



---

MSU Graduate Theses

---

Summer 2019

## Sequence Analysis of the REN1 Genomic Region from the Grapevine 'Kishmish Vatkana'


Dani Joseph

Missouri State University, Dani357@live.missouristate.edu

As with any intellectual project, the content and views expressed in this thesis may be considered objectionable by some readers. However, this student-scholar's work has been judged to have academic value by the student's thesis committee members trained in the discipline. The content and views expressed in this thesis are those of the student-scholar and are not endorsed by Missouri State University, its Graduate College, or its employees.

---

Follow this and additional works at: <https://bearworks.missouristate.edu/theses>

 Part of the [Bioinformatics Commons](#), [Biology Commons](#), [Biotechnology Commons](#), [Genetics Commons](#), [Molecular Genetics Commons](#), [Other Genetics and Genomics Commons](#), and the [Plant Pathology Commons](#)

### Recommended Citation

Joseph, Dani, "Sequence Analysis of the REN1 Genomic Region from the Grapevine 'Kishmish Vatkana'" (2019). *MSU Graduate Theses*. 3426.

<https://bearworks.missouristate.edu/theses/3426>

This article or document was made available through BearWorks, the institutional repository of Missouri State University. The work contained in it may be protected by copyright and require permission of the copyright holder for reuse or redistribution.

For more information, please contact [BearWorks@library.missouristate.edu](mailto:BearWorks@library.missouristate.edu).

**SEQUENCE ANALYSIS OF THE *REN1* GENOMIC REGION FROM THE  
GRAPEVINE ‘KISHMISH VATKANA’**

A Master’s Thesis

Presented to

The Graduate College of  
Missouri State University

In Partial Fulfillment

Of the Requirements for the Degree  
Master of Science, Biology

By

Dani Joseph

August 2019

© 2019, Dani Joseph

# SEQUENCE ANALYSIS OF THE *REN1* GENOMIC REGION FROM THE GRAPEVINE ‘KISHMISH VATKANA’

Biology

Missouri State University, August 2019

Master of Science

Dani Joseph

## ABSTRACT

The *REN1* region of the grapevine ‘Kishmish Vatkana’ was mapped as the locus that confers resistance to the economically important disease, grape powdery mildew. The purpose of this work was to extend the nucleotide sequence information of this region. By sequencing a heretofore unknown bacterial artificial chromosome clone, the sequence information of this region was extended by 46,890 nucleotides. Sequencing was performed using the third-generation sequencing method, named Oxford Nanopore Technology (ONT). In order to improve the accuracy of the sequence data, a modified ONT library preparation method was developed. ONT sequencing of a library prepared with the modified protocol generated a single 69,750 nucleotide-long contig of the entire BAC insert. Comparative analysis of the Sanger, Illumina, and ONT sequence data of the same stretch of DNA has revealed, however, that ONT sequence quality still lagged behind that of the former two. Several gene-finding programs were able to identify in the BAC insert three new genes encoding a TCP-type transcription factor, a cinnamoyl-CoA reductase enzyme and CC-NBS-LRR-type resistance protein. Orthologs of all three of these proteins have been implicated in plant disease resistance. The newly acquired sequence information also made it possible to determine that the BAC insert represented the resistance haplotype of the ‘Kishmish vatkana’ *REN1* region. Nonetheless, evidence that any of the three defense-related alleles discovered here contribute to the powdery mildew resistance phenotype will require more experimental work.

**KEYWORDS:** *REN1*, ‘Kishmish Vatkana’, plant disease resistance, Oxford Nanopore Technology, gene discovery, defense genes, haplotype

**SEQUENCE ANALYSIS OF THE *REN1* GENOMIC REGION FROM THE  
GRAPEVINE ‘KISHMISH VATKANA’**

By

Dani Joseph

A Master's Thesis  
Submitted to the Graduate College  
Of Missouri State University  
In Partial Fulfillment of the Requirements  
For the Degree of Master of Science, Biology

August 2019

Approved:

Laszlo G Kovacs, PhD, Thesis Committee Chair

Kyoungtae Kim, PhD, Committee Member

Lloyd Smith, PhD, Committee Member

Julie Masterson, PhD, Dean of the Graduate College

In the interest of academic freedom and the principle of free speech, approval of this thesis indicates the format is acceptable and meets the academic criteria for the discipline as determined by the faculty that constitute the thesis committee. The content and views expressed in this thesis are those of the student-scholar and are not endorsed by Missouri State University, its Graduate College, or its employees.

## ACKNOWLEDGEMENTS

First and foremost, I would like to thank Almighty God for giving me the strength, wisdom and guidance to complete this project.

I would like to deeply express my sincere gratitude to my thesis advisor, Dr. Laszlo G. Kovacs for his continuous support of this project, his motivation and patience. I am also obliged to acknowledge my thesis committee members, Dr. Kyoungtae Kim and Dr. Lloyd Smith for their encouragement and insightful comments to my thesis and research. I am indebted to be greatly thankful to Dr. Courtney Coleman for her insightful ideas and support throughout my research. I also greatly appreciate support from the Graduate College and the Biology Department at Missouri State University for making this research project possible.

I dedicate this thesis to my family. My uncle Dr. George Mathew, whose support, help and encouragement throughout the entire time directed me to the research field and to work on this project. I am deeply grateful to my best parents Celin Mathew, Jose Mathew and my loving sister Dr. Sani Joseph, who earnestly want me to be successful in my academics and research. I also thankfully remember my cousin brother Dr. Doyil T. Vengayil, who passed away during this thesis preparation time, helped me to embark on academic research.

