY. We used an algorithm (32) which allowed us to detect nonhomologous recombinant reads in the VP pool. The reads in the sRNA pool were too short to be included in the analyses. The results of our study implicated that roughly 7 to 15% of the VP reads obtained in this experiment represent recom-

Downloaded from http://jvi.asm.org/ on May 7, 2015 by guest

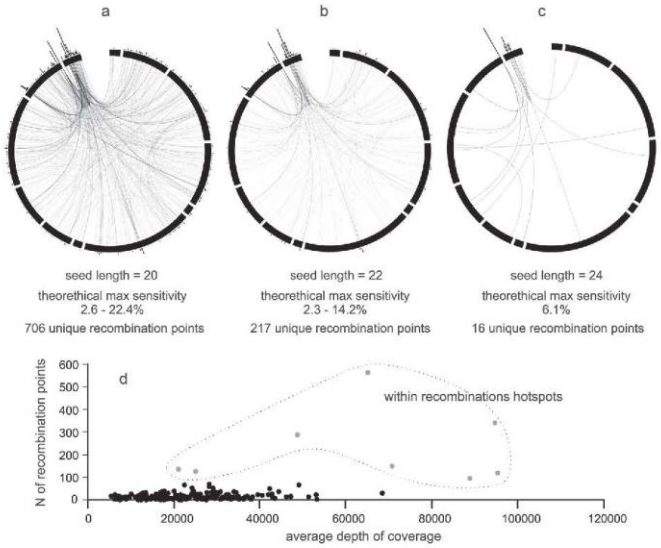


FIG 6 Detection of recombination hotspots. (a to c) Recombination points detected by ViReMa (links connect 3′ and 5′ recombination breakpoints) in the PVY genome (black squares designate genes and UTRs) and their frequency in 10-nt-wide windows across the PVY genome (histogram above the genome segments) with seed length set to 20 nt (a), 22 nt (b), and 24 nt (c). Longer seed length increases the specificity and decreases the sensitivity of the method. The number of recombination points stated under each circular diagram represents the number of detected unique recombination events. (b) Number of recombination points (seed length set to 20 nt) within each of the 10-nt-windows across the PVY genome plotted against the corresponding depth of coverage.

binant viral sequences originating from nonhomologous recombination events. This is congruent with the data obtained for flock house virus (FHV), where a similar methodology was used (32). A relatively high proportion of nonhomologous recombinant sequences in virions highlights an important but rather understudied part of the variability in virus populations. Further studies should be made that employ an approach which would allow the straightforward distinction of *in vivo* recombinants from artifactual recombinations that may have been introduced through library preparation. Here, we have used a conservative approach to reliably distinguish *in planta* recombination events from possible artifactual recombinations by stringently filtering the recombination events based on their occurrence in both of the sequenced technical replicates. This allowed us to perform a high-resolution mapping of nonhomologous recombination events on the viral genome, which indicated hot spots for recombination events in the 3′-untranslated region of the PVY genome as well as in the 5′ and 3′ ends of the coat protein gene. The mechanism and relevance behind this interesting pattern should be further explored.

In the potato-PVY system, both of the sequence pools investigated allowed reliable determination of low-frequency SNPs in virus populations. We showed that sequencing virus-derived small RNAs captures a highly detailed picture of the within-plant virus mutational landscape compared to the RNA packed in viral particles. The deep resolution obtained with each of the approaches would allow us to track in detail the fluctuations of variant frequencies in population studies of plant virus evolution and emergence. Both tested methods have their advantages and disadvantages; small RNA sequencing would allow an easier, quicker, and more generic method to explore within-plant virus population structure, whereas deep sequencing of RNA isolated from viral particles provides an additional insight into the recombination events, which are an important but often overlooked source of diversity in viral populations. The presented comparison can serve as a firm baseline for the employment of both methods in within-plant virus population studies.

ACKNOWLEDGMENTS

The work was supported by the Slovenian Research Agency through grant L4-5525. D.K. is a recipient of a Ph.D. research grant from the Slovenian Research Agency.

We thank Lidija Matičič and Polona Kogovšek for help with the preparation of plant material and virus isolates.

REFERENCES

1. Domingo E, Sheldon J, Perales C. 2012. Viral quasispecies evolution. Microbiol Mol Biol Rev 76:159–216. http://dx.doi.org/10.1128/MMBR.05023-11.
2. McElroy K, Thomas T, Luciani F. 2014. Deep sequencing of evolving pathogen populations: applications, errors, and bioinformatic solutions. Microb Inform Exp 4:1. http://dx.doi.org/10.1186/2042-5783-4-1.
3. Wright CF, Morelli MJ, Thébaud G, Knowles NJ, Herzyk P, Paton DJ, Haydon DT, King DP. 2011. Beyond the consensus: dissecting within-host viral population diversity of foot-and-mouth disease virus by using next-generation genome sequencing. J Virol 85:2266–2275. http://dx.doi.org/10.1128/JVI.01396-10.
4. Routh A, Domitrovic T, Johnson JE. 2012. Host RNAs, including transposons, are encapsidated by a eukaryotic single-stranded RNA virus. Proc Natl Acad Sci U S A 109:1907–1912. http://dx.doi.org/10.1073/pnas.1116168109.

Downloaded from http://jvi.asm.org/ on May 7, 2015 by guest

5. Morelli MJ, Wright CF, Knowles NJ, Juleff N, Paton DJ, King DP, Haydon DT. 2013. Evolution of foot-and-mouth disease virus intra-sample sequence diversity during serial transmission in bovine hosts. Vet Res 44:1–15. http://dx.doi.org/10.1186/1297-9716-44-1.
6. Ohshima K, Akaishi S, Kajiyama H, Koga R, Gibbs AJ. 2010. Evolutionary trajectory of turnip mosaic virus populations adapting to a new host. J Gen Virol 91:788–801. http://dx.doi.org/10.1099/vir.0.016055-0.
7. Kim T, Youn MY, Min BE, Choi SH, Kim M, Ryu KH. 2005. Molecular analysis of quasispecies of Kyuri green mottle mosaic virus. Virus Res 110:161–167. http://dx.doi.org/10.1016/j.virusres.2005.02.007.
8. Acosta-Leal R, Bryan BK, Rush CM. 2010. Host effect on the genetic diversification of beet necrotic yellow vein virus single-plant populations. Phytopathology 100:1204–1212. http://dx.doi.org/10.1094/PHYTO-04-10-0103.
9. Ali A, Roossinck MJ. 2010. Genetic bottlenecks during systemic movement of Cucumber mosaic virus vary in different host plants. Virology 404:279–283. http://dx.doi.org/10.1016/j.virol.2010.05.017.
10. Schneider W, Roossinck M. 2001. Genetic diversity in RNA virus quasispecies is controlled by host-virus interactions. J Virol 75:6566–6571. http://dx.doi.org/10.1128/JVI.75.14.6566-6571.2001.
11. Martínez F, Lafforgue G, Morelli MJ, González-Candelas F, Chua N-H, Daròs J-A, Elena SF. 2012. Ultradeep sequencing analysis of population dynamics of virus escape mutants in RNAi-mediated resistant plants. Mol Biol Evol 29:3297–3307. http://dx.doi.org/10.1093/molbev/mss135.
12. Fabre F, Montarry J, Coville J, Senoussi R, Simon V, Moury B. 2012. Modelling the evolutionary dynamics of viruses within their hosts: a case study using high-throughput sequencing. PLoS Pathog 8:e1002654. http://dx.doi.org/10.1371/journal.ppat.1002654.
13. Montarry J, Doumayrou J, Simon V, Moury B. 2011. Genetic background matters: a plant-virus gene-for-gene interaction is strongly influenced by genetic contexts. Mol Plant Pathol 12:911–920. http://dx.doi.org/10.1111/j.1364-3703.2011.00724.x.
14. Morroni M, Jacquemond M, Tepfer M. 2013. Deep sequencing of recombinant virus populations in transgenic and nontransgenic plants infected with cucumber mosaic virus. Mol Plant Microbe Interact 26:801–811. http://dx.doi.org/10.1094/MPMI-02-13-0057-R.
15. Tugume AK, Mukasa SB, Kalkkinen N, Valkonen JPT. 2010. Recombination and selection pressure in the ipomovirus sweet potato mild mottle virus (Potyviridae) in wild species and cultivated sweet potato in the centre of evolution in East Africa. J Gen Virol 91:1092–1108. http://dx.doi.org/10.1099/vir.0.016089-0.
16. Yu X-Q, Jia J-L, Zhang C-L, Li X-D, Wang Y-J. 2010. Phylogenetic analyses of an isolate obtained from potato in 1985 revealed potato virus X was introduced to China via multiple events. Virus Genes 40:447–451. http://dx.doi.org/10.1007/s11262-010-0468-5.
17. Satya VR, DiCarlo J. 2014. Edge effects in calling variants from targeted amplicon sequencing. BMC Genomics 15:1073. http://dx.doi.org/10.1186/1471-2164-15-1073.
18. Llave C. 2010. Virus-derived small interfering RNAs at the core of plant-virus interactions. Trends Plant Sci 15:701–707. http://dx.doi.org/10.1016/j.tplants.2010.09.001.
19. Kreuze JF, Perez A, Untiveros M, Quispe D, Fuentes S, Barker I, Simon R. 2009. Complete viral genome sequence and discovery of novel viruses by deep sequencing of small RNAs: a generic method for diagnosis, discovery and sequencing of viruses. Virology 388:1–7. http://dx.doi.org/10.1016/j.virol.2009.03.024.
20. Wu Q, Luo Y, Lu R, Lau N, Lai EC, Li W-X, Ding S-W. 2010. Virus discovery by deep sequencing and assembly of virus-derived small silencing RNAs. Proc Natl Acad Sci U S A 107:1606–1611. http://dx.doi.org/10.1073/pnas.0911353107.
21. Seguin J, Rajeswaran R, Malpica-López N, Martin RR, Kasschau K, Dolja VV, Otten P, Farinelli L, Pooggin MM. 2014. De novo reconstruction of consensus master genomes of plant RNA and DNA viruses from siRNAs. PLoS One 9:e88513. http://dx.doi.org/10.1371/journal.pone.0088513.
22. Steyer A, Gutiérrez-Aguire I, Kolenc M, Koren S, Kutnjak D, Pokorn M, Poljšak-Prijatelj M, Racki N, Ravnikar M, Sagadin M, Fratnik Steyer A, Toplak N. 2013. High similarity of novel orthoreovirus detected in a child hospitalized with acute gastroenteritis to mammalian orthoreoviruses found in bats in Europe. J Clin Microbiol 51:3818–3825. http://dx.doi.org/10.1128/JCM.01531-13.
23. Candresse T, Filloux D, Muhire B, Julian C, Galzi S, Fort G, Bernardo P, Daugrois J-H, Fernandez E, Martin DP, Varsani A, Roumagnac P. 2014. Appearances can be deceptive: revealing a hidden viral infection with deep sequencing in a plant quarantine context. PLoS One 9:e102945. http://dx.doi.org/10.1371/journal.pone.0102945.
24. Rupar M, Ravnikar M, Tušek-Žnidarič M, Kramberger P, Glais L, Gutiérrez-Aguirre I. 2013. Fast purification of the filamentous Potato virus Y using monolithic chromatographic supports. J Chromatogr A 1272:33–40. http://dx.doi.org/10.1016/j.chroma.2012.11.058.
25. Scholthof K-BG, Adkins S, Czosnek H, Palukaitis P, Jacquot E, Hohn T, Hohn B, Saunders K, Candresse T, Ahlquist P, Hemenway C, Foster GD. 2011. Top 10 plant viruses in molecular plant pathology. Mol Plant Pathol 12:938–954. http://dx.doi.org/10.1111/j.1364-3703.2011.00752.x.
26. Chen Y-R, Zheng Y, Liu B, Zhong S, Giovannoni J, Fei Z. 2012. A cost-effective method for Illumina small RNA-Seq library preparation using T4 RNA ligase 1 adenylated adapters. Plant Methods 8:41. http://dx.doi.org/10.1186/1746-4811-8-41.
27. Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18:821–829. http://dx.doi.org/10.1101/gr.074492.107.
28. Core Team R. 2013. R: a language and environment for statistical computing. 3.0.2. R Foundation for Statistical Computing, Vienna, Austria.
29. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079. http://dx.doi.org/10.1093/bioinformatics/btp352.
30. Wilm A, Aw PPK, Bertrand D, Yeo GHT, Ong SH, Wong CH, Khor CC, Petric R, Hibberd ML, Nagarajan N. 2012. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic Acids Res 40:11189–11201. http://dx.doi.org/10.1093/nar/gks918.
31. Henikoff S, Henikoff JG. 1992. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A 89:10915–10919. http://dx.doi.org/10.1073/pnas.89.22.10915.
32. Routh A, Johnson JE. 2014. Discovery of functional genomic motifs in viruses with ViReMa–a virus recombination mapper–for analysis of next-generation sequencing data. Nucleic Acids Res 42:e11. http://dx.doi.org/10.1093/nar/gkt916.
33. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra Ma. 2009. Circos: an information aesthetic for comparative genomics. Genome Res 19:1639–1645. http://dx.doi.org/10.1101/gr.092759.109.
34. Sorefan K, Pais H, Hall AE, Kozomara A, Griffiths-Jones S, Moulton V, Dalmay T. 2012. Reducing ligation bias of small RNAs in libraries for next generation sequencing. Silence 3:4. http://dx.doi.org/10.1186/1758-907X-3-4.
35. Li R, Gao S, Hernandez AG, Wechter WP, Fei Z, Ling K-S. 2012. Deep sequencing of small RNAs in tomato for virus and viroid identification and strain differentiation. PLoS One 7:e37127. http://dx.doi.org/10.1371/journal.pone.0037127.
36. Aregger M, Borah BK, Seguin J, Rajeswaran R, Gubaeva EG, Zvereva AS, Windels D, Vazquez F, Blevins T, Farinelli L, Pooggin MM. 2012. Primary and secondary siRNAs in geminivirus-induced gene silencing. PLoS Pathog 8:e1002941. http://dx.doi.org/10.1371/journal.ppat.1002941.
37. Wang Y, Cheng X, Wu X, Wang A, Wu X. 2014. Characterization of complete genome and small RNA profile of pagoda yellow mosaic associated virus, a novel badnavirus in China. Virus Res 188:103–108. http://dx.doi.org/10.1016/j.virusres.2014.04.006.
38. Fuentes S, Heider B, Tasso RC, Romero E, Zum Felde T, Kreuze JF. 2012. Complete genome sequence of a potyvirus infecting yam beans (Pachyrhizus spp.) in Peru. Arch Virol 157:773–776. http://dx.doi.org/10.1007/s00705-011-1214-6.
39. Routh A, Ordoukhanian P, Johnson JE. 2012. Nucleotide-resolution profiling of RNA recombination in the encapsidated genome of a eukaryotic RNA virus by next-generation sequencing. J Mol Biol 424:257–269. http://dx.doi.org/10.1016/j.jmb.2012.10.005.
40. Froissart R, Roze D, Uzest M, Galibert L, Blanc S, Michalakis Y. 2005. Recombination every day: abundant recombination in a virus during a single multi-cellular host infection. PLoS Biol 3:e89. http://dx.doi.org/10.1371/journal.pbio.0030089.
41. Tromas N, Zwart MP, Poulain M, Elena SF. 2014. Estimation of the in vivo recombination rate for a plant RNA virus. J Gen Virol 95:724–732. http://dx.doi.org/10.1099/vir.0.060822-0.

Downloaded from http://jvi.asm.org/ on May 7, 2015 by guest

2.4 Time-Sampled Population Sequencing Reveals the Interplay of Selection and Genetic Drift in Experimental Virus Evolution (Submitted Manuscript)

Denis Kutnjak, Santiago F. Elena, Maja Ravnikar

Submitted manuscript

In this study, we have employed deep sequencing of sRNAs to track the dynamics of PVY evolution during the experimental evolution in different potato genotypes. Using the time-sampled population sequencing approach we were able to follow the trajectories of variants in the viral populations in time and use this information to cluster the variants into groups, finally reconstructing the haplotypes. The experiment revealed a pattern of interplay between natural selection and genetic drift in evolution of viral populations and indicated possible diversifying selection acting on the virus in potato genotypes efficient in accumulation of salicylic acid.

The PhD candidate designed the experiment, performed the serial passages of the virus on different potato genotypes, sampled the plants, performed RNA isolations and virus titres quantification. He performed analysis of sequencing data and population genetics analyses. He drafted the manuscript.

Time-sampled population sequencing reveals the interplay of selection and genetic drift in experimental virus evolution

Denis Kutnjak[1,2*], Santiago F. Elena[3,4,5], Maja Ravnikar[1*]

[1] Department of Biotechnology and Systems Biology, National Institute of Biology, 1000 Ljubljana, Slovenia

[2] Jožef Stefan International Postgraduate School, 1000 Ljubljana, Slovenia

[3] Instituto de Biología Molecular y Celular de Plantas, Consejo Superior de Investigaciones Científicas-Universidad Politécnica de Valencia, Ingeniero Fausto Elio s/n, 46022 València, Spain

[4] Instituto de Biología Integrativa y de Sistemas, Consejo Superior de Investigaciones Científicas-Universitat de València, Catedrático Agustín Escardino 9, 46980 Paterna, València, Spain

[5] The Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA

* For correspondence: denis.kutnjak@nib.si, maja.ravnikar@nib.si

Abstract

RNA viruses are one of the fastest evolving biological entities. Within their hosts, they exist as genetically diverse populations, which are sculpted by different evolutionary mechanisms, such as mutation, natural selection and genetic drift. To estimate the effect of selection and genetic drift on the evolution of *Potato virus Y* (PVY) populations, we have performed experimental evolution of the virus in several potato genotypes, which differ in defense response to the virus. Using deep sequencing of small RNAs, we followed the temporal dynamics of standing and newly emerged variation, reconstructed theoretical haplotypes in the starting population and revealed the interplay between selection and genetic drift during the evolution of the virus. We detected imprints of strong selection sweeps, narrow genetic bottlenecks, followed by the shift in selected haplotypes and diversifying evolution of the virus in the less permissive environment (less sensitive potato genotypes, efficient in the accumulation of salicylic acid).

27 **Introduction**

28 Mutation, recombination, natural selection, and genetic drift are the processes that shape the
29 diversity and evolution of virus populations. Due to their error-prone replication, RNA viruses
30 mutate rapidly and thus exist within hosts as mutant swarms (referred to also as quasispecies),
31 derived from ancestral infecting genomes (Andino and Domingo, 2015). Fast replication,
32 coupled with high mutation and recombination rates (Froissart *et al.*, 2005; Simon-Loriere and
33 Holmes, 2011; Tromas *et al.*, 2014), along with large population sizes, enable RNA viruses to
34 quickly adapt to changing environments and it also render them interesting for studies of basic
35 evolutionary processes, *e.g.*, mutation, selection and genetic drift. Whereas selection
36 describes the competition between entities with different fitness generated by mutation within
37 the population, random genetic drift represents stochastic changes in population structure due
38 to limited population sizes. The latter are, for viruses, often a consequence of bottlenecks
39 during transmission and within-host movement (Zwart and Elena, 2015). Additionally, it has
40 been suggested that the evolution of large diverse populations, like viral, is strongly affected
41 by linked selection of variants on shared haplotype (*i.e.*, epistasis) and the interactions
42 between different haplotypes (*i.e.*, clonal interference) in the population, and not so much by
43 the effects of new spontaneous mutations (Neher, 2013).

44 For viruses, host switches can represent important ecological changes, posing different, and
45 sometimes conflicting, selection pressures on the virus population (Elena, Fraile and García-
46 Arenal, 2014). These selective forces can leave an imprint in the viral genomes, which can
47 show parallel fixed mutations in viral lineages evolved in the same host species (Longdon *et*
48 *al.*, 2014; Vassilakos *et al.*, 2016). Moreover, evolutionary experiments have shown different
49 degrees of parallel evolution after the serial transmission of viruses in different environments,
50 such as adaptation of *Chikungunya virus* (CHIKV) to different vector species (Stapleford *et al.*,
51 2014), adaptation of *Vesicular stomatitis virus* (VSV) to different mammal cell types (Turner
52 and Elena, 2000; Remold, Rambaut and Turner, 2008), adaptation of *Influenza A virus* (IAV)
53 in the presence of a drug (Foll *et al.*, 2014) and others (Bull *et al.*, 1997; Bedhomme, Lafforgue
54 and Elena, 2012). Recently, evidence for host specific convergent mutations were shown also
55 during the experimental evolution of *Tobacco etch virus* (TEV) in different ecotypes of the same
56 species (Hillung *et al.*, 2014).

57 Detection of mutations and characterization of changes in viral mutant swarms during virus
58 evolution experiments have dramatically improved with the advent of next generation
59 sequencing (NGS), enabling to detect even variants present at very low frequencies within the
60 population (Acevedo, Brodsky and Andino, 2014). Most of the virus evolution studies rely on
61 snapshots of viral diversity in the particular time point. Such an approach has limitations to

detect selection patterns (relying on the ratio between observed synonymous and nonsynonymous mutations). Implementation of time-sampled approaches in evolution experiments (Kessinger, Perelson and Neher, 2013; Acevedo, Brodsky and Andino, 2014; Foll *et al.*, 2014; Bordería *et al.*, 2015; Foll, Shim and Jensen, 2015), on the other hand, enable to follow trajectories of mutations in a larger time frame and thus to infer some population genetics parameters, such as variants selection coefficients (s; a measure of the strength of natural selection) and effective population size (N_e; a measure of the strength of genetic drift) (Foll *et al.*, 2014; Foll, Shim and Jensen, 2015).

In this research, we have focused on one of the ten most important plant viruses, *Potato virus Y* (PVY; genus *Potyvirus*, family: *Potyviridae*) (Scholthof *et al.*, 2011), which cause a substantial yield loss in potato production worldwide (Karasev and Gray, 2013). PVY is as single stranded positive-sense RNA virus, comprising a high diversity of different recombinant strains. Specifically, we have focused on the recombinant NTN strain, which causes the potato tuber necrotic ringspot disease (PTNRD). Response of potato to infection with PVY have been extensively studied at different levels (Baebler *et al.*, 2009; Kogovšek *et al.*, 2010, 2016; Stare *et al.*, 2015) and it was shown that hormone signaling has an important role in perception and response to the virus infection; salicylic acid (SA) representing one of the hallmarks of the response (Singh *et al.*, 2004; Baebler *et al.*, 2011, 2014). Disturbance in SA signaling was shown to affect accumulation of virus and the appearance/severity of the disease symptoms (Baebler *et al.*, 2011; Stare *et al.*, 2015). We have therefore decided to perform experimental evolution of PVY in potato plants accumulating SA (representing a less permissive environment) and potato plants depleted in SA (thus representing a more permissive environment), trying to elucidate the effect of these different environments on virus population structure and its evolution.