## TABLE OF CONTENTS

OVERVIEW	Page 1
CHAPTER 1: DEVELOPING OF IMPROVED NANOPORE METHOD AND SEQUENCING OF THE REN1 REGION	Page 4
Introduction	Page 4
Materials and Methods	Page 8
Results	Page 14
Discussion	Page 22
References	Page 26
CHAPTER 2: SEQUENCE ANALYSIS AND HAPLOTYPING OF PART OF REN1 REGION	Page 29
Introduction	Page 29
Materials and Methods	Page 33
Results	Page 35
Discussion	Page 42
References	Page 46
SUMMARY	Page 51

## LIST OF TABLES

Table 1. Primers used for amplification of each BAC clone	Page 14
Table 2. Summary statistics on sequence data with original and the modified sequencing library preparation methods	Page 18
Table 3. Comparison of ONT-and-Illumina generated sequence data	Page 18
Table 4. Hypothetical coding regions and their coordinates in BAC81D11	Page 38
Table 5. Orthologue information of BAC81D11	Page 38
Table 6. Hypothetical proteins encoded by genes of BAC81D11	Page 39
Table 7. Gene ontology information for genes of BAC81D11	Page 39



## LIST OF FIGURES

Figure 1. Schematic diagram of the physical map of REN1 region	Page 19
Figure 2. Relationship between read length and average read quality of ONT generated data	Page 19
Figure 3. Relationship among phred quality, read length and number of reads of raw ONT sequence data generated	Page 20
Figure 4. Relationship between read length and number of reads of the untrimmed base called data generated	Page 20
Figure 5. Relationship between the number of bases sequenced and read length in data generated	Page 21
Figure 6. Cumulative sequencing yield as a function of read length in Gb in data generated	Page 21
Figure 7. Schematic diagram of the relative position of the Sanger-, Illumina- and ONT- sequenced DNA	Page 22
Figure 8. Alignment of a 304-nucleotide fragment of 81D11 with the corresponding haplotypes from ‘Thompson Seedless’ (TS-1 and TS-2)	Page 40
Figure 9. Map of the BAC81D11 insert showing the position of predicted genes and the retrotransposon fragment	Page 40
Figure 10. Alignment of transcription factor gene of BAC 81D11, TS-1, VT-1 and KV-1 haplotypes	Page 41
Figure 11. Alignment of the transcription factor amino acid sequence of TS-1 and 81D11	Page 41

## OVERVIEW

There is considerable grower and consumer interest in reducing the input of pesticides in grape production. The most promising approach to accomplish this is to cultivate disease-resistant grape varieties that have innate biological resistance against pathogens. Genes that confer resistance against pathogens are available in the vast genetic diversity of the *Vitis* genus and can be introgressed into cultivated varieties through traditional breeding or transgenesis. A number of sources for disease resistance genes have been explored from American and Asian *Vitis* species. One of these resources is the cultivated Central Asian *Vitis vinifera* variety ‘Kishmish Vatkana’ carrying a single resistance locus, named *REN1*, which confers resistance to the economically important pathogen powdery mildew (*Erysiphe necator*). While the *REN1* locus has been mapped to a 1.4 million-base pair region of chromosome 13 using a positional cloning approach, its complete nucleotide sequence has yet to be deciphered. The purpose of this work was to complete the nucleotide sequence information of this entire region. While this goal could not be fully met, the sequencing of a heretofore unknown bacterial artificial chromosome (BAC) clone was accomplished, which extended the sequence information of the *REN1* region by 46,890 nucleotides.

DNA sequence analysis was performed using the third-generation sequencing method Oxford Nanopore Technology (ONT). The first attempt to verify the sequence information of the entire known *REN1* region in a single reaction resulted in poor-quality data. To improve the quality of the data, the rest of the work focused on a single BAC clone, namely BAC81D11, which had not previously been sequenced. Furthermore, a few modifications were also made to the Oxford Nanopore Technology library preparation procedure to ensure the generation of the

longest possible reads by the sequencing reactions. Key among these modifications was the restriction of the BAC clone with an enzyme that cut its target molecule at a single locus. The modified library preparation protocol led to increased data yield, deeper sequence coverage, improved read quality, and extended read length. The reads were then assembled into a single 69,750 nucleotide-long contig, which represented the entire BAC insert. The accuracy of the sequence assembly was confirmed with BAC-end sequencing. Of this new clone, 22,860 nucleotides overlapped with another BAC clone previously sequenced, and the remaining 46,890 nucleotides represented new information. Stretches of DNA sequenced using ONT, Sanger and Illumina methods were aligned and compared. The comparison revealed that the quality of the ONT sequence was lower than the quality of either the Sanger or the Illumina sequence. Nonetheless, the ONT-generated sequence was contiguous and it could be used to subclone corresponding DNA haplotypes from ‘Kishmish Vatkana’ and its disease-resistant and susceptible parents ‘Vassagra Tschernaia’ and ‘Thompson Seedless’, respectively. Comparative analysis of these haplotypes confirmed that the DNA insert of BAC 81D11 represented the haplotype associated with powdery mildew resistance.

The ONT sequence also revealed that the BAC81D11 insert contained three complete open reading frames encoding functional proteins of a TCP-type transcription factor, a cinnamoyl-CoA reductase enzyme, and a coiled coil-nucleotide binding site-leucine rich repeat (CC-NBS-LRR)-type resistance gene. Interestingly, all three of these genes have been implicated in plant defense against pathogens. Orthologs of the TCP-type transcription factor gene have been shown to play a part of immune response in *Arabidopsis thaliana*, while orthologs of the cinnamoyl-CoA reductase 1 (CCR1) gene have been demonstrated to strengthen disease resistance in rice. A gene similar to the CC-NBS-LRR-type resistance gene is a key determinant of resistance against

downy mildew in *A. thaliana*. To determine if either of these grape genes present in BAC 81D11 play a part in powdery mildew resistance in 'Kishmish Vatkana' requires further experimentation.

The significance of this work is that it provides important tools for further research. It presents an advanced ONT library preparation protocol which can be used by other researchers working with BAC libraries. It also opens opportunities for hypothesis-driven experiments. The evidence that the BAC81D11 insert is of the resistance haplotype and the exact demarcation of the coding sequences enables other researchers to clone these genes, transfer them to powdery mildew-susceptible plants, and measure the resistance of the resulting transgenic plants.

The limitations of my data are that the ONT sequence of the BAC81D11 insert is less than 100% accurate and that the gaps between the previously assembled contigs remain to be unfilled. Furthermore, this work does not provide evidence about the contribution of the genes to powdery mildew resistance. Nonetheless, it provides tools that make hypothesis-driven experimentation possible.

# CHAPTER 1

## DEVELOPMENT OF AN IMPROVED OXFORD NANOPORE SEQUENCING METHOD

### Introduction

DNA sequencing technology determines the order of nucleotide sequences in a DNA strand. This technology has become essential to detect coding sequences in genomic DNA and cDNA and thereby has become an essential tool in modern biology. All sequencing procedures include three phases: sample preparation, sequencing reaction, and sequence assembly (Schadt et al. 2010). Based on the method used, productivity, throughput, and read length, sequencing techniques are categorized into first-generation, next-generation, and third-generation methods.

First-generation sequencing methods include the Maxam-Gilbert (also known as the chemical method) and the Sanger technologies. The former, developed in 1977, has fallen out of favor since the 1980s because of its technical complexity and reliance on toxic chemicals. Sanger sequencing (also known as the dideoxy sequence termination method) depends on *in vitro* DNA synthesis and the random incorporation of the 2'-3' dideoxy nucleotide, which terminates the activity of the DNA polymerase (Sanger et al. 1977). As DNA polymerase activity is primed with complimentary primers on all template molecules, all newly synthesized strands have the same 5' terminus. As a result, each fragment length is represented by a cohort of DNA strands, the synthesis of which is terminated at the same position. In order to detect these strands, the ddGTP, ddATP, ddTTP, and ddCTP dideoxynucleotides are covalently attached to different fluorescent dyes. The DNA strands are fractionated using capillary gel electrophoresis and the terminal ddNTPs are identified by their fluorescence in response to laser excitation. The

automation and multiplexing of this method facilitated the scaling-up of this technology to an industrial scale, which ushered in the genomic era of biology. Sanger sequencing was the method that enabled the generation of the first reference genome sequence of *Homo sapiens* and several model species. This method is still considered as the most accurate method of DNA sequencing (Heather and Chain 2016) and is commonly used today for small-scale sequencing. However, the Sanger sequencing technique has limitations, such as low sequence length, low throughput, and high cost. Most of these limitations stem from the fact that the Sanger method requires two separate steps: DNA synthesis and electrophoresis. For large-scale genome-sequencing projects, Sanger has been supplemented by less accurate, but higher-throughput next-generation methods.

Next-generation techniques were developed during the first decade of the 21<sup>st</sup> century. Solexa, Ion Torrent, and Pyrosequencing are some of the major next-generation techniques that have been developed and commercialized. Among them, Solexa emerged as the prominent high-throughput method after being commercialized and renamed Illumina sequencing-by-synthesis (SBS) method. Illumina sequencing starts with sample preparation which includes addition of adaptors to the ends of randomly generated template DNA fragments of an entire genome. An adaptor incorporates indices, priming sites and sequences complementary to the flow cell oligo nucleotides into a single DNA fragment. The major step after sample preparation includes the clonal amplification of template DNA which is performed through bridge amplification. Bridge amplification is repeated several times to exponentially increase the number of all the fragments. After clonal amplification, the reverse strands are washed away, and the forward strands are kept in the flow cell. Addition of the labelled nucleotides and binding of the complementary nucleotide fluorescent with dye is then repeated until the desired length of sequence is obtained. Each nucleotide has its own fluorescence, which is detected and recorded while each nucleotide

is added at the same time of synthesizing the complementary strand (Schadt et al. 2010). The latest Illumina sequencer, named NovaSeq produces three terabase of data in a single flow cell reaction (Costello et al. 2018). The Illumina method ensures low error rate, because it facilitates the sequencing of the same DNA fragment many times and produces data of great sequencing depth. This enables to identification of single nucleotide polymorphisms (SNPs) and short fragment variation at a population scale (Lui et al. 2012). However, the read length obtained from Illumina is still much smaller than that of Sanger method (Metzker 2010, Whiteford et al. 2009, Schatz et al. 2010). The low read length makes the assembly process difficult, but this difficulty can be overcome by aligning large numbers of reads to longer contigs. In highly repetitive genomic regions, however, contigs assembled in this way can lead to ambiguity, because it is impossible to determine the number of times a sequence is repeated in such regions (Alkan et al. 2011). High-throughput sequencing also creates challenges in the storage of the resulting immensely large data files (Schatz et al. 2010). The eukaryotic genome has highly complex repetitive sequences and these sequences constitute about half of the genome. These repetitive sequences require long reads for sequence alignment and assembly to generate a meaningful contigs to be further analyzed (Treangen et al. 2013).

Third-generation sequencing technology (TGS) has been developed to improve read length as compared with Illumina. Unlike NGS technologies, TGS technologies produce sequences of an average of 10,000 nucleotide-long reads (Schatz et al. 2016). There are two main TGS techniques available. The SBS method, involves a single molecule of DNA polymerase being detected as it incorporates each nucleotide into the growing DNA chain. This is also referred to as Single-Molecule Real-Time Sequencing (SMRT) and is applied by Pac-Bio sequencing technology (Levene et al. 2003). The average read length generated per run by

PacBio is  $1.0\text{--}1.5 \times 10^4$  nucleotides and the single pass error rate of this method is 13%. The number of reads generated per run is  $3.5\text{--}7.5 \times 10^4$  (Rhoads et al. 2015). Oxford Nanopore Technology (ONT) sequencing is based on the threading of prepared DNA through a nanopore that is fixed on a flow cell. ONT is simpler compared with other sequencing techniques (Loman et al. 2012).