We have recently demonstrated that virus-derived small interfering RNAs (vsiRNAs) can accurately describe the population structure of PVY in potato (Kutnjak *et al.*, 2015). By combining the high resolution of Illumina small RNA deep sequencing and a time-sampled experimental evolutionary approach, we have now tracked the dynamics of PVY populations during serial passages of 10 separate lineages in 3 different potato (*Solanum tuberosum* L.) genotypes (Fig. 1): Pentland squire (P), Désirée (D) and NahG-Désirée (Dn). P is a tolerant cultivar, accumulating virus, but not showing any symptoms, D sporadically shows symptoms after infection with PVY^{NTN} but it is asymptomatic in our experimental setup, Dn is a transgenic genotype derived from D transformed with the *nahG* gene (salicylate hydroxylase), which converts SA into catechol; it is sensitive to PVY and expresses strong symptoms in a form of necrotic lesions on the leaves (Baebler *et al.*, 2011). The source inoculum used to initiate all of the evolving lineages was derived from P plants (Fig. 1). Rather than starting from a cloned

viral genome, we have constituted a virus inoculum with high genetic variability, which allowed us to study the relative contribution to virus adaption and diversification of standing variation in the starting population and of new variation created during the evolution experiment. We have observed signatures of selection and genetic drift and their interplay in the evolution of PVY. Additionally, we have detected diversifying evolution of the virus in the less permissive host, *i.e.*, in the genotypes accumulating SA.

Results

Unexpected drop in viral titer in one of the serial passages resulted in a strong genetic bottleneck

We have first measured virus accumulation in the three different potato genotypes by measuring its titer 6, 9, 12, 15 and 21 days post inoculation (dpi) in the first passage. Virus accumulation was correlated to the sensitivity of the genotypes; the most sensitive genotype (Dn) accumulating the highest amount of the virus, the least sensitive (P) the lowest. Virus titer reached a plateau in all three genotypes 21 dpi, at the time point at which sampling for virus population sequencing was performed. The differences in virus accumulation in first passage 21 dpi between the three cultivars were between one to two orders of magnitude, the difference between the smallest (P) and biggest values (Dn) being around 60-fold (Fig. 2a, Supplementary table 3).

Viral titer was then measured 21 dpi for each passage (Fig. 2b) and it remained relatively stable for all lineages until the 5th passage, where we observed a significant drop in virus titer in all of the potato genotypes. The underlying reason for the observed drop is unclear. We hypothesize it originated from plant inoculation process or from unexpected variation in some environmental factor affecting only this particular passage. The virus titer for the lineage DnxD1 (alternating host plants between D and Dn) was fluctuating, corresponding to the host cultivar (blue line on Fig. 2b).

We have estimated the effective population sizes (N_e) as well as selection coefficients (s) from time sampled deep sequencing data using a Wright Fisher Approximate Bayesian Computations (WF-ABC) approach (Foll, Shim and Jensen, 2015). s represents a measure of relative fitness of the competing individual against the alternative genotype and as such reflects the strength of natural selection. N_e is, on the other hand, an estimator of a strength of a genetic drift, since it represents a number of virus genomes contributing directly to the progeny, being determined by the size of the population bottlenecks (Zwart and Elena, 2015). To determine if the observed drop in virus titer in the 5th passage was reflected also in a narrower genetic bottleneck, first part of the analysis (N_e estimation) was performed separately for pairs of time points (inoculum-1st passage; 1st passage-4th passage; 4th passage-5th

passage; 5th passage-6th passage). The estimations of N_e for pairs of time points (Supplementary Table 2) showed similar pattern to the one observed in viral titer measurements, suggesting that a strong drop in viral titer at the 5th passage affected N_e in this time point (Fig. 2c). Association between viral titer and N_e was more evident for the very low viral titers and was not as much observed at higher titers (Fig. 2c).

Selection analysis and clustering reveals cohorts of linked SNPs and enable reconstruction of haplotypes in the source inoculum

For the second part of the WF-ABC analysis (estimation of *s* values) we have excluded samples contributing to very low estimated N_e values from the calculations. Posterior distribution of *s* values for majority of single nucleotide polymorphisms (SNPs) were wide (including 0), thus we were unable to reject neutrality for such variants. Low statistical power was probably a consequence of the low number of included time points and possibly underestimated N_e values (see Discussion for details). However, for several variants (1-11 SNPs per lineage) neutrality was rejected (95% highest posterior density (HPD) excluded 0). We have observed loci under both: strong positive and negative selection. The latter could be observed since source inoculum in this experiment contained highly diverse PVY population, enabling selection to fix some variants while removing others. The distribution of posterior mean of *s* for new variants was more narrow (centered around 0) compared to the distribution of *s* values for standing variants (Supplementary Fig. 2). No new variants under significant negative selection could be detected, likely due to the quick elimination of such variants from the population. In contrast, some new variants with highly positive *s* could be detected, however, due to the limited statistical power, neutrality was rejected only for one of these variants using our stringent criteria. The results showed that the distributions of *s* had similar shape for all 10 lineages (Fig. 4). Plotting *s* values of variants onto the PVY genome revealed no clustering pattern in different parts of the genome (Fig. 4).

The sign of selection for a particular variant could differ among lineages, sometimes showing distinct patterns of interaction (positive or negative) between the variants. To better understand the patterns of interaction and possible linkage between the observed variants we have employed a clustering approach and tried to assign SNPs into groups. Using hierarchical clustering of SNPs trajectories, several cohorts of SNPs showing similar frequency trajectories within the lineages (Supplementary Fig. 1) were detected. The cohorts of SNPs could theoretically represent variants linked on the same genome (haplotypes) or/and other type of linkage/interactions between the variants within the virus population. Based on the clustering results and observed fixed haplotypes (the variants reaching frequency of ~100%) in some of

the lineages (P1, P3, D2, and D3; Supplementary Fig. 1a,c,e,f) we have reconstructed 5 haplotypes likely present in the source inoculum (Fig. 5b).

Further, we have used estimated selection coefficients of SNPs to cluster SNPs on one hand and lineages on the other. With this analyses, we identified 7 groups of SNPs (Fig. 5a, columns clustering). We observed that the 7 groups represent partial haplotypes, since pairs of these groups formed the haplotypes reconstructed by first clustering analysis. Integrating the results of two clustering analyses, we have constructed a diagram representing grouping of observed SNPs into theoretical haplotypes (Fig. 5b). This suggested a co-occurrence of two major mutant swarms in the source inoculum, first containing haplotype 1 and 2 and second containing haplotypes 3, 4 and 5. Two haplotypes from the first mutant swarm likely originate from the ancestral haplotype containing dark green group of mutations (on Fig. 5b), whereas the three haplotypes from the second mutant swarm likely originate from the ancestral variant containing mutations from the purple group (Fig. 5b). A theoretical possibility is that the reconstructed haplotypes were not yet present in the inoculum, but emerged later by recombination. However, since we observe at least some of them in several independent lineages, this does not seem a plausible scenario. Based on the observed frequencies of SNPs in the inoculum, we can speculate that one of the mutations, C4220U, occurred independently in each of the two mutant swarms or was introduced into one of them by recombination.

Interestingly, each of the derived haplotypes from the first mutant swarm contains one additional nonsynonymous mutation: A6828G (NIa/E184G) in red group and A6858G (NIa/N194S) in blue group, both occurring at a similar position in the genome, maybe indicating an important interaction between this part of genome and mutations from the ancestral variant (dark green group). Both mutations are positioned close to the active site of the NIa protease (Phan *et al.*, 2002; Mestre *et al.*, 2003) (especially E184G) and could probably affect the substrate specificity of the protease.

Evolution of PVY populations is governed by the interplay between selection and genetic drift

Strong decrease in nucleotide diversity (π) was detected for two of P lineages (P1 and P3) immediately after the first passage (Fig. 3). Detailed inspection of SNPs trajectories in these two lineages showed that a large group of variants was quickly fixed and others were eliminated from the populations (Supplementary Fig. 1a,c). For most of the other lineages π remained relatively stable or slightly decreased until the 5^{th} passage, where we observed a disturbed pattern, likely resulting from strong genetic drift, a consequence of severely reduced N_e (Fig. 2).

The total variability was then separated into two groups: standing variation (variants already present in the source inoculum) and new variation (newly appeared mutations). Separate

202 analysis of standing and new variation showed that standing variation was decreasing through
203 the experiment (Fig 3b) in most lineages, whereas π for new variants was increasing (Fig. 3c).
204 Only in lineage DnxD1, in which the host was alternated between D an Dn, standing variation
205 increased from 1st to 4th passage, possibly indicating conflicting selection pressures in the two
206 genotypes, which prevented rapid selection of specific variants.

207 Again, strong influence of genetic drift disturbed the predominant pattern in the 5th passage,
208 where we noted a drop in π values in most of the lineages also for new variation (Fig. 3c).
209 However, interestingly, in lineage Dn2 π for new variants sharply increased in this passage
210 and kept rising also in the 6th passage (Fig. 3c). Inspecting the pattern of variation for this
211 lineage more in detail, we have observed that several new nonsynonymous mutations start to
212 rise in frequency in this passage (Supplementary Fig. 1h), likely indicating strong impact of
213 selection after the observed bottleneck. Those variants were not included in the selection
214 anlysis with WF-ABC, since the time points with observed severe bottlenecks were excluded
215 from that analysis.

216 Besides the aforementioned new mutations, also several other new mutations appeared during
217 the experiment, some of them later approaching fixation or being fixed (refer to Supplementary
218 Fig. 1 for visual representation of variants frequency dynamics). In P1 new nonsynonymous
219 mutation U3122G (P3/S240A and PIPO/L68R) appeared in second passage and could act as
220 a driver mutation, causing fixation of haplotype 2 in this lineage. In lineage P3 synonymous
221 mutation U2012C, which was present in inoculum at very low frequency (0.6%), started to rise
222 in frequency and was finally fixed as a part of haplotype 5. The same haplotype started rising
223 in frequency in lineage D2, however, its frequency dropped after the observed bottleneck in
224 the 5th passage and new nonsynonymous mutation A375G (P1/N64S) emerged on haplotype
225 1, which was finally fixed in this lineage in passage 6.

226 Similar shift between the haplotypes after the bottleneck event was also observed in other
227 lineages. For example, in lineage D3, haplotype 1 was rising in frequency until the 5th passage,
228 after which it was nearly eliminated from the population and haplotype 3 approached fixation.
229 Finally, the most striking example of such pattern is observed in lineage Dn2; here, haplotype
230 3 was rising in frequency during the earliest three time points. After the bottleneck, it was
231 removed from the population and haplotype 1 started to rise in frequency. As already
232 mentioned, at the same time, 5 new nonsynonymous mutations started to rise in frequency,
233 suggesting strong impact of selection at this time points. On the other hand, the switch between
234 different haplotypes was not observed in lineage P2: here we observed a reduced frequency
235 of the haplotype 1 in passage 5, but the same haplotype rose in frequency again in the 6th
236 passage.

Imprints of diversifying selection in less permissive genotypes

Using hierarchical clustering of lineages (based on *s* values of SNPs present in all of the lineages, *i.e.*, standing variation) we tried to identify specific and similar patterns of selection in different lineages. This analysis revealed a structure within the 10 evolved lineages (Fig. 5a). The three lineages evolved in the potato genotype depleted in SA (Dn) were clustered into the same group, including also a lineage, where the host genotype was alternated between Dn and D plants (DnxD1). Six lineages from other two cultivars (P and D) did not show any genotype specific grouping. The observed pattern suggest that SNPs included in this analysis were assigned more similar *s* values within group of Dn lineages, than in other two genotypes, possibly indicating more homogenous selection pressures within this group or, conversely, diversifying selection in other two genotypes (D and P, both efficient in accumulation of SA).

Discussion

One of the important limitations of presently most used NGS technologies are short lengths of sequencing reads. This poses a problem, when aiming to reconstruct longer sequences, in our case complete viral haplotypes. Several computational approaches have been developed to address this problem (Zagordi *et al.*, 2011; Prosperi and Salemi, 2012; Di Giallonardo *et al.*, 2014; Prabhakaran *et al.*, 2014; Töpfer *et al.*, 2014), however, they mostly utilize paired-end reads of substantial length, are hard to optimize and could produce high number of false positive results (Schirmer, Sloan and Quince, 2014). In our experiment, we have used extremely short, sRNA reads (20-24 nucleotides), which could not be assembled into haplotypes by any of these methods. However, by using simple clustering analysis of SNPs frequency trajectories in time, we have been able to detect several cohorts of linked SNPs and reconstruct theoretical haplotypes in the inoculum. Refining this approach and maybe connecting it with previously mentioned ones could lead to better haplotype reconstruction algorithms in future.

Secondly, time sampled approach allowed us to estimate effective population sizes (N_e) and selection coefficients (s) of observed SNPs, using the WF-ABC method (Foll, Shim and Jensen, 2015). The method calculates N_e on the assumption that most of the variants in the population are neutral and not linked, thus, since we have observed linked selection, it is expected to be biased and would likely underestimate the actual N_e values in our experiment (even more so in P1 and P3 lineages, where we observed strong selection sweeps). Nevertheless, it corroborated the suspected narrow genetic bottlenecks in the 5th passage of the experiment, suggested also by other observations: strong drop in viral titers in inocula and decreased populations nucleotide diversities. Association between the virus titer in the inoculum and estimated effective population sizes seems to be stronger for very low titers, and

not evident for higher N_e and viral titers, the situation similar to the one observed for another potyvirus, TEV (Zwart, Daròs and Elena, 2012). Further, we have used estimated selection coefficients to perform clustering of SNPs and evolutionary lineages and the SNPs clusters matched well the pattern of haplotypes reconstructed by clustering the SNPs frequency trajectories.

We observed strong selection sweep in two of the lineages (P1 and P3), resulting in rapid reduction of genetic diversity. This could be explained by two not mutually exclusive hypotheses. First, since the source inoculum for the experiment was derived from P plants, one could expect there would be some amount of preadaptation to this cultivar present in a mixed viral population (the so-called quasispecies memory (Ruiz-Jarabo *et al.*, 2000)). This would allow selection to rapidly fix adaptive haplotypes. Secondly, we have observed new or low-frequency mutations emerging and finally being fixed in these two lineages, maybe acting as driver mutations for selected haplotypes. Finally, a combination of both could explain the observations; first preadaptation to P cultivar would move the whole viral population closer to its fitness optimum in the fitness landscape, providing a conditions where new spontaneous mutations would have more chances to provide beneficial fitness effects to the virus.

Observation of these and other possible driver mutations (in lineages P1, P3, D2) emphasize the importance of disentangling the linkage between alleles within the viral population. By observing trajectories and/or selection coefficients of separate SNPs we can count many loci being positively selected in our experiment (*e.g.*, in a lineage P1 in Fig. 4), however, this pattern is likely a consequence of linked selection (also named genetic draft (Neher, 2013)), where one beneficial mutation on certain haplotype in a population drive the haplotype to the fixation. Even though it was suggested that the evolution in large and diverse viral populations would be highly influenced by interactions between the variants and not so much by spontaneous mutations, we observed several examples, where new spontaneous mutations likely "helped" their background haplotypes to rapidly outcompete the others.

A strong genetic bottleneck observed in the 5th passage of our experiment in most of the evolutionary lineages enabled us to further investigate patterns of evolution by observing an interesting interplay between selection and genetic drift. A reduction in genetic diversity was observed in most of the lineages in this time point and, in some cases, haplotypes approaching fixation were eliminated from the population and alternative haplotypes start to rise in frequency after the bottleneck. It has been suggested that genetic drift can translocate virus population in a hypothetical fitness landscape: from a first local fitness peak to another local peak by stochastically eliminating part of the viral population, which was approaching the first local peak (Zwart and Elena, 2015). In such case, strong genetic drift does not necessarily

lead to the maladaptive genotypes, but it helps to shift the population otherwise stuck in the fitness landscape. We observed this phenomenon in our experiment in several lineages as a shift in positively selected genotype before and after the bottleneck event (D2, D3 and Dn2). Most interestingly, in one of the lineages (Dn2) we have observed a rise in number and frequency of new nonsynonymous mutations after the bottleneck, implying a strong positive selection after the genetic diversity of the population have been reduced. Thus, the interplay of selection and genetic drift seems to be especially important in very diverse virus populations, where interactions between the individuals of the mutant swarm can affect the fitness of the virus (population) (Bordería *et al.*, 2015). In such cases, strong genetic drift can provide a stochastic subsample of the population, which establish new interactions and as such translocate the virus in the rugged evolutionary landscape, but also, the reduced complexity of the mutant swarm relaxes the constrains in the fitness landscape, due to the weaker interactions within the population, and enable a rise of new positively selected mutations. Even though the narrow genetic bottleneck observed in 5th passage of our experiment was in fact an experimental artefact, it provided unexpectedly valuable information about the importance of natural selection and genetic drift and their interplay in the evolution of the viral populations. These findings are even more interesting in the light of the awareness that viral transmission events in the nature almost necessarily involve a narrow genetic bottlenecks (Zwart and Elena, 2015), bringing this experiment setup closer to the expected chain of events in the nature.

Finally, we have observed possible diversifying selection in PVY lineages evolved in cultivars efficient in accumulation of SA. Salicylic acid is one of the most important components of the plant response to the biotic stress and it was shown to be involved in suppression of all the main stages of viral infection (replication, cell-to-cell and long-distance movement)(Singh *et al.*, 2004). It has been extensively studied in potato-PVY interaction and shown to be a trigger of multiple changes in activity of downstream metabolic pathways. Recently, a research of sRNA expression levels in the same experimental system as ours (PVY + D/Dn) showed that there are considerably more microRNAs differentially expressed in the D plants after infection with PVY than in Dn plants (M. Križnik *et al.*, manuscript in preparation, 2016), indicating a stronger activation of RNA silencing mechanism in the presence of salicylic acid, a connection suggested also for several other pathosystems (Yu *et al.*, 2003; Pruss *et al.*, 2004; Singh *et al.*, 2004; Alamillo, Saenz and Garcia, 2006). The pattern of diversification in our experiment can be explained by two non-mutually exclusive scenarios: (*i*) diversifying selection in the presence of SA and (*ii*) stronger within-plant virus population bottlenecks occurring in genotypes accumulating SA. The first could have its basis in likely redundant effects of the SA to the virus, including a strong response and reprograming of plant transcriptome (Stare *et al.*, 2015) and RNA silencing, which could hypothetically introduce nonhomogeneous selection

343 pressures on the virus population. These could, in combination with stochastic movement of
344 different viral variants from cell to cell, recently demonstrated for another plant positive-strand
345 RNA virus (Miyashita *et al.*, 2015), result in a stronger selection of different variants in different
346 individuals. The second scenario (stronger effect of genetic drift within a plant) could be
347 supported by observed smaller viral titers in P and D plants, though the differences are not
348 dramatic, yet reproducible. Stronger genetic drift could cause more random fixation of
349 haplotypes in these plants in comparison to Dn plants, which are depleted in SA and
350 accumulate higher titers of the virus. In any case, the observed result (higher degree of
351 diversification in the genotypes accumulating SA) suggest quicker diversification of the virus
352 in somehow less permissive environments. Even though the purposed mechanisms would
353 reduce viral genetic diversity within the host, it could result in an increased diversity of the virus
354 on a higher level (among individuals in the ecosystem), which can hypothetically increase the
355 possibility for the occurrence of new emergent viral isolates.

356 Rather than starting the evolution experiment from scratch using an infectious clone, we have
357 constituted a complex population mix of PVY^{NTN} and observed how it evolved in different potato
358 genotypes on a short time scale. Coupling of this approach and time sampled sRNA deep
359 sequencing provided a different look into the evolutionary mechanisms in comparison to virus
360 evolution experiments using molecular clones as a starting point. In conclusion, we
361 demonstrated that time-sampled population sequencing data enable reconstruction of
362 theoretical haplotypes in a complex starting population mix. We tracked variants' and
363 haplotypes' trajectories in time to estimate the impact of selection and genetic drift on the
364 evolution of the virus. Following the general trends of PVY evolution in the experiment, a
365 narrow genetic bottleneck in one of the passages was detected, coinciding with the shift in the
366 selected virus variants in many lineages, indicating the power of genetic drift to change the
367 coordinates of virus population in its fitness landscape. Fixation of selected haplotypes was
368 often connected with the occurrence and fixation of new mutations, likely acting as drivers for
369 fixation of group of variants connected to the same haplotype. Comparing the evolution of the
370 virus in different potato genotypes, we detected an imprint of diversifying selection of the virus
371 in less permissive genotypes, which are efficient in accumulation of SA. These findings have
372 important implications for understanding of mechanisms that govern the evolution of PVY, as
373 well as other viruses, including the patterns of virus evolution in similar genetic background
374 with different sensitivity for the virus.

375 **Methods**

376 *Potato virus Y experimental evolution setup and sampling*

377 The potato plants used in the experiment were first propagated in stem node culture (*in vitro*).
378 14 days after node segmentation, they were transferred to soil in a growth chamber, and kept
379 at 21 ±2 °C in the light and 18 ±1 °C in the dark, at a relative humidity of 75% ±2%, with 70-90
380 µmol/m^2/s^2 radiation (L36W/77 lamp, Osram, Germany) and a 16-h photoperiod. After growing
381 in soil for 28 days they were either use to prepare the inoculum (systemically infected plantlets)
382 or infected with the inoculum (healthy plantlets) as described below.

383 Twenty-four *Solanum tuberosum* L. cv. Pentland squire plants systemically infected with PVY
384 (strain NTN; isolate NIB V 151) were used to prepare the source inoculum. At the same time
385 two leaf tissue discs (r = 5 mm, ~15 mg per disc) were sampled from each plant, pooled and
386 frozen for RNA isolation. For inoculum preparation, the rest of PVY infected plant material was
387 grinded in 0.04 M sodium phosphate buffer with 0.01 M sodium diethyldithiocarbamate
388 (DIECA) (Sigma–Aldrich) in 1:3 (w/v) tissue to buffer ratio. This inoculum was used to infect
389 24 plants per three different potato genotypes (Pentland squire - P, Désirée - D, NahG- Désirée
390 - Dn; the latter being transgenic cultivar depleted in salicylic acid). First three fully developed
391 leaves of each plant were rub-inoculated with Carborundum (0.062 mm) (VWR International)
392 using ~75 µl of inoculum per leaf. Inoculum was washed from the inoculated leaves after 10
393 min, non-inoculated leaves were protected from splashing during the washing step. For the
394 first generation of plants, sampling was performed 6, 9, 12, 15, and 21 dpi by taking one leaf
395 tissue disc from first three non-inoculated leaves per each plant. Six plants per cultivar per time
396 point were sampled (some of the plants were sampled twice). The plant material was
397 immediately frozen for subsequent RNA isolation. At 21 dpi additional 3 plants per each cultivar
398 were harvested and their tissue was equally divided in half. First half was frozen and later used
399 for RNA isolation and NGS. Second half was stored at -20 °C and later used to prepare
400 inoculum for infection of the next generation of plants. Each of this 9 plants (3 per cultivar) was
401 used to initiate a lineage in the experiment, plus an additional lineage in which the host cultivar
402 was alternated between D and Dn genotypes, in total amounting 10 lineages.