The ONT nanopore is either a biological or artificial pore protein with a 10-Å opening, fixed on an insulating membrane in the MinION flow cell. The membrane is submerged in saline solution with an electric gradient across it to maintain a positive charge at the internal pore region which attracts negatively charged DNA strands. The stretch of DNA traversing the pore is five nucleotides long. In this way, 5-mers are recognized as a single “event” and their electric current variation during their passage is detected as a signature current for each 5-mer combination. This signature current is then “translated” into nucleotide sequences by the software named MinKNOW (de Lannoy et al. 2017).

The average read length produced by ONT is  $2\text{--}5 \times 10^3$  nucleotides. ONT sequencing is real time and highly cost-effective, does not require PCR, produces long reads, and it can be performed in a small portable device. Along with these remarkable advantages, ONT has a major disadvantage, which is high error rate (Rhoads et al. 2015, de Lannoy et al. 2017).

This study involves sequencing of BAC clones from the *REN1* region of the powdery mildew (PM) disease-resistant grapevine ‘Kishmish Vatkana’ (KV) using ONT hardware and software. The study also includes a comparison of Sanger, Illumina and ONT sequences of a 2.8 kb region from the *REN1* locus. The grapevine KV is a seedless *V. vinifera* table grape variety from Uzbekistan (Hoffman et al. 2008). The genome of this variety contains a resistance locus named *Resistance to Erysiphe necator 1 (REN1)*, which contains one or more resistance gene(s)



against the fungal disease powdery mildew, which is caused by the biotrophic fungus *Erysiphe necator* (Hoffman et al. 2008). The ultimate goal of studying KV and other disease-resistant grape genotypes is to introgress the gene(s) responsible for resistance into grapevines for the development of new grape cultivars, the cultivation of which will require a reduced input of fungicides. The *REN1* region in KV has been delimited to a 1.4 cM (1.4 mega base pairs) region using simple sequence repeat (SSR) marker-based linkage mapping (Coleman et al. 2009). The presence of nucleotide binding site leucine-rich repeat (NBS-LRR) genes in the *REN1* region of KV is identified within a genetic interval flanked by SSR markers (Coleman et al. 2009). These genes are key candidates for encoding the biological information responsible for PM resistance. An attempt had been previously made to sequence the *REN1* region (Coleman 2016). This attempt was only partially successful, because the sequenced DNA fragments failed to cover the entire *REN1* region. As a result, two DNA stretches of unknown length in the physical map remained unsequenced (Figure 1). The purpose of this current work was to fill in those two gaps with BAC clones and sequence information.

## **Materials and Methods**

**Validation of KV BAC colonies.** DNA fragments of the KV *REN1* region from chromosome 13 were obtained from eleven BAC library clones of KV genomic DNA constructed by Lucigen in 2008. These BAC clones were previously identified as positive for *REN1*-specific SSR markers (Coleman 2016), but their validation was necessary before subjecting them to sequencing. To confirm that the BACs contained the predicted insert, a colony PCR was performed using insert-specific SSR primers designed by Coleman et al. (2009). The primer information is shown in Table 1. The BAC clones, namely 81D11, 75I23,

77I15, 67E16, 88C13, 24G05, 70KO3, 41N11, 07D12, 39C19, and 50J14, were stored as glycerol stocks in a -80°C freezer. To perform colony PCR, selected BAC colonies from the 384-well plate were streaked to a nutrient agar plate made with Terrific Broth (TB) medium (Sambrook et al. 2006), which was supplemented with 12.5 µg/mL chloramphenicol and 0.01% L-arabinose. Following a 12-hour growth at 37°C, fast DNA extraction specifically designed for colony PCR was performed. Brief DNA extraction was performed by swiping an isolated colony with a pipette tip and transferring the bacteria into 100 µL deionized nuclease-free water in a 1.5-mL Eppendorf tube. The bacteria were lysed at 94-99°C for 10 min, immediately transferred to ice for 2 min and centrifuged for 5 min at maximum speed. Subsequently, 2 µL of the supernatant containing DNA was used as template for PCR. Resistance haplotype-specific SSR primers were used in the PCR reactions with a total reaction volume of 10 µL (0.1 U enzyme, 1X buffer with KCl, 0.2 µM of each primer, 2.5 mM MgCl<sub>2</sub>, 0.5 mM each dNTP, and about 4 ng bacterial DNA). A 25-µL PCR recipe was also used, which differed from the previous protocol in the use of 0.2 U of Taq polymerase enzyme per reaction. The touch-down (TD) PCR protocol was used to minimize non-specific primer annealing and amplification. The initial denaturation cycle was 94°C for 2 min, which was followed by 10 cycles of denaturation at 94°C for 30 sec, annealing at 62°C for 30 sec with 1°C decrease at each cycle at each cycle, then primer extension at 72°C for 45 sec. This was then followed by 28 cycles of denaturation at 94°C for 30 sec, annealing at the optimal annealing temperature for 30 sec, then primer extension at 72°C for 45 sec. The final primer extension temperature was 72°C for 15 min.

**BAC DNA Extraction and Purification.** BAC clones that were confirmed as part of the *REN1* resistant haplotype using colony PCR were selected for DNA extraction and ONT sequencing. First, the seven BAC clones validated with the 10 µL PCR recipe were selected for

DNA extraction and sequencing. To extract BAC DNA, all the seven BACs were inoculated in TB broth (Ledent et al. 1993) and grown for 12 hours in a 37°C incubator shaker. BAC DNA extraction from the selected seven BAC cultures was performed using the FosmidMax DNA Purification kit (Epicentre, USA, Chicago, IL, USA), following the manufacturer's protocol. The extracted DNA was precipitated with ammonium acetate (2M final concentration) and 100% ethanol was added to the solution at a 2:1 ratio. This solution was mixed and kept at -20°C overnight for DNA precipitation and then incubated at -80°C for 30 minutes, with subsequent centrifugation at 4°C for 20 min. The precipitated DNA samples were washed using 500µL of freshly prepared 70% ethanol and air-dried overnight. The DNA pellet was resuspended in TE buffer and the concentration of the resulting DNA solution was measured using a Qubit Fluorometer (Thermo Fisher Scientific, Waltham, MA).