403 Six serial passages were performed for each lineage (Fig 1). For each new passage (from the
404 2nd passage on), 8 plants per lineage were inoculated following the procedure described above.
405 Virus titer was determined (description in the next section) for the leaf disc samples from 21
406 dpi and this information was used to appropriately dilute the inocula, so that all inocula
407 contained similar concentration of the virus (Fig. 1). All the inocula were sampled and frozen.
408 From the second generation on, leaf tissue discs were sampled at 6 (2 plants per lineage), 9
409 (1 plant per lineage), 12 (1 plant per lineage), 15 (1 plant per lineage), and 21 dpi (2 plants per

lineage). At 21 dpi one plant per lineage was harvested and used for inoculum for next generation of plants and NGS (sampling procedure described above).

Virus titer determination

Virus titer was determined for leaf tissue disc samples sampled on 6, 9, 12, 15, and 21 dpi for 1st generation of plants and on 21 dpi for 2nd – 6th generation.

Total RNA was isolated from plant tissue discs (app. 45 mg) using MagMAX™-96 Total RNA Isolation kits (Life Technologies, USA) and KingFisher™ Purification System (Thermo Scientific) following the manufacturer's instructions for plant samples. Homogenization of the plant material was performed using FastPrep®-24 Instrument (MP Biomedicals, USA).

Absolute quantification of virus titer was performed by RT-qPCR and RT digital droplet PCR (RT-ddPCR) using standard curve approach. The samples were analyzed by two assays: PVY-uni (Kogovšek *et al.*, 2008) and *COX* (Weller *et al.*, 2000). Mix of PVY infected potato samples was used for the standard curve. The absolute number of copies for each dilution point of the standard curve was determined by QX100™ Droplet Digital™ PCR System (BioRad), following the protocol previously described (Gutiérrez-Aguirre *et al.*, 2015), using the same reaction composition and conditions as in respective qPCR reactions. The samples were analyzed by RT-qPCR using AgPath-ID™ one-step RT-PCR kit (Applied Biosystems) on an ABI 7900 HT Fast instrument (Applied Biosystems), following the previously described procedure(Rupar *et al.*, 2015), with manual fluorescence threshold of 0.1 for both assays. Two dilutions (10x and 100x) of isolated RNA, using two technical replicates, were analyzed by both assays to control the dynamics of the replication for each sample. To normalize different input of RNA in reactions, the determined copy number for PVY-uni was normalized by the copy number of calibrator gene: cytochrome oxidase (*COX*), normalizing the number of PVY copies per 1 copy of *COX*. Results are collected in Supplementary Table 3.

Virus population sequencing

Small RNA (sRNA) deep sequencing was employed to determine the population structure of selected samples (Fig 1). Our previous research showed that virus derived sRNAs accurately describe the diversity of PVY in potato (Kutnjak *et al.*, 2015).

Plant material sampled 21 dpi or/and inocula were sequenced for source inoculum, 1st, 4th, 5th, and 6th generation of plants in the experiment (Fig1). For the 1st and 6th generation of plants, RNA was isolated from frozen half of the plant material (other half was used for inoculum preparation) using Trizol reagent (Invitrogen), according to manufacturer's instructions for difficult tissue samples, using 600 µl of Trizol per 100 mg of plant tissue and FastPrep®-24 Instrument (MP Biomedicals, USA) for homogenization. To test if small RNA sequencing can

444 be performed on RNA isolated directly from inocula, we have simultaneously isolated RNA
445 from frozen plant material used for the source inoculum and from source inoculum itself. RNA
446 was isolated from source inoculum using Trizol LS (Invitrogen), following manufacturer
447 protocol for liquid samples. Those two samples were then sequenced and the results showed
448 that high quality small RNA libraries could be obtained also from inoculum itself. The
449 determined virus population structure in both samples was highly similar. Consequently, we
450 have decided to sequence inocula directly for the plants of 4th and 5th passage. Details about
451 each sequenced sample are given in Supplementary Table 1.

452 Due to the narrow genetic bottleneck observed in the 5th and 6th passage of the experiment
453 and consequently a low infection rate we were not able to obtain PVY positive samples for
454 NGS for 6th generation of P1 lineage and 5th and 6th generation of D1 lineage. Plant sample
455 used for sequencing of 6th generation of D2 lineage was sampled at 44 dpi (Fig. 1).

456 Total RNA isolated from either plant material or inocula (explained above) was sent for
457 sequencing to SeqMatic LLC (Fremont, CA, USA), where Illumina small RNA libraries were
458 prepared using TailorMix miRNA Sample Preparation Kit (SeqMatic LLC., Fremont, CA, USA).
459 10-13 samples were pooled in one lane of Illumina HiSeq2000 instrument and sequenced in
460 1x50 bp mode. In each used lane, one of the samples was replicated to be able to later
461 manually determine the frequency cut-off to distinguish between real SNPs and
462 sequencing/PCR/ reverse transcription errors.

463 *Processing of NGS data*

464 Sequencing reads were imported in CLC Genomics Workbench 8 (Qiagen). First, quality
465 control and adaptor trimming was performed, then reads 21-24 nucleotides long were selected
466 (representing small RNAs). Those were exported and mapped to reference genome sequence
467 of PVY (representing consensus genome sequence of PVY population from source inoculum)
468 using BWA 0.7.8 (Li and Durbin, 2009). The mappings were normalized using samtools 1.1 (Li
469 *et al.*, 2009), so that average depth of coverage amounted ~7500x in all but 3 samples, where
470 sequencing depth was lower (see Supplementary Table 1). Subsampling some files with higher
471 coverage to comparably lower coverage (5000x, 2500x), showed that lower coverage in those
472 samples does not affect the sensitivity of further analyses. Mapping files (BAM) were imported
473 back to CLC Genomics Workbench, where SNP calling was performed using Low Frequency
474 Variant Detection tool (significance = 1%, minimum frequency = 0.1%, minimum central quality
475 = 20, minimum neighborhood quality (5 nt) = 15, forward/reverse ratio > 0.001 %). SNP tables
476 were further filtered according to the rules established by comparison of SNP tables of
477 technical replicates. All of the SNPs present at frequency > 0.8% were present in both technical
478 replicates in all of the replicate pairs. Using this value as a cut-off, all of the SNPs which were

present in population at least at 0.8% in one of the time points (inspecting each of the 10 lineages independently) were considered true variants (even for the time points, where their frequency was lower). Secondly, we created a rule to determine true variants in source inoculum: beside all of the SNPs present at frequency > 0.8%, also the SNPs with lower frequencies were considered as true variants if they reached frequency > 0.8% in later time points in at least two independent lineages. All of the variants and their frequencies in all of the samples are shown in Supplementary Table 4. Raw NGS read files (fastq) from this experiment are deposited in NCBI Sequence Read Archive (SRA): accession number SRP083943, or can be found under BioProject number PRJNA340976. The consensus viral genome sequence from the source inoculum is deposited in NCBI GenBank under accession number KX856986.

Nucleotide diversity indices

SNP tables were used as an input for SNPGenie software (Nelson, Moncla and Hughes, 2015), which was used to calculate diversity indices for each population. Consensus sequence of source inoculum sample was used as a reference. Average whole genome nucleotide diversities (for all, synonymous and nonsynonymous SNPs) were calculated separately for all SNPs, SNPs already present in inoculum (standing variation) and newly appeared SNPs (new variation). The results were plotted as line plots (Fig. 3) using R 3.2.3 (R Core Team, 2015).

Wright-Fisher Approximate Bayesian Calculations (WF-ABC)

WF-ABC (Foll, Shim and Jensen, 2015) was used to determine effective population sizes (N_e) of populations and selection coefficients (s) of the SNPs. The program consists of two scripts; wfabc_1 was used to calculate N_e for each lineage. This information was then used as an input for the second script, wfabc_2, which was used to estimate s values associated to observed SNPs.

To test if estimates of N_e from SNPs frequency data somehow reflect the bottlenecks observed in the 5th and 6th passages by RT-qPCR, we have performed wfabc_1 analysis first for all the sequenced time points and then also separately for pairs of time points (inoculum-1st passage; 1st passage-4th passage; 4th passage-5th passage; 5th passage-6th passage). Wfabc_1 was run with default 10000 bootstrap replicates and an average of N_e for all the replicates was calculated.

For the second part of the analysis (selection analysis with wfabc_2) all the time points with observed severe bottlenecks were excluded, thus first three passages were used for lineages P2, D1, D2, D3, Dn1, Dn2, and Dn3, first 4 passages for P1 and all 5 passages were used for lineages P3 and DnxD1.

512 Wfabc_2 was run in haploid mode (-ploidy 1) with 100,000 simulations and 1% acceptance
513 rate, setting uniform prior for *s* between -0.5 and 0.5. Using R 3.2.3, posterior mean, median
514 and 95% highest posterior density interval were calculated for each locus. We considered a
515 variant being non-neutral if the 95% highest posterior density (HPD) interval for *s* excluded
516 zero.

517 The results of the selection analysis were plotted using R 3.2.3. The distributions of posterior
518 mean for *s* values for each lineage were plotted as histograms (Fig. 4, panels on the right).
519 Posterior mean value of estimated *s* for each SNP was plotted on PVY genome coordinates
520 separately for each lineage (Fig. 4, panels on the left). Variants with rejected neutrality (95%
521 highest posterior density interval does not contain 0), variants assigned to different clusters
522 (cluster analysis described in the next section), synonymous/nonsynonymous variants and
523 new/standing variants were encoded in this plot using different symbols and colors.

524 *Analysis of SNP trajectories and clustering*

525 The SNPs trajectories were assigned into groups using R 3.2.3 (heatmap.2 library) hierarchical
526 clustering approach (method = average). The approach used was inspired by analysis of yeast
527 experimental evolution data(Lang *et al.*, 2013). Only the variants reaching 5% in at least one
528 time point were used for this analysis, since the variants with lower frequencies were not
529 informative for cluster analysis. Firstly, SNP trajectories were clustered for each lineage
530 separately. SNP was treated as a vector with *n* dimensions, dimensions representing
531 frequency of the variant in the sequenced time points. Heatmap.2 library was used to cluster
532 SNPs trajectories and visualize them as a heatmap (Supplementary Figure 1).

533 Secondly, results of selection analysis were used to cluster both: SNPs and lineages. Only the
534 SNPs occurring in all lineages were used for this analysis. Each lineage was treated as a
535 vector with *n* dimensions, dimensions representing the selection coefficients of the observed
536 SNPs. *Vice versa*, each SNP was treated as a vector with *n* dimensions, dimensions
537 representing selection coefficient of the SNP in each lineage. R 3.2.3 heatmap.2 library was
538 used to cluster both: lineages and SNPs and visualize the *s* values in a form of a heatmap.
539 Library pvclust (Suzuki and Shimodaira, 2015) was used to asses statistical support of lineages
540 clusters using 1000 bootstrap replicates.

References

Acevedo, A., Brodsky, L. and Andino, R. (2014) 'Mutational and fitness landscapes of an RNA virus revealed through population sequencing.', *Nature*. Nature Publishing Group, 505(7485), pp. 686–90. doi: 10.1038/nature12861.

Alamillo, J. M., Saenz, P. and Garcia, J. A. (2006) 'Salicylic acid-mediated and RNA-silencing defense mechanisms cooperate in the restriction of systemic spread of plum pox virus in tobacco', *Plant Journal*, 48(2), pp. 217–227. doi: 10.1111/j.1365-313X.2006.02861.x.

Andino, R. and Domingo, E. (2015) 'Viral quasispecies', *Virology*. Elsevier, pp. 46–51. doi: 10.1016/j.virol.2015.03.022.

Baebler, Š., Krečič-Stres, H., Rotter, A., Kogovšek, P., Cankar, K., Kok, E. J., Gruden, K., Kovač, M., Žel, J., Pompe-Novak, M. and Ravnikar, M. (2009) 'PVY^{NTN} elicits a diverse gene expression response in different potato genotypes in the first 12 h after inoculation', *Molecular Plant Pathology*, 10(2), pp. 263–275. doi: 10.1111/j.1364-3703.2008.00530.x.

Baebler, Š., Stare, K., Kovač, M., Blejec, A., Prezelj, N., Stare, T., Kogovšek, P., Pompe-Novak, M., Rosahl, S., Ravnikar, M. and Gruden, K. (2011) 'Dynamics of responses in compatible potato - *Potato virus Y* interaction are modulated by salicylic acid.', *PloS one*, 6(12), p. e29009. doi: 10.1371/journal.pone.0029009.

Baebler, Š., Witek, K., Petek, M., Stare, K., Tušek-Žnidarič, M., Pompe-Novak, M., Renaut, J., Szajko, K., Strzelczyk-Żyta, D., Marczewski, W., Morgiewicz, K., Gruden, K. and Hennig, J. (2014) 'Salicylic acid is an indispensable component of the *Ny-1* resistance-gene-mediated response against *Potato virus Y* infection in potato.', *Journal of experimental botany*, 65(4), pp. 1095–109. doi: 10.1093/jxb/ert447.

Bedhomme, S., Lafforgue, G. and Elena, S. F. (2012) 'Multihost experimental evolution of a plant RNA virus reveals local adaptation and host-specific mutations.', *Molecular biology and evolution*, 29(5), pp. 1481–92. doi: 10.1093/molbev/msr314.

Bordería, A. V, Isakov, O., Moratorio, G., Henningsson, R., Agüera-González, S., Organtini, L., Gnädig, N. F., Blanc, H., Alcover, A., Hafenstein, S., Fontes, M., Shomron, N. and Vignuzzi, M. (2015) 'Group selection and contribution of minority variants during virus adaptation determines virus fitness and phenotype.', *PLoS pathogens*, 11(5), p. e1004838. doi: 10.1371/journal.ppat.1004838.

Bull, J. J., Badgett, M. R., Wichman, H. A., Huelsenbeck, J. P., Hillis, D. M., Gulati, A., Ho, C. and Molineux, I. J. (1997) 'Exceptional convergent evolution in a virus', *Genetics*, 147(4), pp. 1497–1507.

Elena, S. F., Fraile, A. and García-Arenal, F. (2014) 'Evolution and emergence of plant viruses', *Advances in Virus Research*, 88, pp. 161–191. doi: 10.1016/B978-0-12-800098-4.00003-9.

Foll, M., Poh, Y. P., Renzette, N., Ferrer-Admetlla, A., Bank, C., Shim, H., Malaspinas, A. S., Ewing, G., Liu, P., Wegmann, D., Caffrey, D. R., Zeldovich, K. B., Bolon, D. N., Wang, J. P., Kowalik, T. F., Schiffer, C. a., Finberg, R. W. and Jensen, J. D. (2014) 'Influenza virus drug resistance: a time-sampled population genetics perspective', *PLoS Genetics*, 10(2), p. e1004185. doi: 10.1371/journal.pgen.1004185.

Foll, M., Shim, H. and Jensen, J. D. (2015) 'WFABC: A Wright-Fisher ABC-based approach for inferring effective population sizes and selection coefficients from time-sampled data', *Molecular Ecology Resources*, 15(1), pp. 87–98. doi: 10.1111/1755-0998.12280.

Froissart, R., Roze, D., Uzest, M., Galibert, L., Blanc, S. and Michalakis, Y. (2005) 'Recombination every day: abundant recombination in a virus during a single multi-cellular host infection.', *PLoS biology*, 3(3), p. e89. doi: 10.1371/journal.pbio.0030089.

588 Di Giallonardo, F., Töpfer, A., Rey, M., Prabhakaran, S., Duport, Y., Leemann, C., Schmutz,
589 S., Campbell, N. K., Joos, B., Lecca, M. R., Patrignani, A., Däumer, M., Beisel, C., Rusert,
590 P., Trkola, A., Günthard, H. F., Roth, V., Beerenwinkel, N. and Metzner, K. J. (2014) 'Full-
591 length haplotype reconstruction to infer the structure of heterogeneous virus populations',
592 *Nucleic Acids Research*, 42(14), pp. 1–12. doi: 10.1093/nar/gku537.

593 Gutiérrez-Aguirre, I., Rački, N., Dreo, T. and Ravnikar, M. (2015) 'Droplet digital PCR for
594 absolute quantification of pathogens', *Plant Pathology: Techniques and Protocols, Methods*
595 *in Molecular Biology*, 1302(6), pp. 331–347. doi: 10.1007/978-1-4939-2620-6.

596 Hillung, J., Cuevas, J. M., Valverde, S. and Elena, S. F. (2014) 'Experimental evolution of an
597 emerging plant virus in host genotypes that differ in their susceptibility to infection.',
598 *Evolution; international journal of organic evolution*, 68(9), pp. 2467–80. doi:
599 10.1111/evo.12458.

600 Karasev, A. V and Gray, S. M. (2013) 'Continuous and emerging challenges of *Potato virus Y*
601 in potato.', *Annual review of phytopathology*, 51, pp. 571–86. doi: 10.1146/annurev-phyto-
602 082712-102332.

603 Kessinger, T. A., Perelson, A. S. and Neher, R. A. (2013) 'Inferring HIV escape rates from
604 multi-locus genotype data', *Frontiers in Immunology*, 4(252), pp. 1–13. doi:
605 10.3389/fimmu.2013.00252.

606 Kogovšek, P., Gow, L., Pompe-Novak, M., Gruden, K., Foster, G. D., Boonham, N. and
607 Ravnikar, M. (2008) 'Single-step RT real-time PCR for sensitive detection and discrimination
608 of Potato virus Y isolates.', *Journal of virological methods*, 149(1), pp. 1–11. doi:
609 10.1016/j.jviromet.2008.01.025.

610 Kogovšek, P., Pompe-Novak, M., Baebler, Š., Rotter, a., Gow, L., Gruden, K., Foster, G. D.,
611 Boonham, N. and Ravnikar, M. (2010) 'Aggressive and mild *Potato virus Y* isolates trigger
612 different specific responses in susceptible potato plants', *Plant Pathology*, 59(6), pp. 1121–
613 1132. doi: 10.1111/j.1365-3059.2010.02340.x.

614 Kogovšek, P., Pompe-Novak, M., Petek, M., Fragner, L., Weckwerth, W. and Gruden, K.
615 (2016) 'Primary metabolism, phenylpropanoids and antioxidant pathways are regulated in
616 potato as a response to *Potato virus Y* infection', *PLoS ONE*, 11(1), pp. 1–20. doi:
617 10.1371/journal.pone.0146135.

618 Kutnjak, D., Rupar, M., Gutierrez-Aguirre, I., Curk, T., Kreuze, J. F. and Ravnikar, M. (2015)
619 'Deep sequencing of virus-derived small interfering RNAs and RNA from viral particles shows
620 highly similar mutational landscapes of a plant virus population', *Journal of Virology*, 89(9),
621 pp. 4760–4769. doi: 10.1128/JVI.03685-14.

622 Lang, G. I., Rice, D. P., Hickman, M. J., Sodergren, E., Weinstock, G. M., Botstein, D. and
623 Desai, M. M. (2013) 'Pervasive genetic hitchhiking and clonal interference in forty evolving
624 yeast populations.', *Nature*. Nature Publishing Group, 500(7464), pp. 571–4. doi:
625 10.1038/nature12344.

626 Li, H. and Durbin, R. (2009) 'Fast and accurate short read alignment with Burrows-Wheeler
627 transform', *Bioinformatics*, 25(14), pp. 1754–1760. doi: 10.1093/bioinformatics/btp324.

628 Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G.
629 and Durbin, R. (2009) 'The sequence alignment/map format and SAMtools.', *Bioinformatics*
630 *(Oxford, England)*, 25(16), pp. 2078–9. doi: 10.1093/bioinformatics/btp352.

631 Longdon, B., Brockhurst, M. A., Russell, C. A., Welch, J. J. and Jiggins, F. M. (2014) 'The
632 evolution and genetics of virus host shifts', *PLoS Pathogens*. Edited by T. C. Hobman,
633 10(11), p. e1004395. doi: 10.1371/journal.ppat.1004395.

634 Mestre, P., Brigneti, G., Durrant, M. C. and Baulcombe, D. C. (2003) 'Potato virus Y NIa

635 protease activity is not sufficient for elicitation of *Ry*-mediated disease resistance in potato.',
636 *The Plant journal : for cell and molecular biology*, 36(6), pp. 755–761. doi:
637 doi:10.1046/j.1365-313X.2003.01917.x.

638 Miyashita, S., Ishibashi, K., Kishino, H. and Ishikawa, M. (2015) 'Viruses roll the dice: the
639 stochastic behavior of viral genome molecules accelerates viral adaptation at the cell and
640 tissue levels', *PLoS Biology*, 13(3), p. e1002094. doi: 10.1371/journal.pbio.1002094.

641 Neher, R. A. (2013) 'Genetic draft, selective interference, and population genetics of rapid
642 adaptation', *Annu. Rev. Ecol. Evol. Syst*, 44, pp. 195–215. doi: 10.1146/annurev-ecolsys-
643 110512-135920.

644 Nelson, C. W., Moncla, L. H. and Hughes, A. L. (2015) 'SNPGenie: Estimating evolutionary
645 parameters to detect natural selection using pooled next-generation sequencing data',
646 *Bioinformatics*, 31(22), pp. 3709–3711. doi: 10.1093/bioinformatics/btv449.

647 Phan, J., Zdanov, A., Evdokimov, A. G., Tropea, J. E., Peters, H. K., Kapust, R. B., Li, M.,
648 Wlodawer, A. and Waugh, D. S. (2002) 'Structural basis for the substrate specificity of
649 tobacco etch virus protease', *Journal of Biological Chemistry*, 277(52), pp. 50564–50572.
650 doi: 10.1074/jbc.M207224200.

651 Prabhakaran, S., Rey, M., Zagordi, O., Beerenwinkel, N. and Roth, V. (2014) 'HIV haplotype
652 inference using a propagating dirichlet process mixture model', *IEEE/ACM Transactions on
653 Computational Biology and Bioinformatics*, 11(1), pp. 182–191. doi:
654 10.1109/TCBB.2013.145.

655 Prosperi, M. C. F. and Salemi, M. (2012) 'QuRe: software for viral quasispecies
656 reconstruction from next-generation sequencing data.', *Bioinformatics (Oxford, England)*,
657 28(1), pp. 132–3. doi: 10.1093/bioinformatics/btr627.

658 Pruss, G. J., Lawrence, C. B., Bass, T., Li, Q. Q., Bowman, L. H. and Vance, V. (2004) 'The
659 potyviral suppressor of RNA silencing confers enhanced resistance to multiple pathogens',
660 *Virology*, 320(1), pp. 107–120. doi: 10.1016/j.virol.2003.11.027.

661 R Core Team (2015) 'R: A language and environment for statistical computing.' Vienna,
662 Austria: R Foundation for Statistical Computing. Available at: https://www.r-project.org/.