**Library Preparation and ONT Sequencing.** Selected BAC DNA samples were prepared for sequencing using the ONT library preparation protocol 1D Genomic DNA-by-Ligation (protocol SQK-LSK108) following manufacturer's guidelines (ONT, Oxford, England). The 1D Genomic DNA-by-Ligation method was designed for the generation of ultra-long sequence reads. The beginning of the library preparation protocol provided by ONT involves a mechanical shearing step, which breaks the sample DNA into fragments. The DNA fragments are then end-prepared by adding adaptors. The resulting library was then purified using XP beads and loaded on a flow cell in an ONT MinION portable sequencer. Analysis of the ONT data was performed using ONT software. Sequencing was allowed to proceed for 48 hours on the flow cell.

**Sequence Analysis.** The ONT Fast5 results were downloaded automatically and were transferred to FastQ format for downstream processing of the raw sequences using the python

software Poretools (Loman and Quinlan, 2014). It is downloaded from <https://github.com/arg5x/poretools>. The FastQ format sequences were then converted to FASTA format using a sequence converter available online at <http://sequenceconversion.bugaco.com/converter/biology/sequences/>. The FastA-formatted raw sequences were then assembled using the *de novo* genome assembler software Canu according to the guidelines provided by Koren et al. (2017). Low-quality and questionable sequences from the raw reads were trimmed and mapped by Canu before sequence assembly. The contigs produced by Canu were analyzed using quality control (QC) tools such as Nanoplot and Pauvre (de Coster et al. 2018). Nanoplot and Pauvre (github tools) are available at <https://github.com/wdecoster/NanoPlot> and <https://github.com/conchoecia/pauvre>, respectively.

**Improved Protocol for Sequencing.** In the improved library construction method, only a single BAC clone, namely BAC81D11, was used. This BAC, 81D11, was specifically selected because it had not been previously sequenced. BAC DNA was extracted using the FosmidMax DNA Purification kit as described above. The BAC plasmid DNA was linearized using the enzyme *FseI*, which cuts BAC81D11 at only a single site. *FseI* was purchased from New England Biolabs (Ipswich, MA, USA). The linearized DNA was purified from cellular debris and proteins using phenol/chloroform extraction. In this procedure, equal amounts of phenol/chloroform were added to the linearized DNA at a 1:1 ratio, mixed thoroughly, then centrifuged. The aqueous supernatant was transferred to a clean Eppendorf tube, and then an equal amount of chloroform was added to further purify the sample. The supernatant containing the DNA was then transferred to a fresh Eppendorf tube. The purified DNA was then precipitated and washed as described above. In this library preparation, I avoided the mechanical

shearing step to avoid breaking the DNA into fragments smaller than the full-length BAC plasmid. All remaining steps of the library preparation were performed using 1D Genomic DNA by ligation with the SQK-LSK108 kit according to the manufacturer's protocol for the flow cell FAH11697 (ONT, Oxford, England). The raw sequences were provided in FastQ format by the MinKNOW software. The assembly of the raw sequences into a long contig was performed in the same manner as described above.

**Vector Sequence Removal.** The vector sequences in the Canu-assembled contig were identified by aligning the contig to all the vector sequences present in the NCBI GenBank database using the Vecscreen tool available at <https://www.ncbi.nlm.nih.gov/tools/vecscreen/>. The sequence repeats and their coordinates in the Canu assembled contig were identified by aligning the contig to itself using the BLAST tool, available at <https://blast.ncbi.nlm.nih.gov/Blast.cgi> (Altschul et al. 1990). The vector sequences and sequence repeats were manually removed from the Canu-assembled contig using the functions within the CLC Genomics Workbench software.

**Validation of BAC 81D11 ONT Sequence.** To amplify the terminal regions of the 81D11 BAC DNA, primers specific to the pSMART vector sequences flanking the cloning site were used. The PCR product from the right SLR4 primer sequence (TTGACCATGTTGGTATGATTT) and left SL1 primer sequence (CAGTCCAGTTACGCTGGAGTC) were used to PCR amplify approximately 1000 nucleotides from each end of the 81D11 DNA sample and sequenced using the Sanger technique. BAC-end sequences from both the right and left ends were aligned to the 81D11 ONT sequence using the BLAST tool to compare the nucleotide sequences between the ONT and the Sanger sequences and to test the accuracy of the ONT sequence and the Canu assembly. Alignment of BAC-end

sequences to the reference grapevine genome sequence was performed to find the coordinates of the corresponding sequence of 81D11 in the reference 12X genome *V. vinifera* PN40024, which is available at

[https://www.ncbi.nlm.nih.gov/genome/gdv/browser/?acc=GCF\\_000003745.3&context=genome](https://www.ncbi.nlm.nih.gov/genome/gdv/browser/?acc=GCF_000003745.3&context=genome)

(Jaillon et al. 2007). To validate the 81D11 ONT sequence again, the Illumina contigs of the BAC clone 71H20, which had been predicted to overlap with BAC 81D11 (Coleman 2016), were aligned with the 81D11 ONT sequence using the BLAST tool. To further validate the coordinates of BAC 71H20 in the reference genome sequence of PN40024, the 71H20 BAC-end Sanger sequences were aligned with the *V. vinifera* reference PN40024 sequence. The nucleotide sequences produced using both the ONT protocol (provided by the manufacturer) and the altered new protocol were compared by using the sequencing statistics generated by Nanoplot and Pauvre software.