663 Remold, S. K., Rambaut, A. and Turner, P. E. (2008) 'Evolutionary genomics of host
664 adaptation in vesicular stomatitis virus', *Molecular Biology and Evolution*, 25(6), pp. 1138–
665 1147. doi: 10.1093/molbev/msn059.

666 Ruiz-Jarabo, C. M., Arias, A., Baranowski, E., Escarmís, C. and Domingo, E. (2000) 'Memory
667 in viral quasispecies.', *Journal of virology*, 74(8), pp. 3543–3547. doi: 10.1128/JVI.74.8.3543-
668 3547.2000.

669 Rupar, M., Faurez, F., Tribodet, M., Gutierrez-Aguirre, I., Delaunay, A., Glais, L., Kriznik, M.,
670 Dobnik, D., Gruden, K., Jacquot, E. and Ravnikar, M. (2015) 'Fluorescently tagged *Potato
671 virus Y*: a versatile tool for functional analysis of plant-virus interactions.', *Molecular plant-
672 microbe interactions: MPMI*, 28(7), pp. 739–750. doi: 10.1094/MPMI-07-14-0218-TA.

673 Schirmer, M., Sloan, W. T. and Quince, C. (2014) 'Benchmarking of viral haplotype
674 reconstruction programmes: an overview of the capacities and limitations of currently
675 available programmes', *Briefings in Bioinformatics*, 15(3), pp. 431–442. doi:
676 10.1093/bib/bbs081.

677 Scholthof, K.-B. G., Adkins, S., Czosnek, H., Palukaitis, P., Jacquot, E., Hohn, T., Hohn, B.,
678 Saunders, K., Candresse, T., Ahlquist, P., Hemenway, C. and Foster, G. D. (2011) 'Top 10
679 plant viruses in molecular plant pathology.', *Molecular plant pathology*, 12(9), pp. 938–54.
680 doi: 10.1111/j.1364-3703.2011.00752.x.

681 Simon-Loriere, E. and Holmes, E. C. (2011) 'Why do RNA viruses recombine?', *Nature*
682 *reviews. Microbiology*. Nature Publishing Group, 9(8), pp. 617–626. doi:
683 10.1038/nrmicro2614.

684 Singh, D. P., Moore, C. A., Gilllliland, A. and Carr, J. P. (2004) 'Activation of multiple antiviral
685 defence mechanisms by salicylic acid', *Molecular Plant Pathology*, 5, pp. 57–63. doi:
686 10.1046/J.1364-3703.2004.00203.X.

687 Stapleford, K. a, Coffey, L. L., Lay, S., Bordería, A. V, Duong, V., Isakov, O., Rozen-Gagnon,
688 K., Arias-Goeta, C., Blanc, H., Beaucourt, S., Haliloğlu, T., Schmitt, C., Bonne, I., Ben-Tal,
689 N., Shomron, N., Failloux, A.-B., Buchy, P. and Vignuzzi, M. (2014) 'Emergence and
690 transmission of arbovirus evolutionary intermediates with epidemic potential.', *Cell host &*
691 *microbe*, 15(6), pp. 706–16. doi: 10.1016/j.chom.2014.05.008.

692 Stare, T., Ramšak, Ž., Blejec, A., Stare, K., Turnšek, N., Weckwerth, W., Wienkoop, S.,
693 Vodnik, D. and Gruden, K. (2015) 'Bimodal dynamics of primary metabolism-related
694 responses in tolerant potato-*Potato virus Y* interaction.', *BMC genomics*, 16(1), p. 716. doi:
695 10.1186/s12864-015-1925-2.

696 Suzuki, R. and Shimodaira, H. (2015) 'pvclust: hierarchical clustering with p-values via
697 multiscale bootstrap resampling. R package version 2.0-0'. Available at: https://cran.r-
698 project.org/package=pvclust.

699 Töpfer, A., Marschall, T., Bull, R. A., Luciani, F., Schönhuth, A. and Beerenwinkel, N. (2014)
700 'Viral quasispecies assembly via maximal clique enumeration', *Lecture Notes in Computer*
701 *Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in*
702 *Bioinformatics)*, e1003515(3), pp. 1–10. doi: 10.1007/978-3-319-05269-4_25.

703 Tromas, N., Zwart, M. P., Poulain, M. and Elena, S. F. (2014) 'Estimation of the in vivo
704 recombination rate for a plant RNA virus', *Journal of General Virology*, 95(3), pp. 724–732.
705 doi: 10.1099/vir.0.060822-0.

706 Turner, P. E. and Elena, S. F. (2000) 'Cost of host radiation in an RNA virus', *Genetics*,
707 156(4), pp. 1465–1470.

708 Vassilakos, N., Simon, V., Tzima, A., Johansen, E. and Moury, B. (2016) 'Genetic
709 determinism and evolutionary reconstruction of a host jump in a plant virus', *Molecular*
710 *Biology and Evolution*, 33(2), pp. 541–553. doi: 10.1093/molbev/msv222.

711 Weller, S. A., Elphinstone, J. G., Smith, N. C., Boonham, N. and Stead, D. E. (2000)
712 'Detection of Ralstonia solanacearum strains with a quantitative, multiplex, realtime,
713 fluorogenic PCR (TaqMan) assay', *Applied and Environmental Microbiology*, 66(7), pp.
714 2853–2858. doi: 10.1128/AEM.66.7.2853-2858.2000.Updated.

715 Yu, D., Fan, B., MacFarlane, S. a and Chen, Z. (2003) 'Analysis of the involvement of an
716 inducible *Arabidopsis* RNA-dependent RNA polymerase in antiviral defense.', *Molecular*
717 *plant-microbe interactions : MPMI*, 16(3), pp. 206–216. doi: 10.1094/MPMI.2003.16.3.206.

718 Zagordi, O., Bhattacharya, A., Eriksson, N. and Beerenwinkel, N. (2011) 'ShoRAH:
719 estimating the genetic diversity of a mixed sample from next-generation sequencing data.',
720 *BMC bioinformatics*, 12(1), p. 119. doi: 10.1186/1471-2105-12-119.

721 Zwart, M. P., Daròs, J.-A. and Elena, S. F. (2012) 'Effects of potyvirus effective population
722 size in inoculated leaves on viral accumulation and the onset of symptoms.', *Journal of*
723 *Virology*, 86(18), pp. 9737–47. doi: 10.1128/JVI.00909-12.

724 Zwart, M. P. and Elena, S. F. (2015) 'Matters of size: genetic bottlenecks in virus infection
725 and their potential impact on evolution', *Annual Review of Virology*, 2(1), pp. 161–179. doi:
726 10.1146/annurev-virology-100114-055135.

727 **Acknowledgements**

728 This study was supported by Slovenian Research Agency through grant L4-5525 and PhD
729 grant to D.K. Work in Valencia was supported by grant BFU2015-65037-P from Spain Ministry
730 of Economy and Competitiveness to S.F.E. and short term scientific mission support to D.K. in
731 the frame of EU funded COST action FA1407. We would like to thank Ion Gutierrez-Aguirre
732 for critical reading of the manuscript, Lidija Matičič for the propagation of potato plantlets and
733 Maja Križnik, Špela Alič, Zala Kogej and Larisa Gregur for help with the inoculation. We also
734 thank Dr. Sabine Rosahl for providing NahG-Désirée potato plants.

735 **Competing interests**

736 Authors declare that no competing interests exist.

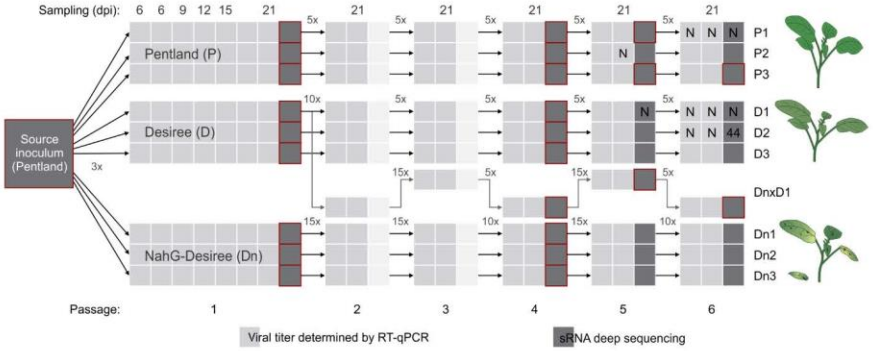
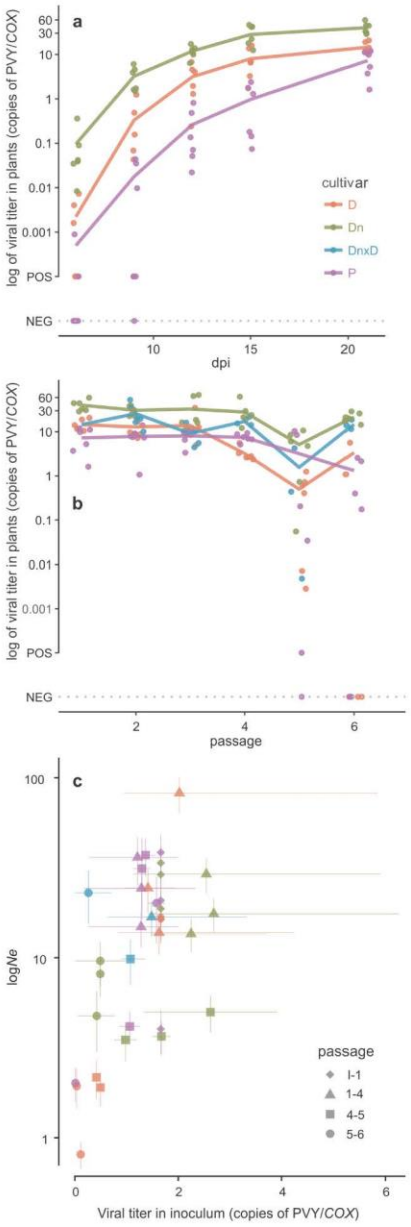


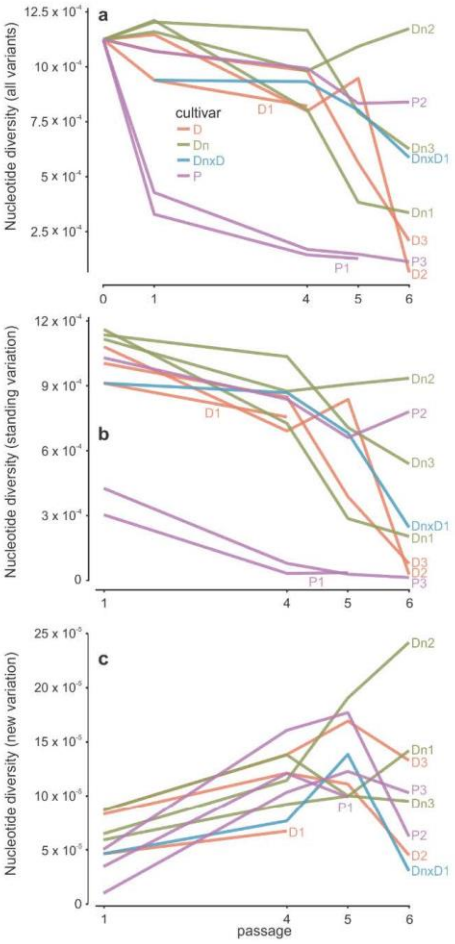
Figure 1: *Setup of the evolution experiment*. Three lineages of three different potato genotypes (Pentland squire, Désirée, NahG- Désirée) were inoculated with the same PVY inoculum (derived from the pool of systemically infected cv. Pentland squire plants: big square). 6 serial passages of the virus were performed for each lineage. For one of the lineages, host plant genotype was alternated between Désirée and NahG-Désirée (lineage DnxD). Small squares represent individual plants, which were sampled either for viral titer determination by RT-qPCR (light grey) or population sequencing (Illumina sRNA deep sequencing, dark grey). Above the scheme sampling time points are designated (dpi – days post inoculation). Coefficients next to the arrows designate dilution coefficient used to approximately normalize virus titer in inocula in each passage. Samples negative for virus are designated with N. In the 6th passage, one of the samples was sampled later (44 dpi) than others, which is designated on the scheme (44). Red outlines designate the samples included into the selection analysis with WF-ABC (some of the samples were excluded due to the observation of very narrow genetic bottlenecks).



751 Figure 2: *Viral titers and effective population*
752 *sizes*. Viral titers measured at 6, 9, 12, 15,
753 and 21 dpi for the 1st passage are shown in
754 (a). Viral titers at 21 dpi in each of the 6
755 passages are shown in (b). Samples
756 positive for PVY, but below the quantitative
757 range are marked as POS, negative
758 samples are marked as NEG. Lines connect
759 average values, which are calculated only
760 for positive samples. In (c) viral titers in the
761 inocula (viral titer in plants at 21 dpi
762 multiplied by dilution coefficient) are plotted
763 against log transformed estimated effective
764 population sizes (N_e) from WF-ABC
765 analysis. Points represent average values,
766 horizontal lines connect minimum and
767 maximum measured value for viral titer for
768 specific lineage/passage, whereas vertical
769 lines represent standard deviation of
770 estimated N_e. For point 1-4, average was
771 calculated for viral titers in passages 1-4.
772 Different individuals were used for
773 estimation of viral titer on one hand and
774 population sequencing on the other (Fig. 1).
775 In all three panels the host genotypes are
776 color coded. In panel (c) different symbols
777 encode different passages.

750

778



779

780 Figure 3: *Virus populations diversity dynamics*. Average population nucleotide diversity (π) was
781 calculated for (a) all variants, (b) variants presents already in the source inoculum (standing variation)
782 and (c) new spontaneous mutations (new variation). Lines are connecting calculated π values through
783 time points for each lineage; host genotypes are color coded in all three panels.

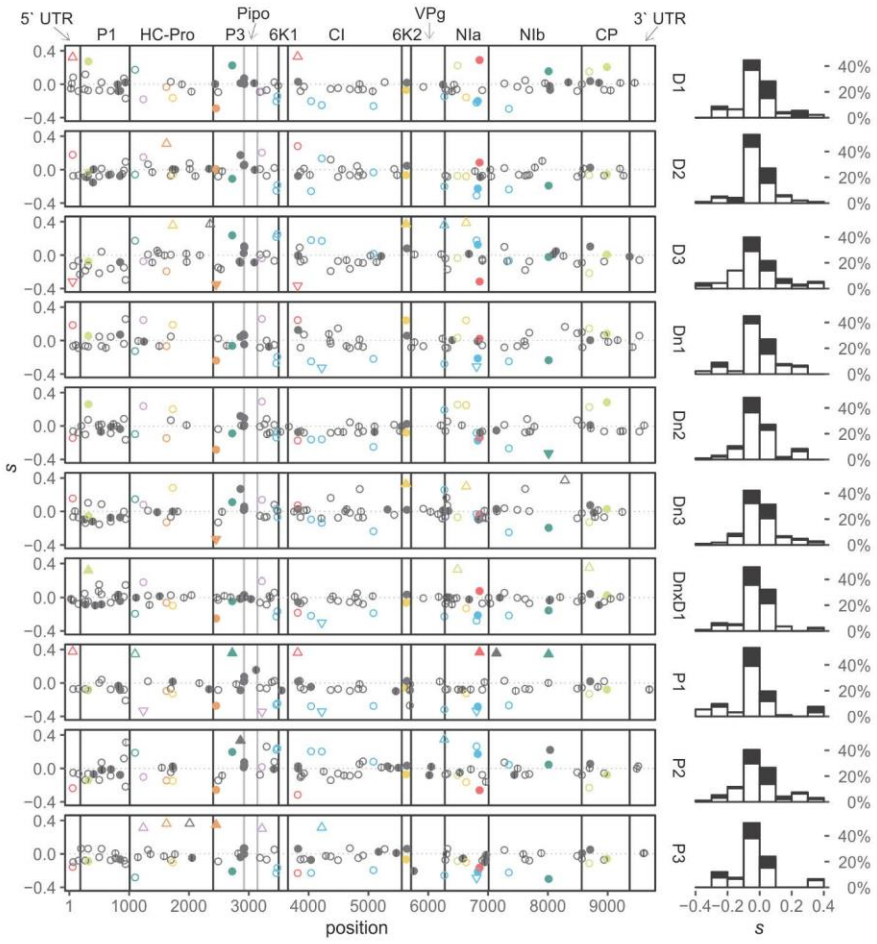


Figure 4: *Distribution of SNPs on PVY genome and their selection coefficients (s) in different lineages.* *s* values were estimated using WF-ABC approach, excluding the time points with observed severe bottlenecks (Fig 1). On the left, posterior means of estimated selection coefficients for SNPs in different lineages are plotted across the PVY genome. Empty signs designate synonymous SNPs, filled signs nonsynonymous SNPs; triangles represent SNPs for which neutrality was rejected (95% HPD interval does not include 0); ticks within the signs mark the new spontaneous mutations (not present in the source inoculum). Colors correspond to 7 groups of SNPs identified by clustering analysis (Fig. 5), grey color designates SNPs, which were not assigned to any group. On the right side, distribution of posterior means of estimated *s* of SNPs for each lineage are shown; white surfaces represent synonymous mutations and black surfaces nonsynonymous mutations.

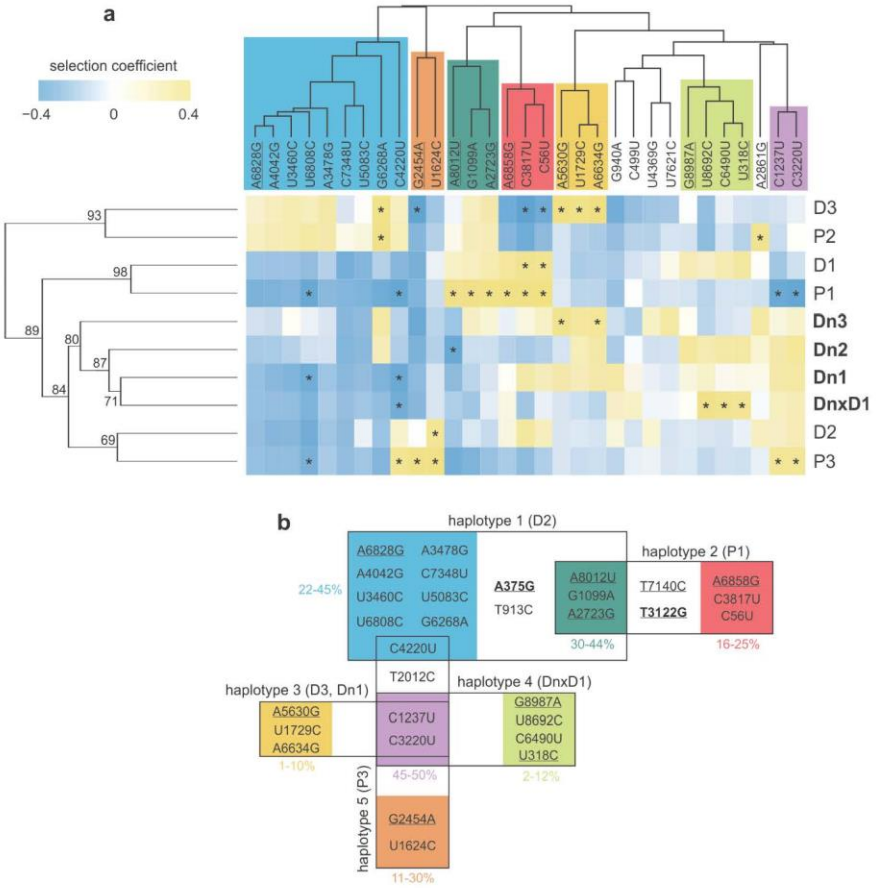


Figure 5: *Clustering analysis and haplotypes reconstruction*. SNPs were clustered into groups based on their inferred selection coefficients in different lineages (a: columns). 7 recognized groups are color-coded. Lineages were clustered based on selection coefficient of the SNPs (a: lines), the ones evolved in potato genotypes depleted in salicylic acid are bolded. Only the SNPs occurring in all lineages were included in this analysis. Selection coefficient of SNPs in every lineage are shown as blue-yellow heat map. Asterisks designate SNPs for which neutrality was rejected (in respective lineage). The numbers next to the branches of the lineages dendrogram represent statistical support of the groups, determined by bootstrapping. Several haplotypes were reconstructed by clustering SNPs frequency trajectories and observing fixed haplotypes. Diagram (b) represent reconstructed haplotypes in source inoculum (lineages in which they were observed are given in parentheses, see Supplementary Figure 1), boxes are delimiting theoretical haplotypes, colors are corresponding to the colors of the groups from clustering analysis shown in (a). New spontaneous mutations (not present in the source inoculum) are bolded, nonsynonymous mutations are underlined. Frequencies shown next to the groups are frequencies observed in the source inoculum.

Chapter 3

Discussion

3.1 Viral Sequence Enrichment Approaches

With the growing interest in virus metagenomics as well as research of virus diversity on other levels, there is an increased need for more efficient sample preparation approaches, which would reduce the time needed for preparation of samples and increase the purity of obtained viral nucleic acids before NGS. Therefore, we have first focused on the sample preparation steps. CIM monolithic chromatography was shown to be a time-efficient alternative to classical virus particles purification techniques [22], [23]. We have coupled it, for the first time, with NGS and confirmed its utility as a virus enrichment technique. In the first publication, CIM monolithic chromatography was successfully employed to purify a putative orthoreovirus from a stool sample and cell-culture supernatant. We compared three alternative purification approaches: (i) stool sample treated with benzonase, viruses concentrated with CIM one-step elution; (ii) cell culture supernatant treated with benzonase, viruses concentrated with CIM one-step elution and (iii) cell culture supernatant without benzonase treatment, viruses purified with CIM gradient elution. We were able to get the best results using the third approach, which yielded the lowest concentration of total RNA, however, this was very pure viral RNA, with only a minute amount of host background sequences. Using reads from this approach for de novo genome assembly we were able to obtain the longest viral contigs, covering nearly the whole viral genome sequence. These results demonstrated not only the efficiency of CIM monolithic chromatography as a viral enrichment approach (before NGS), but also showed that purity of viral nucleic acids is an important factor when trying to assemble new viral genomes from NGS data. We showed that high quality sequencing data can be obtained from very small amounts of viral RNA if it is sufficiently pure. Finally, this sequencing data allowed us to reconstruct a complete viral genome of a putative Orthoreovirus and phylogenetic analysis showed that this is a *Mammalian orthoreovirus*, highly similar to the viruses found in bats in Germany and Italy. This was the first study describing MRV from bat cluster in human, indicating a possible zoonotic transmission, which is the topic of further investigations [45]–[48].

In the second part of the research we have employed small RNA deep sequencing, which is currently one of the most widely adopted NGS techniques for discovery of viruses in plants and invertebrates. Using this method we characterized eight isolates of PVX from South America. The study resulted in a detection of a new, phylogenetically divergent strain of PVX and discovery of PVX isolates from the Eurasian group in this region. Both confirmed the hypothesis that PVX is undersampled in the Andean region of South America and thus, its diversity is underestimated. Moreover, we detected two strains of PVX in one sample, which was the first use of sRNA sequencing for such an application.