#### **Cloning and Sanger Sequencing of the 2,777 nucleotide-Long Region from 81D11. A**

2,777-nucleotide region from the 81D11 ONT sequence was amplified using the primers

5'AAAAAATTAAAATTAAATTTTAAGAAATC3' and

5'CCTTCATATTTTTTTTTTGAATATA3'. The amplified fragment was ligated to the pMiniT

vector and transformed to *Escherichia coli* 10 beta cells using the NEB PCR Cloning Kit (NEB

Bio Labs, USA). The Sanger sequence of this cloned fragment was generated using a set of

additional internal forward primers such as 5'GGTCTTCGCTTTTCATTACCT 3',

5'AAATCTTGAGTTACTATGCCACA3', and 5'CCTAGGGTTTCGACCACAAA3', and

reverse internal primers including 5'AGTTGAATCGCACATTGCTCT3',

5'TCATTTGGTCATTTTAAAGGAGA3' and

5'TGAATGCAAAAGAAAATTAATAGTGG3'. The Sanger sequence is then compared to the

ONT sequence of 81D11 using a BLAST pairwise alignment. This sequence was also compared with the overlapping Illumina sequence contigs of the BAC clone 71H20.

**Table 1** Primers used for amplification of each BAC clone.

BAC	Forward primer Sequence	Reverse primer Sequence
81D11	CTTGGCTAGATAGTGCCTTCA	GAAAATCAAAGGGATAAAGGGTC
75I23	GAAAATCAAAGGGATAAAGGGTC	AAGGATTTGAATGATATCTAATATAGG
77I15	TTCGAGTTTTGTAGATCTATTTTTGG	CCAATCCTATCAACTTGTTCAATG
88C13	TCTACAGCGTCGGCTAGGTT	ATCATCATGTGCACGTCTCC
67E16	TCAACAGTGGCATTAAAAATGG	CGAGCAAATCAGTGGAAGC
24G05	TCAACAGTGGCATTAAAAATGG	CGAGCAAATCAGTGGAAGC
70K03	TTTAAGGGGCAGTGCCTAACT	TTTCTCCTGACCCCGAATG
41N11	TTTAAGGGGCAGTGCCTAACT	TTTCTCCTGACCCCGAATG
07D12	TCATTGCGTGGAAAATTTGTAT	AAACCTGGAATAATCTAGT
39C19	TGCAAATATTATGGTTGGTTTG	CATTGTACCTTGGCCACATT
50J14	GAAAAATCTCTTCATAGTTTTGATTGG	GTTGGTTTGTGTGTACTTTAATTT

## Results

**Rationale.** The purpose of this work is to close the gaps between contigs in the physical map of the *RENI* region in the grapevine KV (Fig. 1). Using Illumina sequencing, it has been established that the *RENI* region contains repeated elements (Coleman et al. 2009). Because the assembly of DNA regions with repeated elements is prone to assembly errors, it was also important to validate the correctness of such regions. Third-generation sequencing is able

produce long reads and is therefore well suited to resolve ambiguities in the nucleotide sequence of repeated elements. Resequencing the BAC clones around gaps 1 and 2 will not only help connect contigs 1, 2, and 3 (Figure 1), but may also improve the accuracy of the contig sequences. In this context, sequencing clone BAC81D11 is central to filling gap 2, because this BAC clone has been putatively localized to the right terminus of contig 1, but its nucleotide sequence has not been determined (Figure 1).

**Validation of KV BAC clones and ONT sequencing.** BAC clones named BAC75I23, BAC77I15, BAC67E16, BAC88C13, BAC24G05, BAC70KO3, BAC41N11, BAC07O2, BAC39C19, and BAC50J14 were validated with polymorphic SSR markers specific to the resistance haplotype, and pBAC81D11 was validated with a non-polymorphic *REN1* specific SSR marker. DNA purified from pBAC77I15, pBAC67E16, pBAC24G05, pBAC70KO3, pBAC41N11, pBAC07O2, and pBAC39C19 were combined and sequenced using the ONT protocol. The sequenced long reads were then assembled into five non-overlapping contigs of lengths 85,751, 8,882, 27,406, 29,498, and 19627 nucleotides long, respectively. Summary statistics of the ONT sequences are listed in Table 2.

**Development of an improved sequencing protocol to sequence BAC81D11.** The extracted BAC81D11 DNA was linearized, phenol/chloroform-purified and sequenced using an altered ONT protocol, as detailed in Materials and Methods. The raw sequences were assembled into a 116,607 nucleotide-long single contig using the software Canu. An examination of the resulting contig sequence revealed that a long stretch of DNA was present twice, which was likely an artifact due to incorporating vector sequences into the contig twice. To correct this problem, the vector sequences from coordinates 8,315 to 45,857 and 115,609 to 116,605 and sequence repeats from coordinates 1 to 45,858 and 115,609 to 116,607 from the 116,607-



nucleotide-long 81D11 contig were removed from the 116,607 nucleotide-long sequence. This resulted in a contig of 69,750 nucleotides, which represented the insert of BAC81D11.

Sequencing quality indices and statistics clearly indicated that the improved library preparation protocol was more efficient than the original ONT protocol (Table 2). To validate the correctness of the BAC insert, I used BAC-end primers to amplify and Sanger-sequence the terminal stretches of BAC81D11. The Sanger data were 747 and 941 nucleotides long on the right and left ends of BAC insert, respectively, and were in agreement with the ONT data. The region that corresponds to the 81D11 sequence in the *V. vinifera* PN40024 reference genome sequence (Jaillon et al. 2007) spans from coordinate 16,836,214 to 16,921,940 on chromosome 13. When aligned with BAC 71H20 (Coleman 2016), the BAC81D11 overlapped from coordinates 1 to 22,860 with 7 different Illumina contigs of the 71H20 insert. The BAC71H20 Illumina contig overlap information with BAC81D11 is listed in Table 3.