This was likely possible because the two strains were divergent enough; for closely related virus strains the discrimination using short sRNA reads would probably not be possible. Sequence data obtained in this study allowed us to design specific PCR primer pairs, which enable quick discrimination between different PVX strains. This demonstrates how accumulation of genome data obtained by NGS technologies can help us to design new or better molecular tools for further investigation of viruses.

3.2 Comparison of Virus Derived sRNAs and RNA from Purified Virus Particles

In the first two publications we demonstrated the use of two different virus enrichment techniques coupled with NGS. We have then decided to compare these two approaches. First we compared them in terms of efficiency of host background removal and, connected to this, in terms of efficiency of de novo assembly of viral consensus genome sequence. Using PVY-potato as a model system, we showed that both approaches result in a strong reduction of the amount of host sequences, though this was more dramatic in the case of sequencing RNA from purified viral particles, where host sequences were almost completely removed (they represented only 2% of total reads in this pool, in comparison, they comprised app. 50% of sRNA libraries). This, probably in combination with slightly longer reads, enabled more efficient de novo reconstruction of consensus viral genome sequence, which was assembled in one step using VP data. Also, reads from the VP pool enabled detection of recombination events, for which sRNA reads were too short. On the other hand, sRNAs showed more uniform distribution across the viral genome; they still enabled reliable detection and also reconstruction of complete viral consensus genomes, however this was performed in several steps.

Secondly, we compared the structure of virus populations reflected in each of the two viral sequence pools. The results showed very similar diversity in both pools, *i.e.*, the same variants were found in both populations at similar frequencies. Some differences were, nevertheless, observed, especially a higher number of low frequency variants in sRNA pool. There are several possible explanations for this observation, discussed in detail in publication 2.3, including the sampling procedure and possible new mutations arising through generation of secondary small RNAs. The comparison presented in publication 2.3 is the first thorough comparison of two different viral sequence pools from the same plants; it will serve as an important baseline for further studies of within-host virus population studies.

Our comparison showed that the choice of the method used for enrichment of viral sequences depends on the type of the study; if one wants to use NGS to obtain a complete genome sequence of an unknown, possibly very divergent virus, purification of virus particles before NGS is a good option, since longer reads and a lower amount of host background sequences would enhance the reconstruction of a viral genome sequence. On the other hand, sRNA deep sequencing represents a more generic solution (protocol does not differ between different plant and viral species), which can be easily routinized and is as such a good option when investigating a larger number of different samples. Other approaches, not directly compared in this thesis, also represent important alternatives, for example, sequencing total RNA with depleted ribosomal RNA [17].

3.3 Experimental Evolution of PVY

Building on the conclusions of the previous publication, we have used sRNA deep sequencing to track the dynamics of PVY populations within different potato genotypes during an experimental evolution. Using a time-sampled approach (defining viral populations' structures in several passages during the experiment) and starting from a highly diverse viral population, we were able to follow the frequency trajectories of variants, which were present in the population from the beginning (standing variation) as well as the ones, which emerged anew during the experiment.

In the first part of the experiment, we observed, in most of the lineages, a rise in diversity of new variants, reflecting new spontaneous mutations emerging during the replication of the virus. On the other hand we observed a drop in diversity of standing variation, likely a consequence of a natural selection. Exception to this pattern was observed for the lineage in which the host genotype was alternated. Here, standing variation slightly increased, likely reflecting conflicting selection pressures in different genotypes, which prevented fixation of specific viral haplotype. This interesting observation suggests that viral populations in a changing environment can sustain higher diversity in comparison to viral populations in a more stable environment. Since the diversity of viral population represents a reservoir of genetic variants important for viral adaptation potential, these findings can have an important implication for understanding of the evolution of specialist (stable host/environment) and generalist (changing host/environment) viruses.

Next, we have unexpectedly observed strongly reduced viral titers in all of the potato genotypes in one of the late passages. These corresponded to a narrow genetic bottleneck reflected in the population sequencing data. This event enabled us to explore in more detail the dynamics of the viral populations subjected to a strong effect of a genetic drift. Such events are common in the nature, accompanying most of the transmission events. The population sequencing data showed disturbed patterns of variation after this event, with a severely reduced diversity of viral populations in most of the lineages. At the same time, a shift in selected variants was observed, indicating a role of genetic drift in redirecting the trajectories of evolution to another peak in a fitness landscape of the virus. Moreover, observation of a strong selection and many new synonymous mutations in some lineages, after the bottleneck, suggested an important role of selection after the population diversity was decreased.

Time-sampled population sequencing approach enabled to calculate estimates of a selection and genetic drift, using population genetics approaches, and, at the same time, to cluster variants' trajectories in groups, which ultimately led to reconstruction of ancestral viral haplotypes in source inoculum. We demonstrated a linkage of groups of variants to the same haplotype and showed that the genetic draft (linked selection) often caused a fixation of a group of variants (linked to the same haplotype) after the emergence of a new, "driver" mutation. A clustering approach, demonstrated in this study, could be further developed and used for the reconstruction of viral haplotypes, especially in combination with the currently available haplotype reconstruction algorithms, which do not take temporal information into account.

When comparing the selection patterns in PVY lineages evolving in different potato genotypes, we noticed slight differences. Most notably, selection pressures appeared to be more uniform in the potato genotype deleted in salicylic acid than in two other potato cultivars (efficient in accumulation of salicylic acid), indicating diversifying selection in the presence of salicylic acid. These could translate to a greater diversity of the virus, when infecting less permissive hosts and consequently a greater potential for adaptation to new environments (*e.g.*, new host species).

Chapter 4

Conclusions

This thesis presents new processing solutions which will impact NGS-aided discovery and research of viruses, and provides important insights into the diversity and evolution of viruses on the level of species, strains and within-host populations. First, CIM monolithic chromatography was successfully employed to accelerate the purification of viral particles before the NGS. Such an approach provides advantage at de novo assembly of viral genomes from NGS reads. It is especially suited when highly pure viral nucleic acids are needed and could be used for most diverse types of samples (*e.g.*, also for samples of water or soil). We demonstrated its utility on cases of human and plant viruses, discovering a novel enteric virus in human, with a possible indication of zoonotic transmission.

An alternative virus sequence enrichment approach, small RNA deep sequencing, was used to characterize a genome sequence of a novel strain of potato virus X, broadening the knowledge about the diversity of this virus in South America. We also demonstrated the utility of this method to distinguish between two strains of the virus in a mixed infection.

By comparing two viral sequence enrichment approaches, sRNA deep sequencing and CIM purification of VP, it was revealed that both enriched viral sequence pools (sRNAs and RNA from VP) show very similar diversity of PVY populations within host plants. This provided a baseline to use sRNA deep sequencing to describe the course of the experimental evolution of PVY within differently permissive hosts, revealing an interesting pattern of the interplay between selection and genetic drift during the virus evolution and possibly diversifying selection acting on the virus in less permissive host genotypes. New insights into evolution patterns detected on the level of virus populations will help us to better understand the mechanism of viral evolution and emergence in wild and anthropogenic ecosystems.

Appendix A

Supplementary Material of Included Publications

A.1 Supplementary Material for Publication 2.1

Table S1: Nucleotide (Nu) and amino acid (AA) identity between the Slovenian SI-MRV01 orthoreovirus strain and the most related strains obtained from GenBank.

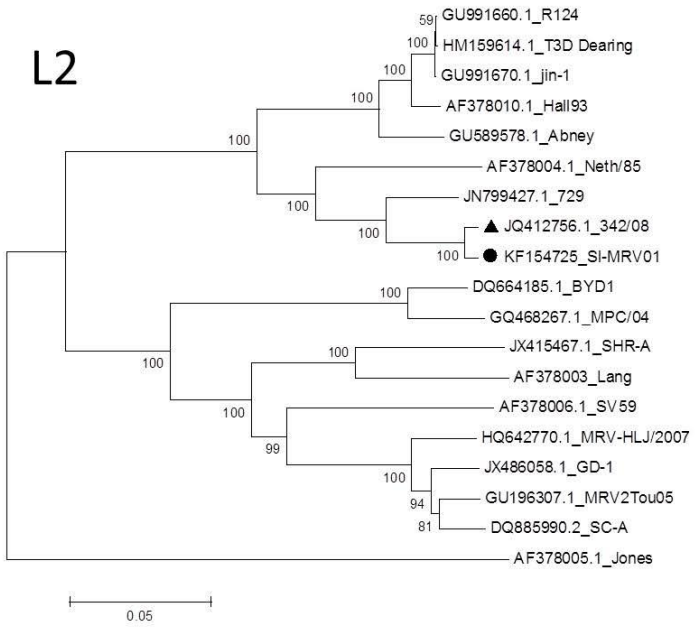
	SI-MRV01																					
	L1		L2		L3		M1		M2		M3		S1		S2		S3		S4			
	Nu	*AA*	*Nu*	*AA*	*Nu*	*AA*	*Nu*	*AA*	*Nu*	*AA*	*Nu*	*AA*	*Nu*	*AA*	*Nu*	*AA*	*Nu*	*AA*	*Nu*	*AA*		
342/08	**97.9**	**99.2**	**99.0**	**99.6**	**98.4**	**99.7**	**98.6**	**99.0**	**98.7**	**99.4**	**98.7**	**99.7**	**98.2**	**98.4**	93.8	**98.5**	**98.2**	**98.9**	**98.8**	**99.7**		
MRV-HUJ/2007	88.6	99.1	76.1	93.2	83.6	98.6	86.9	96.3	89.6	98.5	94.4	97.0	80.9	89.8	**94.5**	98.0	89.7	98.8	90.9	97.4		
MPC/04	88.8	98.5	76.9	92.6	83.4	97.9	87.0	95.6	93.5	99.4	95.0	98.0	81.5	90.7	94.4	98.0	84.7	97.4	90.8	97.4		
729	91.1	98.6	94.4	98.6	83.9	98.5	86.4	95.5	84.1	98.3	90.7	97.0	40.9	23.9	84.8	97.6	89.6	98.8	91.5	98.3		
SC-A	90.6	98.5	75.9	92.7	83.7	97.9	86.7	95.7	93.0	99.1	90.4	97.5	80.7	88.7	93.9	98.0	89.3	98.2	91.8	98.8		
GD-1	90.3	98.3	76.2	93.0	83.6	98.1	85.6	95.1	92.7	99.1	83.5	94.7	81.1	90.7	87.2	97.1	92.0	98.5	76.2	87.4		
R124	89.3	98.7	87.2	97.6	87.2	98.3	87.6	96.4	90.9	98.5	85.0	96.1	80.0	88.1	84.0	98.0	85.1	98.0	87.0	96.3		
T3D	89.3	98.7	87.2	97.6	87.3	98.5	87.6	96.1	90.8	98.5	85.0	95.9	80.0	88.3	83.9	98.0	85.1	98.0	86.9	96.3		
BYD1	89.3	98.5	77.1	92.4	87.4	98.0	87.3	96.1	89.6	98.7	84.5	94.7	42.2	26.1	85.0	96.4	83.9	96.8	83.9	96.6		
SHR-A	90.2	99.0	75.8	91.4	86.9	98.7	88.0	96.7	90.9	98.7	85.1	96.2	40.9	21.7	86.7	97.8	87.2	98.5	76.1	87.9		
Jin-1	89.3	98.8	87.2	97.6	87.3	98.6	87.6	96.4	90.8	98.5	85.0	96.2	79.9	87.6	83.9	97.8	85.1	98.0	86.9	96.0		
Ndelle	89.4	98.1	/	/	/	/	/	/	89.3	98.4	/	/	71.5	67.6	84.0	96.1	/	/	89.8	95.8		
Abney	88.9	98.5	87.1	97.3	87.4	98.8	86.8	96.0	91.0	98.7	85.0	96.2	79.6	87.9	86.1	97.3	84.5	97.4	87.1	97.4		
Lang	89.1	98.3	75.3	92.3	87.3	98.6	87.9	97.1	84.5	98.3	84.3	96.1	41.3	21.7	86.2	97.8	93.9	98.2	86.2	96.3		
Jones	75.2	92.1	72.7	86.3	76.6	95.8	72.3	81.5	76.9	97.4	70.5	83.4	40.7	20.0	76.4	93.0	73.3	86.8	77.1	92.7		

/ - not available

Table S2: Details of mapping parameters used in the CLC Genomics Workshop program.

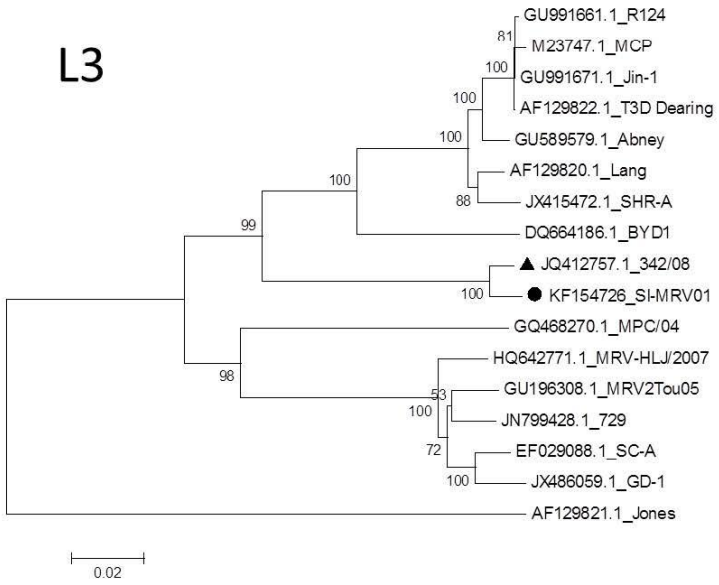
	mapping to new Orthoreovirus reference sequence	**mapping to *Macaca mulatta* genome**	**de novo assembly (map reads back to contigs)**
Missmatch cost	1	1	2
Insertion cost	2	2	3
Deletion cost	2	2	3
Length fraction	0.5	0.9	0.5
Similarity fraction	0.8	0.99	0.8
Non-specific match handling	map randomly	map randomly	-

Fig. S1: Phylogenetic analysis of mammalian orthoreoviruses for nucleotide sequence of L2 genome segment.



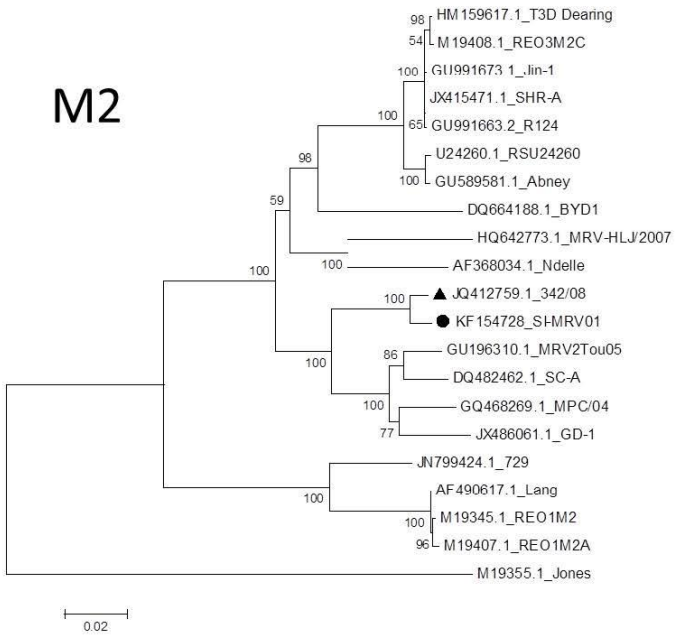
● Slovenian MRV isolate SI-MRV01; ▲ German MRV isolate 342/08 identified in bat

Fig. S2: Phylogenetic analysis of mammalian orthoreoviruses for nucleotide sequence of L3 genome segment.



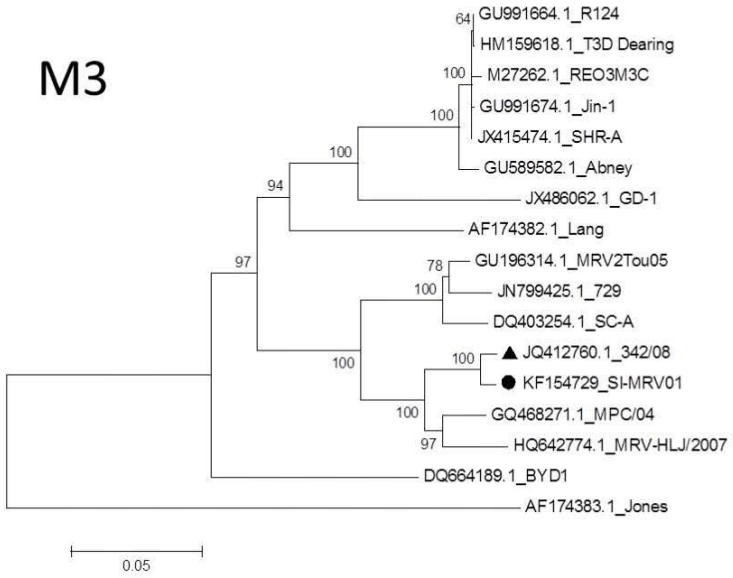
● Slovenian MRV isolate SI-MRV01; ▲ German MRV isolate 342/08 identified in bat

Fig. S3: Phylogenetic analysis of mammalian orthoreoviruses for nucleotide sequence of M2 genome segment.



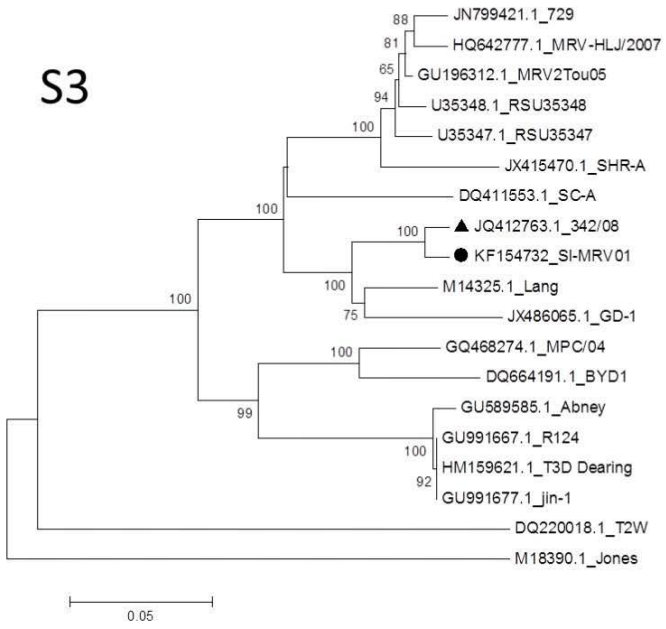
● Slovenian MRV isolate SI-MRV01; ▲ German MRV isolate 342/08 identified in bat

Fig. S4: Phylogenetic analysis of mammalian orthoreoviruses for nucleotide sequence of M3 genome segment.



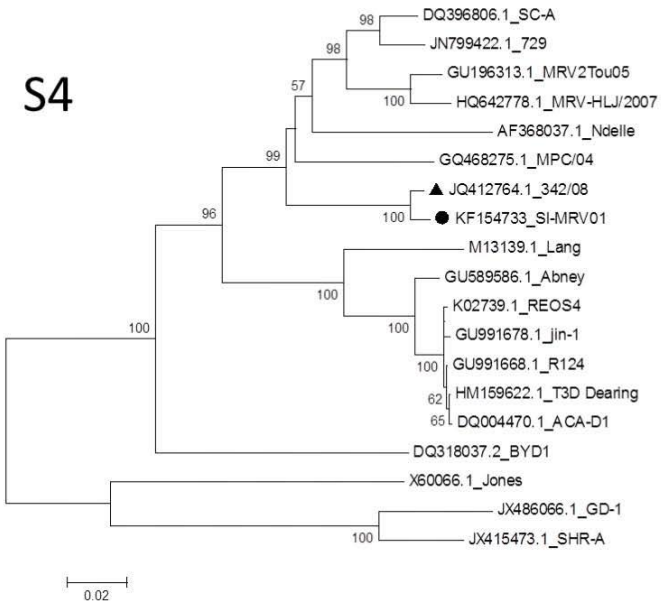
● Slovenian MRV isolate SI-MRV01; ▲ German MRV isolate 342/08 identified in bat

Fig. S5: Phylogenetic analysis of mammalian orthoreoviruses for nucleotide sequence of S3 genome segment.



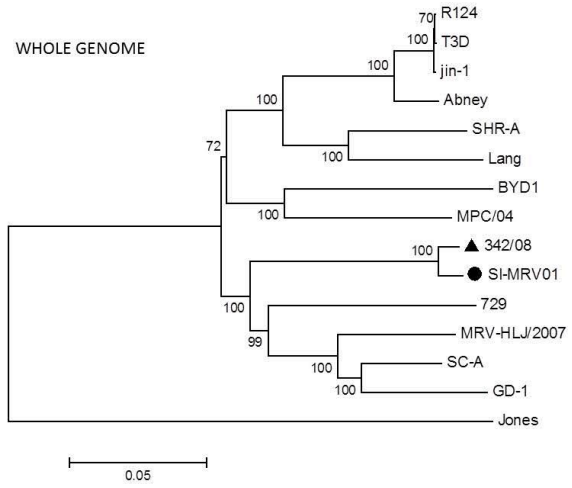
● Slovenian MRV isolate SI-MRV01; ▲ German MRV isolate 342/08 identified in bat

Fig. S6: Phylogenetic analysis of mammalian orthoreoviruses for nucleotide sequence of S4 genome segment.



● Slovenian MRV isolate SI-MRV01; ▲ German MRV isolate 342/08 identified in bat

Fig. S7: Phylogenetic analysis of mammalian orthoreoviruses for whole genome nucleotide sequence.



● Slovenian MRV isolate SI-MRV01; ▲ German MRV isolate 342/08 identified in bat

A.2 Supplementary Material for Publication 2.2

Table S1: Information about primers used for Sanger sequencing validation of PVX genome sequences and discrimination between different groups of isolates

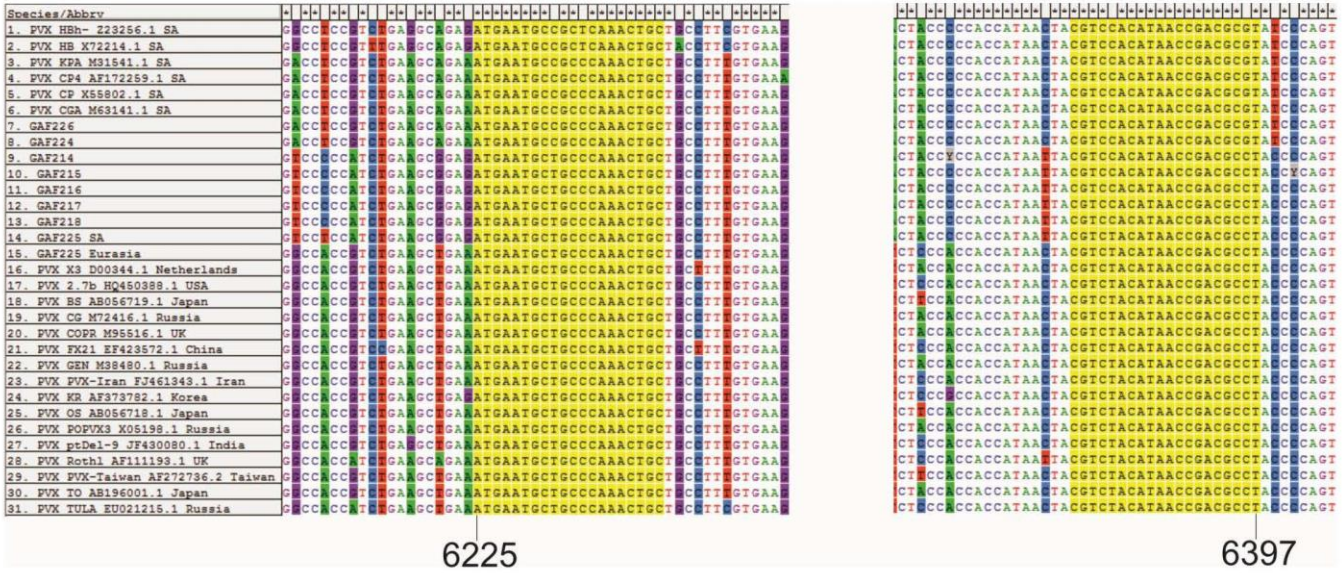
Name	Sequence (5'->3')		annealing position in PVX genome	length of amplicon (bp)	Samples sequenced
	forward primer	reverse primer			
PVX-CIP1	GTWAGYCTTCCYCGCCTCA	TCAAAGTCRAASCCGCARCT	4424-4896	473	GAF226
PVX-CIP2	CCCCACTTCTACTTRGARACCT	GACYGCTATGATTGTTGCCAC	4818-5491	673	GAF226
PVX-CIP4	TTCCYTCRGACACTATGGCT	ATACTATGAAACTGGGRTASGCGTC	5872-6415	543	GAF215; GAF218
PVX-CIP6	AGGCACACGGACACTCAAAT	CATGTCCTCAGGCGTTGTTT	3207-3757	550	GAF226
PVX-SA1	CCCGACGGAAATCATCCCA	CACTTCAGTAGCCGGGCCTA	1871-2645	774	GAF224
PVX-SA_CIP2	CTCAGCTGCACCAATTGAGGAAATA	ACGTAGTGGAYCCTGCGGAC	5072-5697	625	GAF214-218, GAF225 (SA)
Primers used for discrimination between different groups of isolates					
universal for PVX - all groups	PVX-uni	ATGAATGCYGCYCAAACTGC	ASGCGTCGGTTATGTRGACG	6225-6397	172	-
specific for PVX group I	PVX-EuA	YTGCAAGCGTGGAAAGCYT	GCYTGRTGTTTCRCCTCCCTAC	1377-1654	277	GAF225 (EuA)
specific for PVX group II-3	PVX-SA_CIP1	TGCATTTGCCGCTGTACTCCT	AGTTGYGGTAGCTGACGTAGTG	5391-5711	320	-

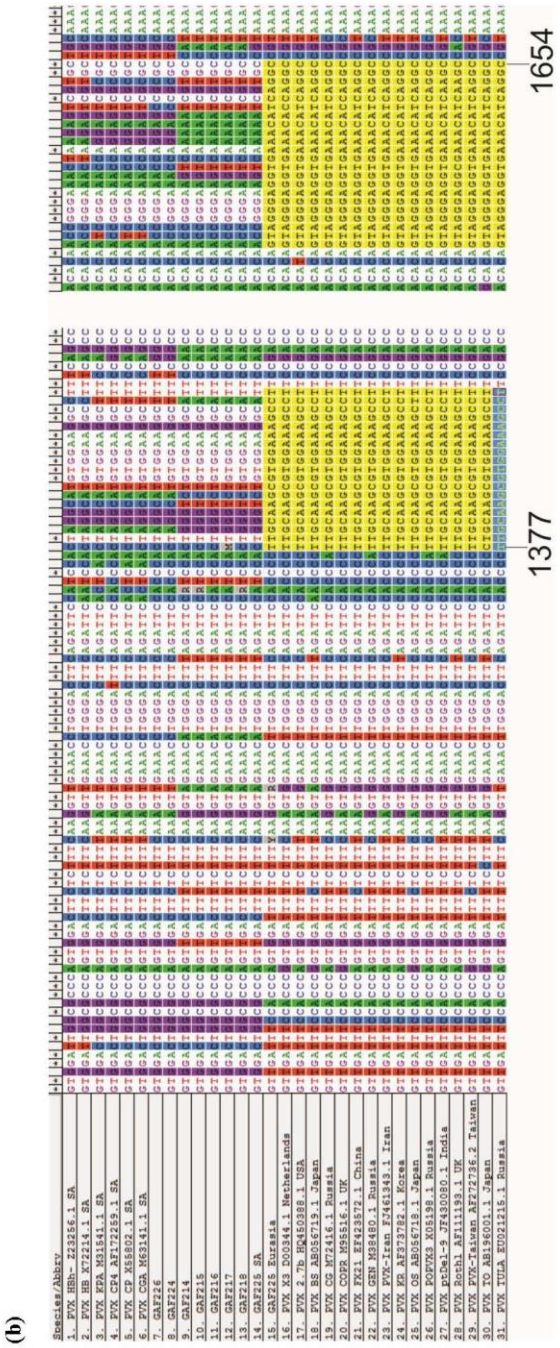
Table S2: Other potexviruses with sequenced complete genomes, included in phylogenetic analysis as an outgroup.

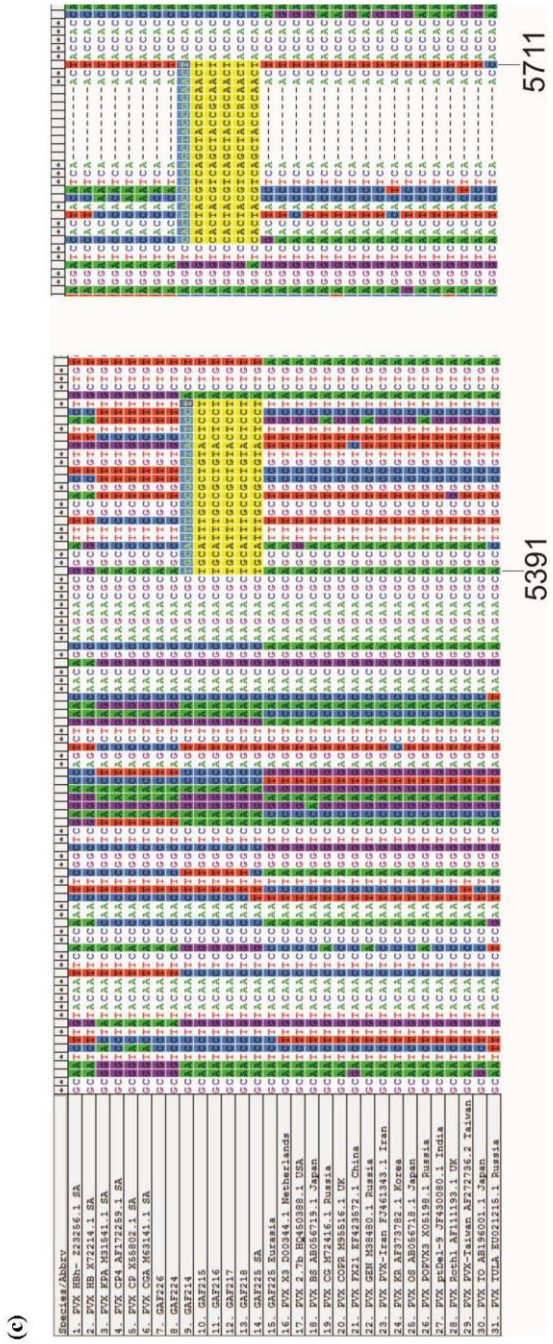
Virus	NCBI GenBank accession number
Potato aucuba mosaic virus	NC 003632.1
Pepino mosaic virus	NC 004067.1
Plantagoasiatica mosaic virus	NC 003849.1
Lily virus X	NC 007192.1
Nerine virus X	NC 007679.1
Shallot virus X	NC 003795.1
Tulip virus X	NC 004322.1
White clover mosaic virus	NC 003820.1
Narcissus mosaic virus	NC 001441.1
Strawberry mild yellow edge virus	NC 003794.1
Foxtail mosaic virus	NC 001483.1
Scallion virus X	NC 003400.1
Hosta virus X	NC 011544.1
Cymbidium mosaic virus	NC 001812.1
Papaya mosaic virus	NC 001748.1
Cactus virus X	NC 002815.2
Bamboo mosaic virus	NC 001642.1
Cassava common mosaic virus	NC 001658.1
Clover yellow mosaic virus	NC 001753.1
Alternanthera mosaic virus	NC 007731.1

Figure S3: Annealing positions of primers used for discrimination between different PVX groups of isolates: a. universal for PVX (PVX-uni); b. specific for PVX group I (PVX-EuA); c. specific for PVX group II-3 (PVX-SA_CIP1); Yellow box shows the position of the primer. The starting and ending position of expected amplicon in the PVX genome is indicated.

(a)




(b)
Species/Abbrv
1. FVX_HBh-_Z23256.1_SA
2. FVX_HB_X72214.1_SA
3. FVX_HPA_M31541.1_SA
4. FVX_CP4_AF172259.1_SA
5. FVX_CP_X65802.1_SA
6. FVX_CGA_M63141.1_SA
7. GAF226
8. GAF224
9. GAF214
10. GAF215
11. GAF216
12. GAF217
13. GAF218
14. GAF225_SA
15. GAF225_Eurasia
16. FVX_K3_D00344.1_Netherlands
17. FVX_2.7b_HQ450388.1_USA
18. FVX_SS_AB056719.1_Japan
19. FVX_CG_M72416.1_Russia
20. FVX_COPR_M35516.1_UK
21. FVX_FX21_EF423572.1_China
22. FVX_GEN_M38480.1_Russia
23. FVX_FVX-Iran_FJ461343.1_Iran
24. FVX_KR_AF373782.1_Korea
25. FVX_OS_AB056718.1_Japan
26. FVX_POFVX3_K05198.1_Russia
27. FVX_prDel-9_JF490080.1_India
28. FVX_Rochl_AF111393.1_UK
29. FVX-Taiwan_AY272736.2_Taiwan
30. FVX_TO_AB360001.1_Japan
31. FVX_YUGA_EU021215.1_Russia
1377
1654

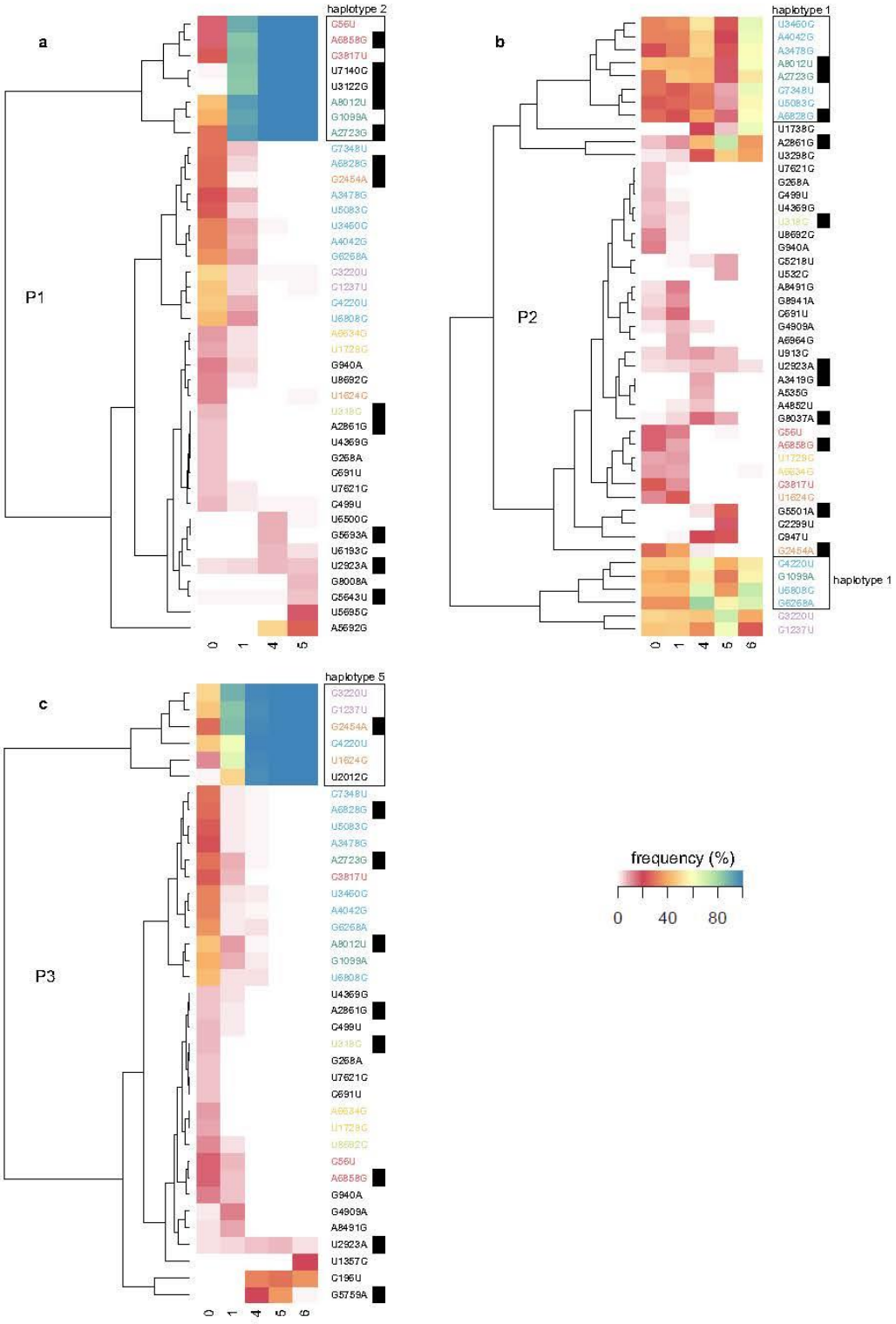

(c)
Species/Abbrv
1. PVX HBh- 232256.1 SA
2. PVX HB X72214.1 SA
3. PVX HPA M31541.1 SA
4. PVX CP4 AF172259.1 SA
5. PVX CP X56802.1 SA
6. PVX CCA M63141.1 SA
7. GAP216
8. GAP214
9. GAP215
10. GAP215
11. GAP216
12. GAP217
13. GAP218
14. GAP226 SA
15. GAP226 Eurasia
16. PVX X3 D00344.1 Nederlands
17. PVX 2.7b HQ450388.1 USA
18. PVX B8 AB056719.1 Japan
19. PVX C6 M72416.1 Russia
20. PVX C6PR M95516.1 UK
21. PVX FK01 EF423572.1 China
22. PVX GEN M38480.1 Russia
23. PVX PVX-Iran FJ461343.1 Iran
24. PVX KR AF373782.1 Korea
25. PVX O8 AB056718.1 Japan
26. PVX POPVX3 X05190.1 Russia
27. PVX pinWal-9 JF430080.1 India
28. PVX Rothl AF111198.1 UK
29. PVX PVX-Taiwan AF272736.2 Taiwan
30. PVX TO AB196001.1 Japan
31. PVX TUCA EU021215.1 Russia
5711
5391

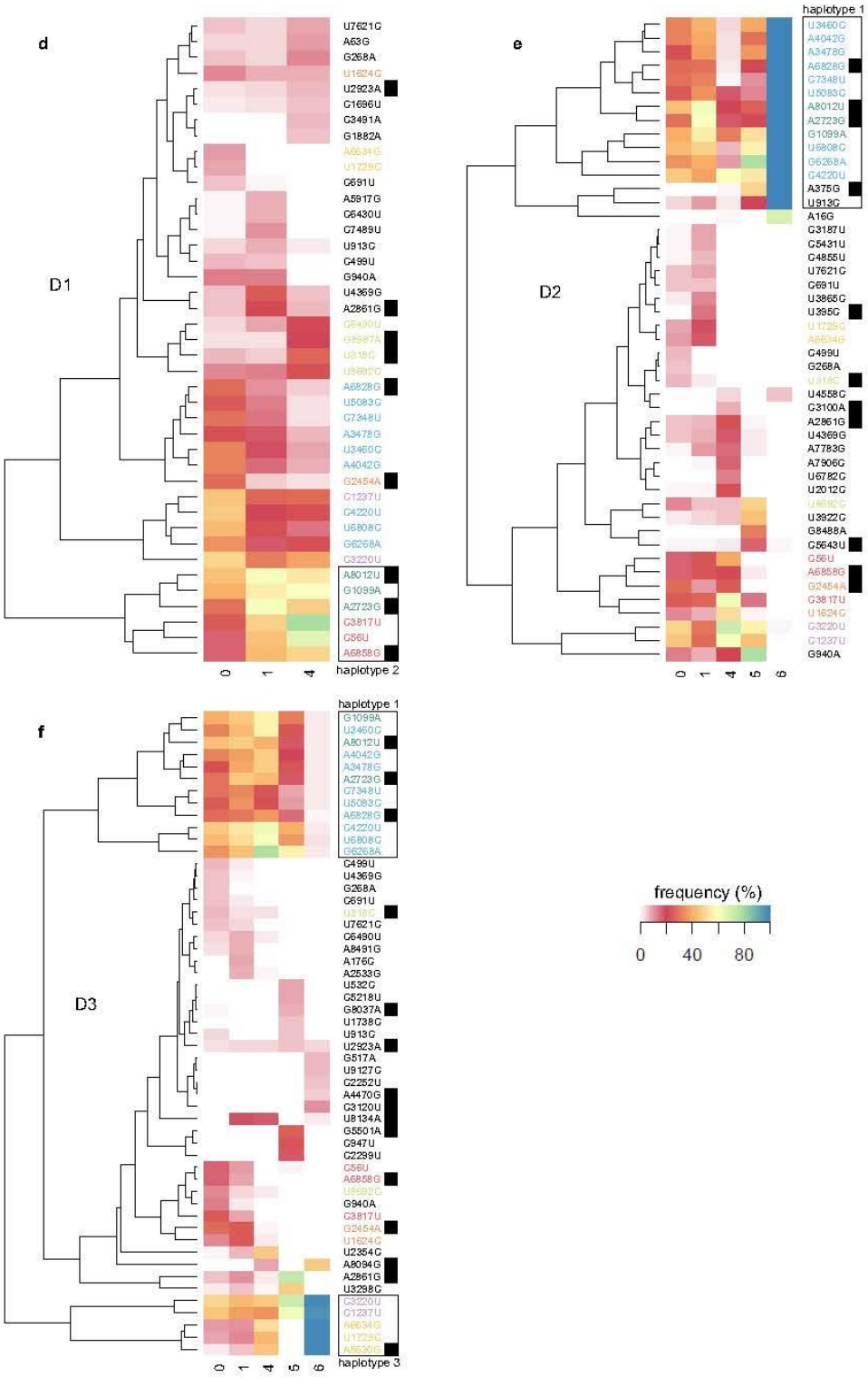
A.3 Supplementary Material for Publication 2.3

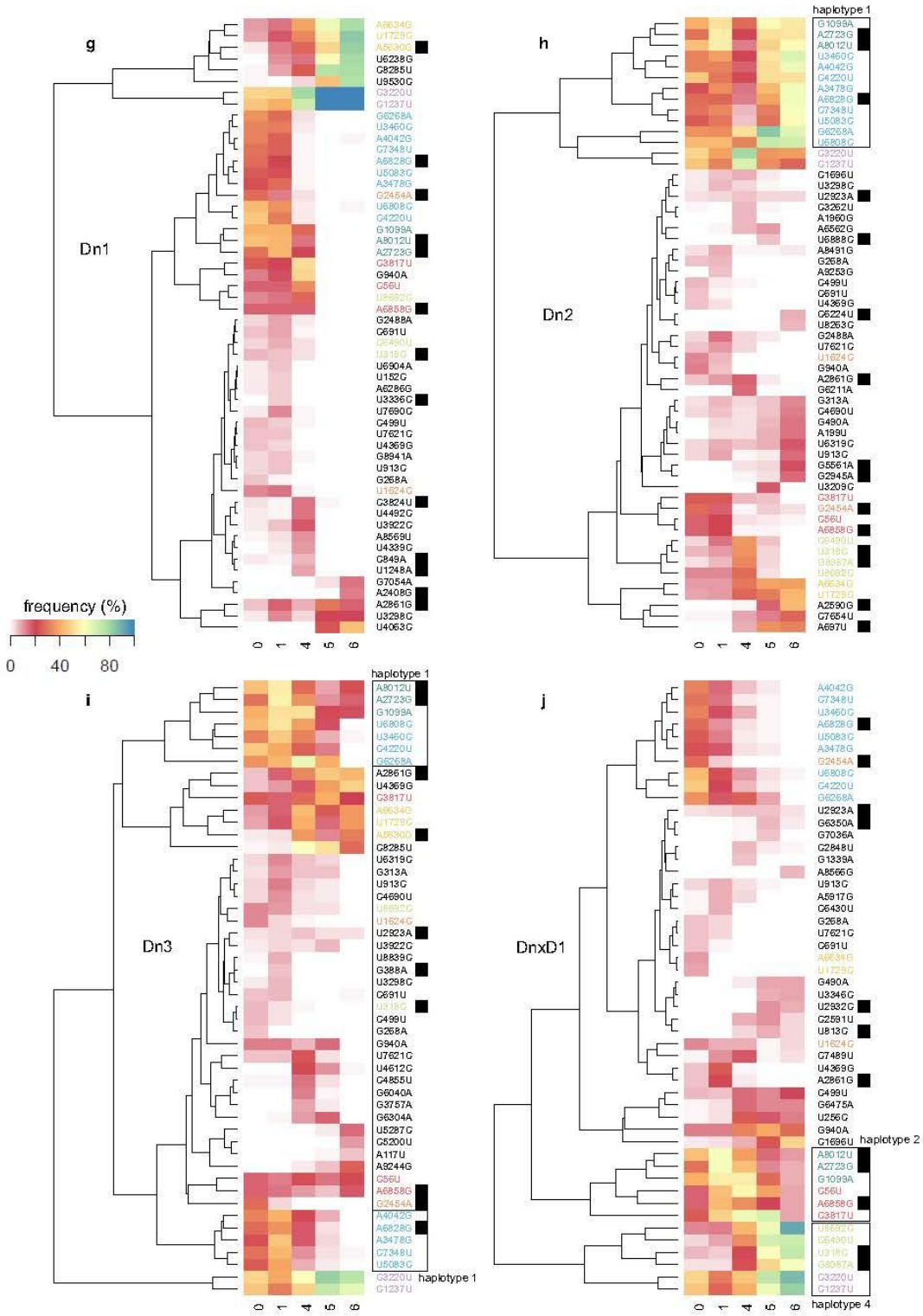
Supplementary material of the publication 2.3 contains only a large table with frequencies of all of the variants detected in the study. It is available online at: http://jvi.asm.org/content/early/2015/02/05/JVI.03685-14/suppl/DCSupplemental.