**Comparison of data from traditional and modified ONT sequencing protocols.** A bivariate graph showing sequencing quality plotted against read length is depicted in Figure 2. Read quality was filtered, and read length, phred quality, and the number of read data within the specified range and within a cut-off read quality value of 5 are shown in Figure 3. The extended length of most reads generated with the modified protocol is confirmed by Figure 4, which shows the number of reads plotted against read length. This is further corroborated by data in Figure 5, which is a histogram of total number of bases sequenced weighted by the number of nucleotides per read-length bin. This histogram clearly demonstrates that a substantially larger quantity of sequence data was generated in the improved protocol than in the original ONT protocol and that more of the data were generated as longer reads in the improved protocol (Figure 5). Figure 6 depicts a graph in which the cumulative yield is plotted against log-

transformed read length, which demonstrates that a substantially greater quantity of the sequence data was generated by reads in the 10 to 40 kb length by the improved protocol than by the original ONT protocol, further substantiating the higher efficiency of the modified protocol than the original ONT protocol.

**Cloning and Sanger sequencing of a 2,777 nucleotide-long region from 81D11.** The comparison of 2.777 kb ONT fragment from BAC 81D11, and its corresponding Sanger sequence (ONT coordinates 14,339-17,115) generated one mismatch and 31 indels (insertion or deletion) after aligning both sequences using BLAST tool. The Sanger sequence length was 2,807 for the 2,777-nucleotide long ONT sequence. The corresponding Illumina-sequence contigs from the overlapping BAC 71H2O was aligned with the Sanger sequence and produced one indel, one insertion in Illumina contig (Figure 7). These alignments clearly demonstrate that the improved ONT protocol was able to generate a continuous sequence in a region that could only be captured by two non-overlapping contigs by the Illumina method. Sequencing accuracy is more difficult to compare, because of the gap in the Illumina contigs. Nonetheless, the alignments reveal that, aside from the gap between the Illumina contigs, the Illumina and the Sanger sequences differ only at one indel, while the ONT and Sanger sequence differs at as many as 31 indels. The 2,807-nucleotide sequence is correctly assembled by the Canu software as demonstrated by the continuity of the corresponding Sanger sequence. The entire Illumina contigs which have an overlap with the 81D11 region is also aligned along with the 81D11 ONT sequence. The alignment produced 7 contigs with 5 gaps. Among 7 Illumina contigs, 5 of them were non overlapping contigs. The total number of nucleotides present in the gaps produced by the Illumina contigs are 982-nucleotides according to the 81D11 ONT sequence.

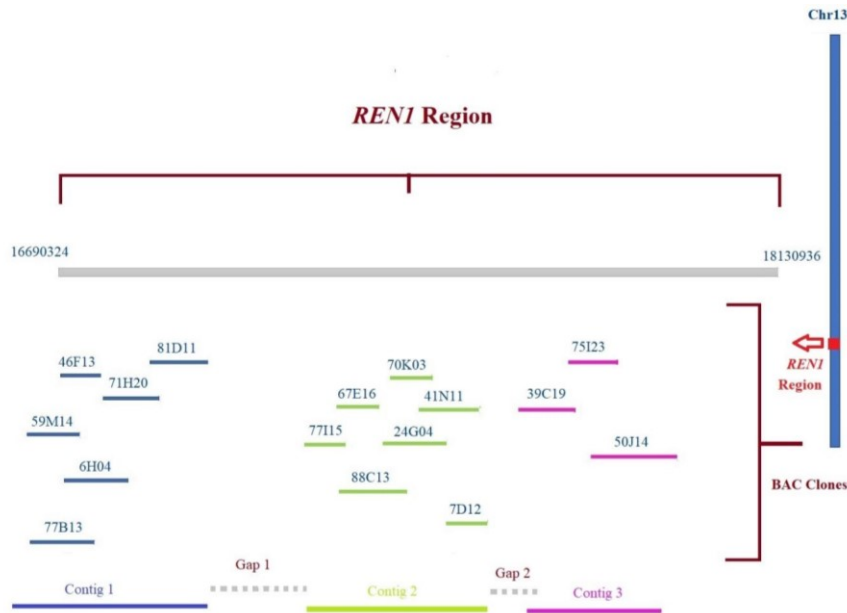
**Table 2** Summary statistics on sequence data with ONT and modified library preparation.

Parameters	ONT Method	Modified Method
Total Number of Nucleotides Sequenced	53,232,565 nt	1,373,305,247 nt
Mean Read Length in Nucleotides	3,326.2 nt	4,986 nt
Total number of raw sequences generated	16,004	275,433
Mean Read Length in Nucleotides	3,326.2 nt	4,985.99 nt
Median Read Length in Nucleotides	2,257.5 nt	2,120 nt
Median Read Length ( $\geq 1000$ nt)	4,046 nt	6,287.5 nt
Mean read quality	4.8	7.3
Median Read Quality	4.6	6.4
N50 Value	4,997.04 nt	12,895 nt
L50 Value	2,873	29,022
Sequencing Depth	2.93X	17818.8X