A.4 Supplementary Material for Publication 2.4

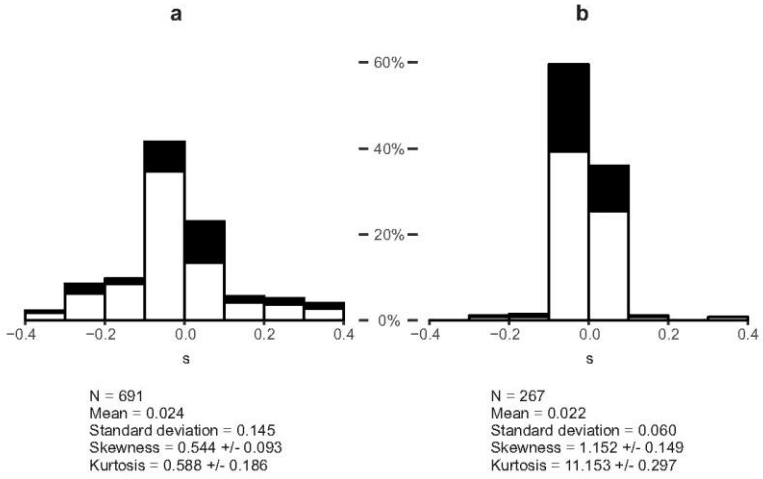
Supplementary figure 1: Clustering of SNPs trajectories. The SNP trajectories were assigned into groups using R 3.2.3 (heatmap.2 library) hierarchical clustering approach (method=average). Only the variants reaching 5% in at least one time point were used for this analysis, since the variants with lower frequencies were not informative for cluster analysis. SNPs trajectories were clustered for each lineage separately (a-j). SNP was treated as a vector with *n* dimensions, dimensions representing frequency of the variant in the sequenced time points. Heatmap.2 library was used to cluster SNPs trajectories and visualize them as a heatmap. SNPs labels are colored according to the groups defined in the analysis shown in the main text on Fig. 5. Reconstructed haplotypes are boxed. Black squares next to the SNPs labels designate nonsynonymous mutations.


a
P1
haplotype 2
C56U
A6858G
G3817U
U7140G
U3122G
A8012U
G1099A
A2723G
G7348U
A6828G
G2454A
A3478G
U5083G
U3460G
A4042G
G6268A
C3220U
G1237U
C4220U
U6808G
A6834G
U1729G
G940A
U8692G
U1624G
U318G
A2861G
U4369G
G268A
C691U
U7621G
C499U
U6500G
G5693A
U6193G
U2923A
G8008A
G5643U
U5695G
A5692G
0 1 4 5
b
P2
haplotype 1
U3460G
A4042G
A3478G
A8012U
A2723G
G7348U
U5083G
A6828G
U1738G
A2861G
U3298G
U7621G
G268A
C499U
U4369G
U318G
U8692G
G940A
C5218U
U532G
A8491G
G8941A
C691U
G4909A
A6964G
U913G
U2923A
A3419G
A535G
A4852U
G8037A
C56U
A6858G
U1729G
A6834G
G3817U
U1624G
G5501A
C2299U
G947U
G2454A
C4220U
G1099A
U6808G
G6268A
G3220U
G1237U
haplotype 1
0 1 4 5 6
c
P3
haplotype 5
G3220U
G1237U
G2454A
C4220U
U1624G
U2012G
G7348U
A6828G
U5083G
A3478G
A2723G
G3817U
U3460G
A4042G
G6268A
A8012U
G1099A
U6808G
U4369G
A2861G
C499U
U318G
G268A
U7621G
C691U
A6834G
U1729G
U8692G
C56U
A6858G
G940A
G4909A
A8491G
U2923A
U1357G
C195U
G5759A
0 1 4 5 6
frequency (%)
0 40 80


d
D1
U7621C
A63G
G268A
U1624C
U2923A
C1696U
G3491A
G1882A
A6634G
U1729C
C691U
A5917G
C6430U
C7489U
U913C
C499U
G940A
U4369G
A2861G
C6490U
G8987A
U318C
U8892C
A6828G
U5083C
C7348U
A3478G
U3460C
A4042G
G2454A
G1237U
C4220U
U6808C
G6268A
G3220U
A8012U
G1098A
A2723G
G3817U
C56U
A6858G
0
1
4
haplotype 2
e
D2
haplotype 1
U3460C
A4042G
A3478G
A6828G
C7348U
U5083C
A8012U
A2723G
G1098A
U6808C
G6268A
C4220U
A375G
U913C
A16G
G3187U
C5431U
C4855U
U7621C
C691U
U3865G
U395C
U1729C
A6634G
C499U
G268A
U318C
U4558C
G3100A
A2861G
U4369G
A7783G
A7906C
U6782C
U2012C
U3922C
G8488A
C5643U
C56U
A6858G
G2454A
G3817U
U1624C
G3220U
G1237U
G940A
0
1
4
5
6
f
D3
haplotype 1
G1098A
U3460C
A8012U
A4042G
A3478G
A2723G
C7348U
U5083C
A6828G
C4220U
U6808C
G6268A
C499U
U4369G
G268A
C691U
U318C
U7621C
C6490U
A8491G
A176G
A2533G
U522C
C5218U
G8037A
U1738C
U813C
U2923A
G517A
U9127C
C2252U
A4470G
C3120U
U8134A
G5501A
G947U
C2299U
C56U
A6858G
U6892C
G940A
G3817U
G2454A
U1624C
U2354C
A8094G
A2861G
U3298C
G3220U
G1237U
A6634G
U1729C
A5630G
0
1
4
5
6
haplotype 3
frequency (%)
0
40
80


g
Dn1
frequency (%)
0 40 80
h
Dn2
haplotype 1
i
Dn3
haplotype 1
j
DnxD1
haplotype 2
haplotype 4

Supplementary figure 2: Distribution of inferred selection coefficients of SNPs for (a) standing variation (present already in the source inoculum) and (b) new variation and their descriptive statistics. The distributions were visualized using R 3.2.3, descriptive statistics were calculated using SPSS 16. White surfaces represent synonymous mutations, whereas black surfaces represent nonsynonymous mutations.



Supplementary Table 1: Samples included in population sequencing and their metadata.

Sample number	Lineage	Passage	Plant designation	Cultivar	Sampling time (dpi)	Type of sample	Average sequencing depth
inoc-1	source inoculum	0	pool	Pentland squire	21	leaf tissue	app. 7500x[1]
inoc-2	source inoculum	0	pool	Pentland squire	21	leaf tissue	app. 7500x[1]
inoc-3	source inoculum	0	pool	Pentland squire	21	inoculum	app. 7500x[1]
DKA127-1	D1	1	D 0.1	Desiree	21	leaf tissue	app. 7500x[1]
DKA127-2	D1	1	D 0.1	Desiree	21	leaf tissue	app. 7500x[1]
DKA1146	D1	4	D 3.1.1	Desiree	21	inoculum	app. 7500x[1]
DKA128	D2	1	D 0.2	Desiree	21	leaf tissue	app. 7500x[1]
DKA1147	D2	4	D 3.2.1	Desiree	21	inoculum	app. 7500x[1]
DKA1157	D2	5	D 4.2.1	Desiree	21	inoculum	2159x
DKA1003	D2	6	D 5.2.7	Desiree	44	leaf tissue	app. 7500x[1]
DKA129	D3	1	D 0.3	Desiree	21	leaf tissue	app. 7500x[1]
DKA1148-1	D3	4	D 3.3.1	Desiree	21	inoculum	app. 7500x[1]
DKA1148-2	D3	4	D 3.3.1	Desiree	21	inoculum	app. 7500x[1]
DKA1158	D3	5	D 4.3.1	Desiree	21	inoculum	app. 7500x[1]
DKA938	D3	6	D 5.3.3	Desiree	21	leaf tissue	app. 7500x[1]
DKA133	Dn1	1	Dn 0.1	Desiree NahG	21	leaf tissue	app. 7500x[1]
DKA1149	Dn1	4	Dn 3.1.1	Desiree NahG	21	inoculum	app. 7500x[1]
DKA1159	Dn1	5	Dn 4.1.8	Desiree NahG	21	inoculum	app. 7500x[1]
DKA943	Dn1	6	Dn 5.1.1	Desiree NahG	21	leaf tissue	app. 7500x[1]
DKA134	Dn2	1	Dn 0.2	Desiree NahG	21	leaf tissue	app. 7500x[1]
DKA1150	Dn2	4	Dn 3.2.1	Desiree NahG	21	inoculum	app. 7500x[1]
DKA1160	Dn2	5	Dn 4.2.1	Desiree NahG	21	inoculum	app. 7500x[1]
DKA946-1	Dn2	6	Dn 5.2.1	Desiree NahG	21	leaf tissue	app. 7500x[1]
DKA946-2	Dn2	6	Dn 5.2.1	Desiree NahG	21	leaf tissue	app. 7500x[1]
DKA135	Dn3	1	Dn 0.3	Desiree NahG	21	leaf tissue	app. 7500x[1]
DKA1151	Dn3	4	Dn 3.3.1	Desiree NahG	21	inoculum	app. 7500x[1]
DKA1161	Dn3	5	Dn 4.3.1	Desiree NahG	21	inoculum	app. 7500x[1]
DKA949	Dn3	6	Dn 5.3.1	Desiree NahG	21	leaf tissue	app. 7500x[1]
DKA1152	DnxD1	4	Dn x D 3.1.1	Desiree NahG	21	inoculum	app. 7500x[1]
DKA1162	DnxD1	5	Dn x D 4.1.1	Desiree	21	inoculum	app. 7500x[1]
DKA961	DnxD1	6	Dn x D 5.1.1	Desiree NahG	21	leaf tissue	app. 7500x[1]
DKA121	P1	1	P 0.1	Pentland squire	21	leaf tissue	app. 7500x[1]
DKA1143	P1	4	P 3.1.7	Pentland squire	21	inoculum	app. 7500x[1]
DKA1153	P1	5	P 4.1.1	Pentland squire	21	inoculum	3385x
DKA123	P2	1	P 0.3	Pentland squire	21	leaf tissue	app. 7500x[1]
DKA1144	P2	4	P 3.3.1	Pentland squire	21	inoculum	app. 7500x[1]
DKA1154	P2	5	P 4.3.1	Pentland squire	21	inoculum	4112x
DKA928	P2	6	P 5.3.1	Pentland squire	21	leaf tissue	app. 7500x[1]
DKA124	P3	1	P 0.4	Pentland squire	21	leaf tissue	app. 7500x[1]
DKA1145	P3	4	P 3.4.1	Pentland squire	21	inoculum	app. 7500x[1]
DKA1155-1	P3	5	P 4.4.1	Pentland squire	21	inoculum	app. 7500x[1]
DKA1155-2	P3	5	P 4.4.1	Pentland squire	21	inoculum	app. 7500x[1]
DKA931	P3	6	P 5.4.1	Pentland squire	21	leaf tissue	app. 7500x[1]

[1]mapping files were normalized to average depth of 7500x

Supplementary Table 2: Results of the Ne analysis using WF-ABC for all the sequenced time points and consecutive pairs of time points. Average and standard deviation (SD) for 1000 bootstrap replicates from wfabc_1 analysis are shown.

	all		In.->1		1->4		4->5		5->6	
	average	*SD*	*average*	*SD*	*average*	*SD*	*average*	*SD*	*average*	*SD*
P1	17.7	3.3	20.8	7.2	14.8	3.5	31.3	15.0	NA	NA
P2	6.7	0.9	38.4	9.9	24.3	5.9	4.2	0.9	2.0	0.4
P3	14.6	2.7	4.0	1.1	36.1	11.0	37.2	8.8	20.1	8.1
D1	21.6	4.4	16.3	4.4	24.4	6.6	NA	NA	NA	NA
D2	2.9	0.3	16.4	4.0	13.8	3.4	1.9	0.4	0.8	0.1
D3	5.6	0.9	16.9	4.4	82.2	18.6	2.2	0.5	1.9	0.5
Dn1	7.9	1.2	29.0	7.1	13.6	2.9	3.5	0.9	4.8	1.8
Dn2	10.4	1.3	33.6	8.6	17.5	3.8	3.7	0.8	9.6	2.7
Dn3	12.6	1.6	18.7	4.7	29.1	6.5	5.0	1.2	8.1	2.1
DnxD1	15.6	2.2	16.3	4.4	16.8	3.7	9.8	2.8	22.9	7.5

Supplementary Table 3: Detailed results of the virus titer measurements using RT-qPCR - source data for Fig. 2.

Supplementary Table 4: Extensive table with all the variants detected in the experiment and their frequencies in the analyzed samples.

There is additional data presented in the supplementary tables 3 and 4. Their titles are listed above, however, they are omitted here, since they represent extensive supplementary data matrices, not necessarily needed for the understanding of the study.

Appendix B

Permissions for Reproduction of Included Publications

B.1 Permission for Reproduction of Publication 2.1



17. 8. 2016 Rightslink® by Copyright Clearance Center

Copyright Clearance Center RightsLink®

Home | Create Account | Help | Live Chat

AMERICAN SOCIETY FOR MICROBIOLOGY

Title: High Similarity of Novel Orthoreovirus Detected in a Child Hospitalized with Acute Gastroenteritis to Mammalian Orthoreoviruses Found in Bats in Europe

Author: Andrej Steyer, Ion Gutiérrez-Aguire, Marko Kolenc et al.

Publication: Journal of Clinical Microbiology

Publisher: American Society for Microbiology

Date: Nov 1, 2013

Copyright © 2013, American Society for Microbiology

LOGIN

If you're a copyright.com user, you can login to RightsLink using your copyright.com credentials. Already a RightsLink user or want to learn more?

Permissions Request

Authors in ASM journals retain the right to republish discrete portions of his/her article in any other publication (including print, CD-ROM, and other electronic formats) of which he or she is author or editor, provided that proper credit is given to the original ASM publication. ASM authors also retain the right to reuse the full article in his/her dissertation or thesis. For a full list of author rights, please see: http://journals.asm.org/site/misc/ASM_Author_Statement.xhtml

BACK | CLOSE WINDOW

Copyright © 2016 Copyright Clearance Center, Inc. All Rights Reserved. Privacy statement. Terms and Conditions.
Comments? We would like to hear from you. E-mail us at customercare@copyright.com

B.2 Permission for Reproduction of Publication 2.2

ELSEVIER LICENSE
TERMS AND CONDITIONS

Aug 17, 2016

This Agreement between Denis Kutnjak ("You") and Elsevier ("Elsevier") consists of your license details and the terms and conditions provided by Elsevier and Copyright Clearance Center.

License Number	3931230774083
License date	Aug 17, 2016
Licensed Content Publisher	Elsevier
Licensed Content Publication	Virus Research
Licensed Content Title	Complete genome sequences of new divergent potato virus X isolates and discrimination between strains in a mixed infection using small RNAs sequencing approach
Licensed Content Author	Denis Kutnjak,Rocio Silvestre,Wilmer Cuellar,Wilmer Perez,Giovanna Müller,Maja Ravnikar,Jan Kreuze
Licensed Content Date	13 October 2014
Licensed Content Volume Number	191
Licensed Content Issue Number	n/a
Licensed Content Pages	6
Start Page	45
End Page	50
Type of Use	reuse in a thesis/dissertation
Portion	full article
Format	both print and electronic
Are you the author of this Elsevier article?	Yes
Will you be translating?	No
Order reference number	
Title of your thesis/dissertation	Discovery and diversity studies of viruses with next generation sequencing
Expected completion date	Dec 2016
Estimated size (number of pages)	100
Elsevier VAT number	GB 494 6272 12
Requestor Location	Denis Kutnjak National Instiute of Biology Vecna pot 111 Ljubljana, 1000 Slovenia Attn: Denis Kutnjak
Customer VAT ID	SI83534784
Total	0.00 EUR

Terms and Conditions

INTRODUCTION

1. The publisher for this copyrighted material is Elsevier. By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the Billing and Payment terms and conditions established by Copyright Clearance Center, Inc. ("CCC"), at the time that you opened your Rightslink account and that are available at any time at http://myaccount.copyright.com).

GENERAL TERMS

2. Elsevier hereby grants you permission to reproduce the aforementioned material subject to the terms and conditions indicated.
3. Acknowledgement: If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source, permission must also be sought from that source. If such permission is not obtained then that material may not be included in your publication/copies. Suitable acknowledgement to the source must be made, either as a footnote or in a reference list at the end of your publication, as follows:
"Reprinted from Publication title, Vol /edition number, Author(s), Title of article / title of chapter, Pages No., Copyright (Year), with permission from Elsevier [OR APPLICABLE SOCIETY COPYRIGHT OWNER]." Also Lancet special credit - "Reprinted from The Lancet, Vol. number, Author(s), Title of article, Pages No., Copyright (Year), with permission from Elsevier."
4. Reproduction of this material is confined to the purpose and/or media for which permission is hereby given.
5. Altering/Modifying Material: Not Permitted. However figures and illustrations may be altered/adapted minimally to serve your work. Any other abbreviations, additions, deletions and/or any other alterations shall be made only with prior written authorization of Elsevier Ltd. (Please contact Elsevier at permissions@elsevier.com)
6. If the permission fee for the requested use of our material is waived in this instance, please be advised that your future requests for Elsevier materials may attract a fee.
7. Reservation of Rights: Publisher reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.
8. License Contingent Upon Payment: While you may exercise the rights licensed immediately upon issuance of the license at the end of the licensing process for the transaction, provided that you have disclosed complete and accurate details of your proposed use, no license is finally effective unless and until full payment is received from you (either by publisher or by CCC) as provided in CCC's Billing and Payment terms and conditions. If full payment is not received on a timely basis, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and publisher reserves the right to take any and all action to protect its copyright in the materials.
9. Warranties: Publisher makes no representations or warranties with respect to the licensed material.
10. Indemnity: You hereby indemnify and agree to hold harmless publisher and CCC, and

their respective officers, directors, employees and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.
11. No Transfer of License: This license is personal to you and may not be sublicensed, assigned, or transferred by you to any other person without publisher's written permission.
12. No Amendment Except in Writing: This license may not be amended except in a writing signed by both parties (or, in the case of publisher, by CCC on publisher's behalf).
13. Objection to Contrary Terms: Publisher hereby objects to any terms contained in any purchase order, acknowledgment, check endorsement or other writing prepared by you, which terms are inconsistent with these terms and conditions or CCC's Billing and Payment terms and conditions. These terms and conditions, together with CCC's Billing and Payment terms and conditions (which are incorporated herein), comprise the entire agreement between you and publisher (and CCC) concerning this licensing transaction. In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall control.
14. Revocation: Elsevier or Copyright Clearance Center may deny the permissions described in this License at their sole discretion, for any reason or no reason, with a full refund payable to you. Notice of such denial will be made using the contact information provided by you. Failure to receive such notice will not alter or invalidate the denial. In no event will Elsevier or Copyright Clearance Center be responsible or liable for any costs, expenses or damage incurred by you as a result of a denial of your permission request, other than a refund of the amount(s) paid by you to Elsevier and/or Copyright Clearance Center for denied permissions.

LIMITED LICENSE

The following terms and conditions apply only to specific license types:
15. **Translation**: This permission is granted for non-exclusive world **English** rights only unless your license was granted for translation rights. If you licensed translation rights you may only translate this content into the languages you requested. A professional translator must perform all translations and reproduce the content word for word preserving the integrity of the article.
16. **Posting licensed content on any Website**: The following terms and conditions apply as follows: Licensing material from an Elsevier journal: All content posted to the web site must maintain the copyright information line on the bottom of each image; A hyper-text must be included to the Homepage of the journal from which you are licensing at http://www.sciencedirect.com/science/journal/xxxxx or the Elsevier homepage for books at http://www.elsevier.com; Central Storage: This license does not include permission for a scanned version of the material to be stored in a central repository such as that provided by Heron/XanEdu.
Licensing material from an Elsevier book: A hyper-text link must be included to the Elsevier homepage at http://www.elsevier.com . All content posted to the web site must maintain the copyright information line on the bottom of each image.

Posting licensed content on Electronic reserve: In addition to the above the following clauses are applicable: The web site must be password-protected and made available only to bona fide students registered on a relevant course. This permission is granted for 1 year only. You may obtain a new license for future website posting.
17. **For journal authors:** the following clauses are applicable in addition to the above:
Preprints:
A preprint is an author's own write-up of research results and analysis, it has not been

private sharing rights for others' research accessed under that agreement. This includes use for classroom teaching and internal training at the institution (including use in course packs and courseware programs), and inclusion of the article for grant funding purposes.
Gold Open Access Articles: May be shared according to the author-selected end-user license and should contain a CrossMark logo, the end user license, and a DOI link to the formal publication on ScienceDirect.
Please refer to Elsevier's posting policy for further information.
18. **For book authors** the following clauses are applicable in addition to the above:
Authors are permitted to place a brief summary of their work online only. You are not allowed to download and post the published electronic version of your chapter, nor may you scan the printed edition to create an electronic version. **Posting to a repository:** Authors are permitted to post a summary of their chapter only in their institution's repository.
19. **Thesis/Dissertation**: If your license is for use in a thesis/dissertation your thesis may be submitted to your institution in either print or electronic form. Should your thesis be published commercially, please reapply for permission. These requirements include permission for the Library and Archives of Canada to supply single copies, on demand, of the complete thesis and include permission for Proquest/UMI to supply single copies, on demand, of the complete thesis. Should your thesis be published commercially, please reapply for permission. Theses and dissertations which contain embedded PJAs as part of the formal submission can be posted publicly by the awarding institution with DOI links back to the formal publications on ScienceDirect.

Elsevier Open Access Terms and Conditions
You can publish open access with Elsevier in hundreds of open access journals or in nearly 2000 established subscription journals that support open access publishing. Permitted third party re-use of these open access articles is defined by the author's choice of Creative Commons user license. See our open access license policy for more information.
Terms & Conditions applicable to all Open Access articles published with Elsevier:
Any reuse of the article must not represent the author as endorsing the adaptation of the article nor should the article be modified in such a way as to damage the author's honour or reputation. If any changes have been made, such changes must be clearly indicated.
The author(s) must be appropriately credited and we ask that you include the end user license and a DOI link to the formal publication on ScienceDirect.
If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source it is the responsibility of the user to ensure their reuse complies with the terms and conditions determined by the rights holder.
Additional Terms & Conditions applicable to each Creative Commons user license:
CC BY: The CC-BY license allows users to copy, to create extracts, abstracts and new works from the Article, to alter and revise the Article and to make commercial use of the Article (including reuse and/or resale of the Article by commercial entities), provided the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, indicates if changes were made and the licensor is not represented as endorsing the use made of the work. The full details of the license are available at http://creativecommons.org/licenses/by/4.0.
CC BY NC SA: The CC BY-NC-SA license allows users to copy, to create extracts, abstracts and new works from the Article, to alter and revise the Article, provided this is not done for commercial purposes, and that the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, indicates if changes were made and the licensor is not represented as endorsing the use made of the work. Further, any new works must be made available on the same conditions. The full

details of the license are available at http://creativecommons.org/licenses/by-nc-sa/4.0.
CC BY NC ND: The CC BY-NC-ND license allows users to copy and distribute the Article, provided this is not done for commercial purposes and further does not permit distribution of the Article if it is changed or edited in any way, and provided the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, and that the licensor is not represented as endorsing the use made of the work. The full details of the license are available at http://creativecommons.org/licenses/by-nc-nd/4.0.
Any commercial reuse of Open Access articles published with a CC BY NC SA or CC BY NC ND license requires permission from Elsevier and will be subject to a fee.
Commercial reuse includes:

- Associating advertising with the full text of the Article
- Charging fees for document delivery or access
- Article aggregation
- Systematic distribution via e-mail lists or share buttons

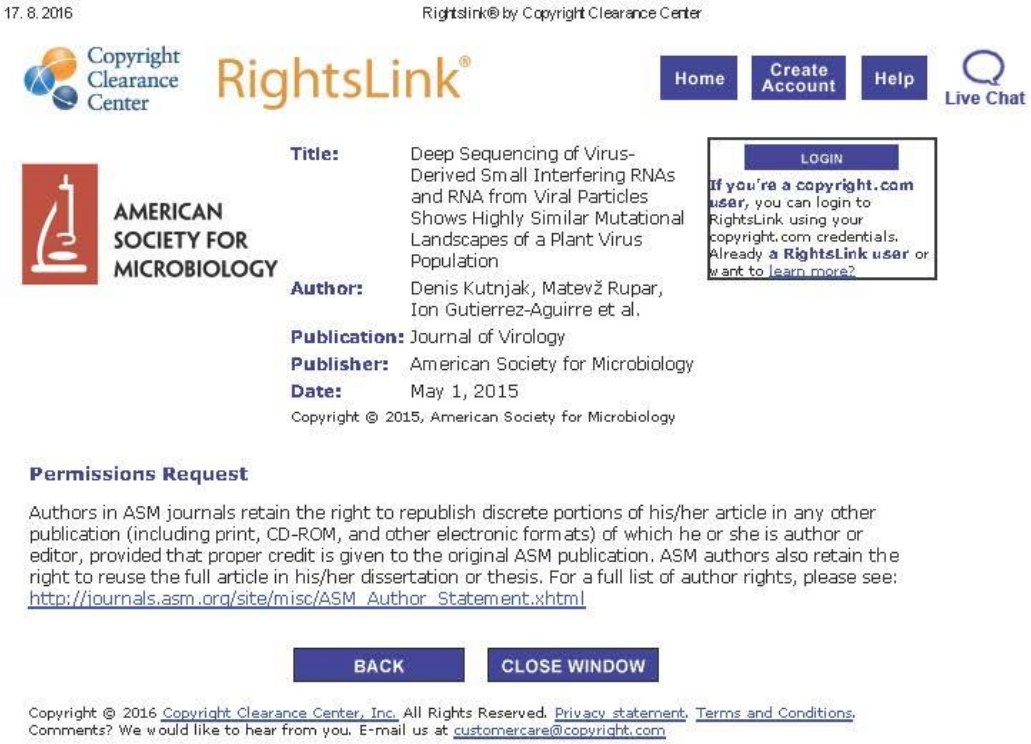
Posting or linking by commercial companies for use by customers of those companies.

20. **Other Conditions**:

v1.8

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.

B.3 Permission for Reproduction of Publication 2.3



17. 8. 2016 Rightslink® by Copyright Clearance Center

Copyright Clearance Center RightsLink®

Home | Create Account | Help | Live Chat

AMERICAN SOCIETY FOR MICROBIOLOGY

Title: Deep Sequencing of Virus-Derived Small Interfering RNAs and RNA from Viral Particles Shows Highly Similar Mutational Landscapes of a Plant Virus Population

Author: Denis Kutnjak, Matevž Rupar, Ion Gutierrez-Aguirre et al.

Publication: Journal of Virology

Publisher: American Society for Microbiology

Date: May 1, 2015

Copyright © 2015, American Society for Microbiology

LOGIN

If you're a copyright.com user, you can login to RightsLink using your copyright.com credentials. Already a RightsLink user or want to learn more?

Permissions Request

Authors in ASM journals retain the right to republish discrete portions of his/her article in any other publication (including print, CD-ROM, and other electronic formats) of which he or she is author or editor, provided that proper credit is given to the original ASM publication. ASM authors also retain the right to reuse the full article in his/her dissertation or thesis. For a full list of author rights, please see: http://journals.asm.org/site/misc/ASM_Author_Statement.xhtml

BACK CLOSE WINDOW

Copyright © 2016 Copyright Clearance Center, Inc. All Rights Reserved. Privacy statement. Terms and Conditions.
Comments? We would like to hear from you. E-mail us at customercare@copyright.com

References

[1] E. V. Koonin and V. V. Dolja, "A virocentric perspective on the evolution of life," *Curr. Opin. Virol.*, vol. 3, no. 5, pp. 546–557, 2013.

[2] J. Flint, V. R. Racaniello, G. F. Rall, and A. M. Skalka, *Principles of Virology, Volume I: Molecular Biology*, 4th ed., no. 1. Washington: American Society of Microbiology, ASM Press, 2015.

[3] C. a. Suttle, "Marine viruses — major players in the global ecosystem," *Nat. Rev. Microbiol.*, vol. 5, no. 10, pp. 801–812, 2007.

[4] A. López-Bueno, J. Tamames, D. Velázquez, A. Moya, A. Quesada, and A. Alcamí, "High diversity of the viral community from an Antarctic lake," *Science*, vol. 326, no. 5954, pp. 858–861, 2009.

[5] T. W. Whon, M.-S. Kim, S. W. Roh, N.-R. Shin, H.-W. Lee, and J.-W. Bae, "Metagenomic characterization of airborne viral DNA diversity in the near-surface atmosphere.," *J. Virol.*, vol. 86, no. 15, pp. 8221–31, 2012.

[6] "http://www.ictvonline.org/taxonomyReleases.asp," *International Committee on Taxonomy of Viruses*, 2015. [Online]. Available: http://www.ictvonline.org/taxonomyReleases.asp. [Accessed: 23-Aug-2016].

[7] S. J. Anthony, J. H. Epstein, K. A. Murray, I. Navarrete-Macias, C. M. Zambrana-Torrelio, A. Solovyov, R. Ojeda-Flores, N. C. Arrigo, A. Islam, S. Ali Khan, P. Hosseini, T. L. Bogich, K. J. Olival, M. D. Sanchez-Leon, W. B. Karesh, T. Goldstein, S. P. Luby, S. S. Morse, J. A. K. Mazet, P. Daszak, and W. I. Lipkin, "A strategy to estimate unknown viral diversity in mammals.," *MBio*, vol. 4, no. 5, pp. e00598-13, 2013.

[8] R. Sanjuán and P. Domingo-Calap, "Mechanisms of viral mutation," *Cell. Mol. Life Sci.*, pp. 1–16, 2016.

[9] E. Domingo, J. Sheldon, and C. Perales, "Viral quasispecies evolution.," *Microbiol. Mol. Biol. Rev.*, vol. 76, no. 2, pp. 159–216, Jun. 2012.

[10] N. Vassilakos, V. Simon, A. Tzima, E. Johansen, and B. Moury, "Genetic determinism and evolutionary reconstruction of a host jump in a plant virus," *Mol. Biol. Evol.*, vol. 33, no. 2, pp. 541–553, 2016.

[11] N. Boonham, J. Kreuze, S. Winter, R. van der Vlugt, J. Bergervoet, J. Tomlinson, and R. Mumford, "Methods in virus diagnostics: From ELISA to next generation sequencing," *Virus Res.*, vol. 186, pp. 20–31, 2014.

[12] M. L. Metzker, "Sequencing technologies - the next generation.," *Nat. Rev. Genet.*, vol. 11, no. 1, pp. 31–46, 2010.

[13] A. D. Radford, D. Chapman, L. Dixon, J. Chantrey, A. C. Darby, and N. Hall,

"Application of next-generation sequencing technologies in virology," *Journal of General Virology*, vol. 93. pp. 1853–1868, Sep-2012.

[14] X. Ge, Y. Li, X. Yang, H. Zhang, P. Zhou, Y. Zhang, and Z. Shi, "Metagenomic analysis of viruses from bat fecal samples reveals many novel viruses in insectivorous bats in China," *J Virol*, vol. 86, no. 8, pp. 4620–4630, 2012.

[15] R. J. Hall, J. Wang, A. K. Todd, A. B. Bissielo, S. Yen, H. Strydom, N. E. Moore, X. Ren, Q. S. Huang, P. E. Carter, and M. Peacey, "Evaluation of rapid and simple techniques for the enrichment of viruses prior to metagenomic virus discovery," *J. Virol. Methods*, vol. 195, pp. 194–204, 2014.

[16] Q. Wu, S.-W. Ding, Y. Zhang, and S. Zhu, "Identification of Viruses and Viroids by Next-Generation Sequencing and Homology-Dependent and Homology-Independent Algorithms," *Annu. Rev. Phytopathol.*, vol. 53, no. 1, pp. 452–44, 2014.

[17] I. Adams and A. Fox, "Diagnosis of Plant Viruses Using Next-Generation Sequencing and Metagenomic Analysis," in *Current Research Topics in Plant Virology*, A. Wang and X. Zhou, Eds. Springer International Publishing Switzerland, 2016, pp. 323–335.

[18] A. Routh, T. Domitrovic, and J. E. Johnson, "Host RNAs, including transposons, are encapsidated by a eukaryotic single-stranded RNA virus.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 109, no. 6, pp. 1907–12, Feb. 2012.

[19] M. J. Roossinck, P. Saha, G. B. Wiley, J. Quan, J. D. White, H. Lai, F. Chavarría, G. Shen, and B. a. Roe, "Ecogenomics: using massively parallel pyrosequencing to understand virus ecology," *Mol. Ecol.*, vol. 19, pp. 81–88, 2010.

[20] I. P. Adams, R. H. Glover, W. A. Monger, R. Mumford, E. Jackeviciene, M. Navalinskiene, M. Samuitiene, and N. Boonham, "Next-generation sequencing and metagenomic analysis: a universal diagnostic tool in plant virology.," *Mol. Plant Pathol.*, vol. 10, no. 4, pp. 537–45, Jul. 2009.

[21] J. F. Kreuze, A. Perez, M. Untiveros, D. Quispe, S. Fuentes, I. Barker, and R. Simon, "Complete viral genome sequence and discovery of novel viruses by deep sequencing of small RNAs: a generic method for diagnosis, discovery and sequencing of viruses.," *Virology*, vol. 388, no. 1, pp. 1–7, May 2009.

[22] M. Rupar, M. Ravnikar, M. Tušek-Žnidarič, P. Kramberger, L. Glais, and I. Gutiérrez-Aguirre, "Fast purification of the filamentous Potato virus Y using monolithic chromatographic supports.," *J. Chromatogr. A*, vol. 1272, pp. 33–40, Jan. 2013.

[23] P. Kramberger, M. Peterka, J. Boben, M. Ravnikar, and A. Štrancar, "Short monolithic columns-A breakthrough in purification and fast quantification of tomato mosaic virus," *J. Chromatogr. A*, vol. 1144, no. 1, pp. 143–149, 2007.

[24] Q. Wu, Y. Luo, R. Lu, N. Lau, E. C. Lai, W.-X. Li, and S.-W. Ding, "Virus discovery by deep sequencing and assembly of virus-derived small silencing RNAs.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 107, no. 4, pp. 1606–11, Jan. 2010.

[25] J. F. Kreuze, "siRNA Deep Sequencing and Assembly: Piecing Together Viral Infections," in *Detection and Diagnostics of Plant Pathogens*, 2014, pp. 21–38.

[26] C. B. Jabara, C. D. Jones, J. Roach, J. A. Anderson, and R. Swanstrom, "Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID.," *Proc.*

Natl. Acad. Sci. U. S. A., vol. 108, no. 50, pp. 20166–71, 2011.

[27] A. Acevedo, L. Brodsky, and R. Andino, "Mutational and fitness landscapes of an RNA virus revealed through population sequencing.," *Nature*, vol. 505, no. 7485, pp. 686–90, Jan. 2014.

[28] J. Seguin, R. Rajeswaran, N. Malpica-López, R. R. Martin, K. Kasschau, V. V Dolja, P. Otten, L. Farinelli, and M. M. Pooggin, "De Novo Reconstruction of Consensus Master Genomes of Plant RNA and DNA Viruses from siRNAs.," *PLoS One*, vol. 9, no. 2, p. e88513, Jan. 2014.

[29] V. I. Maliogka, B. Salvador, A. Carbonell, P. Sáenz, D. S. A. N. León, J. C. Oliveros, M. A. O. Delgadillo, J. A. García, and C. Simón-mateo, "Virus variants with differences in the P1 protein coexist in a Plum pox virus population and display particular host-dependent pathogenicity features," vol. 13, pp. 877–886, 2012.

[30] S. F. Elena and R. Sanjuán, "Virus Evolution: Insights from an Experimental Approach," *Annu. Rev. Ecol. Evol. Syst.*, vol. 38, no. 1, pp. 27–52, 2007.

[31] R. J. Orton, C. F. Wright, M. J. Morelli, N. Juleff, G. Thébaud, N. J. Knowles, B. Valdazo-González, D. J. Paton, D. P. King, and D. T. Haydon, "Observing micro-evolutionary processes of viral populations at multiple scales," *Philos. Trans. R. Soc. B Biol. Sci.*, vol. 368, no. 1614, p. 20120203, 2013.

[32] S. F. Elena, "Evolutionary transitions during RNA virus experimental evolution," *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, vol. 371, p. 20150441, 2016.

[33] C. J. R. Illingworth, "Fitness Inference from Short-Read Data: Within-Host Evolution of a Reassortant H5N1 Influenza Virus.," *Mol. Biol. Evol.*, vol. 32, no. 11, pp. 3012–3026, 2015.

[34] M. Foll, Y. P. Poh, N. Renzette, A. Ferrer-Admetlla, C. Bank, H. Shim, A. S. Malaspinas, G. Ewing, P. Liu, D. Wegmann, D. R. Caffrey, K. B. Zeldovich, D. N. Bolon, J. P. Wang, T. F. Kowalik, C. a. Schiffer, R. W. Finberg, and J. D. Jensen, "Influenza virus drug resistance: a time-sampled population genetics perspective," *PLoS Genet.*, vol. 10, no. 2, p. e1004185, 2014.

[35] A. V Bordería, O. Isakov, G. Moratorio, R. Henningsson, S. Agüera-González, L. Organtini, N. F. Gnädig, H. Blanc, A. Alcover, S. Hafenstein, M. Fontes, N. Shomron, and M. Vignuzzi, "Group selection and contribution of minority variants during virus adaptation determines virus fitness and phenotype.," *PLoS Pathog.*, vol. 11, no. 5, p. e1004838, 2015.

[36] M. Schirmer, W. T. Sloan, and C. Quince, "Benchmarking of viral haplotype reconstruction programmes: an overview of the capacities and limitations of currently available programmes," *Brief. Bioinform.*, vol. 15, no. 3, pp. 431–442, 2014.

[37] G. I. Lang, D. P. Rice, M. J. Hickman, E. Sodergren, G. M. Weinstock, D. Botstein, and M. M. Desai, "Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations.," *Nature*, vol. 500, no. 7464, pp. 571–4, 2013.

[38] S. K. Remold, A. Rambaut, and P. E. Turner, "Evolutionary genomics of host adaptation in vesicular stomatitis virus," *Mol. Biol. Evol.*, vol. 25, no. 6, pp. 1138–1147, 2008.

[39] K. A. Stapleford, L. L. Coffey, S. Lay, A. V Bordería, V. Duong, O. Isakov, K.

Rozen-Gagnon, C. Arias-Goeta, H. Blanc, S. Beaucourt, T. Haliloğlu, C. Schmitt, I. Bonne, N. Ben-Tal, N. Shomron, A.-B. Failloux, P. Buchy, and M. Vignuzzi, "Emergence and transmission of arbovirus evolutionary intermediates with epidemic potential," *Cell Host Microbe*, vol. 15, no. 6, pp. 706–16, Jun. 2014.

[40] J. Hillung, J. M. Cuevas, S. Valverde, and S. F. Elena, "Experimental evolution of an emerging plant virus in host genotypes that differ in their susceptibility to infection.," *Evolution*, vol. 68, no. 9, pp. 2467–80, Sep. 2014.

[41] T. Stare, Ž. Ramšak, A. Blejec, K. Stare, N. Turnšek, W. Weckwerth, S. Wienkoop, D. Vodnik, and K. Gruden, "Bimodal dynamics of primary metabolism-related responses in tolerant potato-Potato virus Y interaction.," *BMC Genomics*, vol. 16, no. 1, p. 716, 2015.

[42] Š. Baebler, K. Stare, M. Kovač, A. Blejec, N. Prezelj, T. Stare, P. Kogovšek, M. Pompe-Novak, S. Rosahl, M. Ravnikar, and K. Gruden, "Dynamics of responses in compatible potato-Potato virus Y interaction are modulated by salicylic acid.," *PLoS One*, vol. 6, no. 12, p. e29009, Jan. 2011.

[43] P. Kogovšek, M. Pompe-Novak, M. Petek, L. Fragner, W. Weckwerth, and K. Gruden, "Primary metabolism, phenylpropanoids and antioxidant pathways are regulated in potato as a response to Potato virus Y infection," *PLoS One*, vol. 11, no. 1, pp. 1–20, 2016.

[44] I. Gutiérrez-Aguirre, M. Banjac, A. Steyer, M. Poljšak-Prijatelj, M. Peterka, A. Štrancar, and M. Ravnikar, "Concentrating rotaviruses from water samples using monolithic chromatographic supports," *J. Chromatogr. A*, vol. 1216, no. 13, pp. 2700–2704, 2009.

[45] C. Kohl and A. Kurth, "European bats as carriers of viruses with zoonotic potential," *Viruses*, vol. 6, no. 8, pp. 3110–3128, 2014.

[46] L. Wang, S. Fu, L. Cao, W. Lei, Y. Cao, J. Song, Q. Tang, H. Zhang, Y. Feng, W. Yang, and G. Liang, "Isolation and identification of a natural reassortant mammalian orthoreovirus from least horseshoe bat in China," *PLoS One*, vol. 10, no. 3, p. e0118598, 2015.

[47] D. Lelli, A. Moreno, A. Steyer, T. Naglič, C. Chiapponi, A. Prosperi, F. Faccin, E. Sozzi, and A. Lavazza, "Detection and characterization of a novel reassortant mammalian orthoreovirus in bats in Europe," *Viruses*, vol. 7, no. 11, pp. 5844–5854, 2015.

[48] Z. Li, D. Liu, X. Ran, C. Liu, D. Guo, X. Hu, J. Tian, X. Zhang, Y. Shao, S. Liu, and L. Qu, "Characterization and pathogenicity of a novel mammalian orthoreovirus from wild short-nosed fruit bats," *Infect. Genet. Evol.*, vol. 43, pp. 347–353, 2016.

Bibliography

Publications Related to the Thesis

Journal Articles

A. Steyer, I. Gutiérrez-Aguire, M. Kolenc, S. Koren, D. Kutnjak, M. Pokorn, M. Poljšak-Prijatelj, N. Racki, M. Ravnikar, M. Sagadin, A. Fratnik Steyer, and N. Toplak, "High similarity of novel orthoreovirus detected in a child hospitalized with acute gastroenteritis to mammalian orthoreoviruses found in bats in Europe.," J. Clin. Microbiol., vol. 51, no. 11, pp. 3818–25, 2013.

D. Kutnjak, R. Silvestre, W. Cuellar, W. Perez, G. Müller, M. Ravnikar, and J. Kreuze, "Complete genome sequences of new divergent potato virus X isolates and discrimination between strains in a mixed infection using small RNAs sequencing approach.," Virus Res., vol. 191C, pp. 45–50, Jul. 2014.

D. Kutnjak, M. Rupar, I. Gutierrez-Aguirre, T. Curk, J. F. Kreuze, and M. Ravnikar, "Deep sequencing of virus-derived small interfering RNAs and RNA from viral particles shows highly similar mutational landscapes of a plant virus population," J. Virol., vol. 89, no. 9, pp. 4760–4769, 2015.

Biography

Denis Kutnjak was born on January 22, 1988 in Celje, Slovenia. He finished his primary and secondary education in Velenje, Slovenia. He started his studies at the Biotechnical Faculty, University of Ljubljana, Slovenia, in 2006. In 2011, he visited the Institute of Botany at the University of Innsbruck, Austria, where he has conducted the laboratory part of the research for his graduation thesis, under the supervision of Assist. Prof. Božo Frajman, PhD, and Prof. Peter Schönswetter, PhD. In 2012 he was awarded a Diploma degree in Biology for the thesis entitled "Phylogeography and conservation genetics of *Cerastium dinaricum*" from the University of Ljubljana. He received the Prešeren award of the Biotechnical Faculty for his thesis.

In 2012 he was selected as a PhD candidate at the National Institute of Biology, Ljubljana, Slovenia under the supervision of Prof. Maja Ravnikar, PhD, at the Department of Biotechnology and Systems Biology. At the same time he started his postgraduate studies in Nanosciences and Nanotechnologies at the Jožef Stefan International Postgraduate School, Ljubljana, Slovenia, also under the supervision of Prof. Maja Ravnikar, PhD.

His research focuses on different aspects of virus diversity, using NGS as an investigation tool. During the PhD studies he has visited two foreign laboratories, first in 2012, for a 2-month research stay at the Virology department of the International Potato Center in Lima, Peru, where he was trained in small RNA deep sequencing sample preparation and analysis, under the supervision of Jan F. Kreuze, PhD. In 2016, he visited the laboratory of the Evolutionary Systems Virology Group at the Instituto de Biología Molecular y Celular de Plantas, Polytechnic University of Valencia, Spain, performing population genetics analyses of viral populations, under the supervision of Prof. Santiago F. Elena, PhD. He has attended several workshops on the topic of bioinformatics, virus metagenomics, NGS for plant research, molecular phylogenetics, within-species diversity studies and presented his work at several international conferences.

He is currently involved in the Slovenian Research Agency (L4-5525) and EU-funded projects (ArimNet Emeramb). He is a management committee substitute for Slovenia in the EU-funded COST project FA1407: "DIVAS – deep investigation of virus associated sequences".