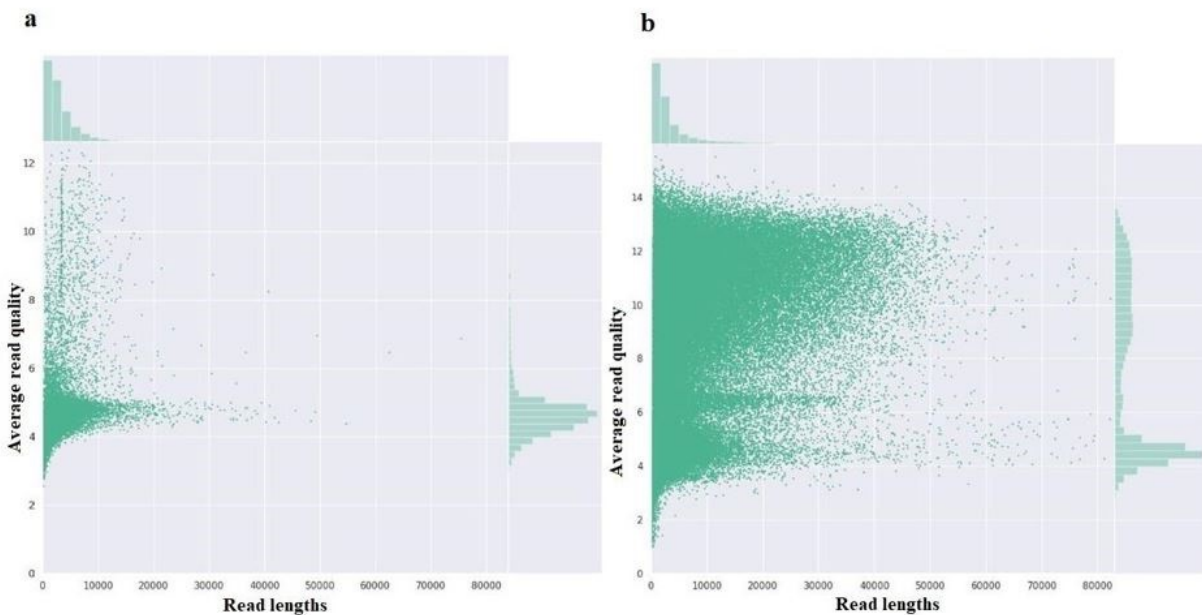
**Table 3** Comparison of ONT- and Illumina-generated sequence data.

71H20 Contig Number	Coordinates*	Percentage Identity
Contig-14	1 – 7,148	98%
Contig-22	7,154-7,903	99%
Contig-12	8,730-10,381	98%
Contig-26	10,763-12,186	98%
Contig-09	12,184-15,160	99%
Contig-03	15,228-17,111	99%
Contig-02	17,152-22,860	99%

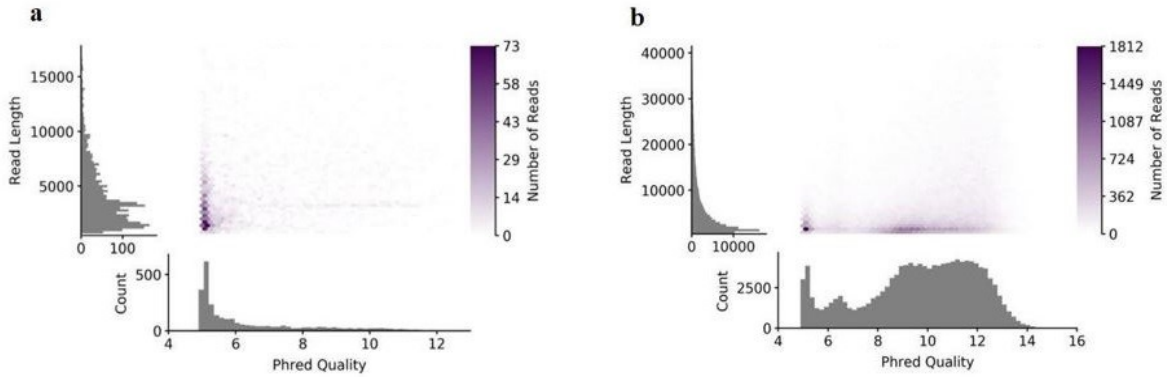
\* Coordinates based on the 81D11 sequence.



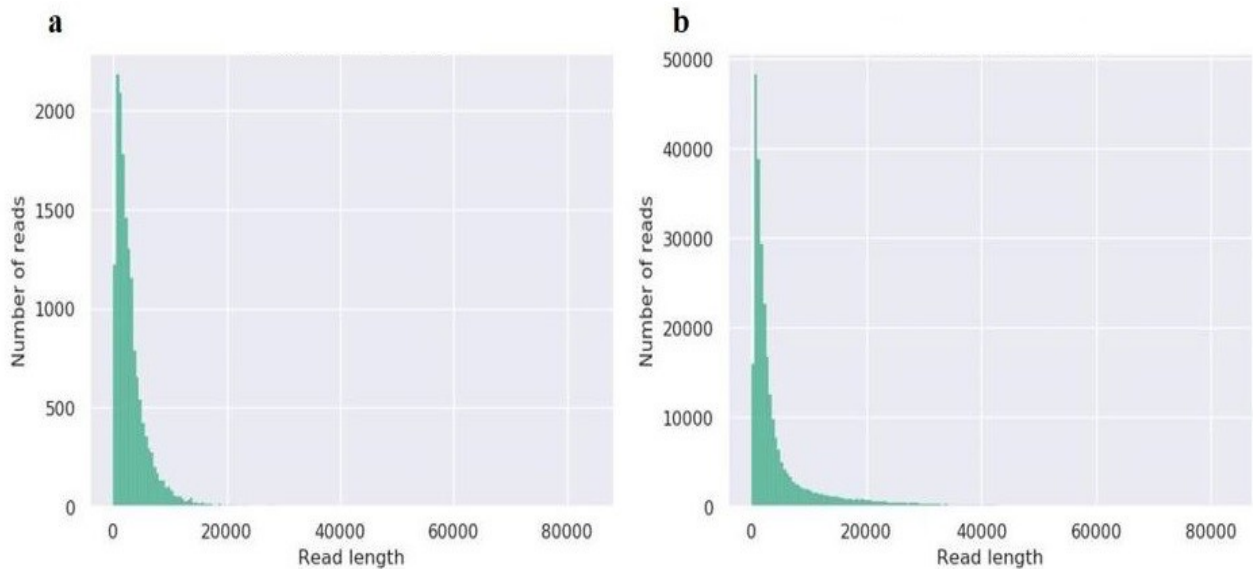
**Figure 1** Schematic diagram of the physical map of *REN1* region. The vertical bar on the right represents the entire chromosome 13 with the red color representing the approximate position of the *REN1* region. The *REN1* region is enlarged in the horizontal bar with coordinates 16,690,324 to 18,130,936. Each colored bar under the enlarged *REN1* region represents each contig. Map is not drawn to scale.



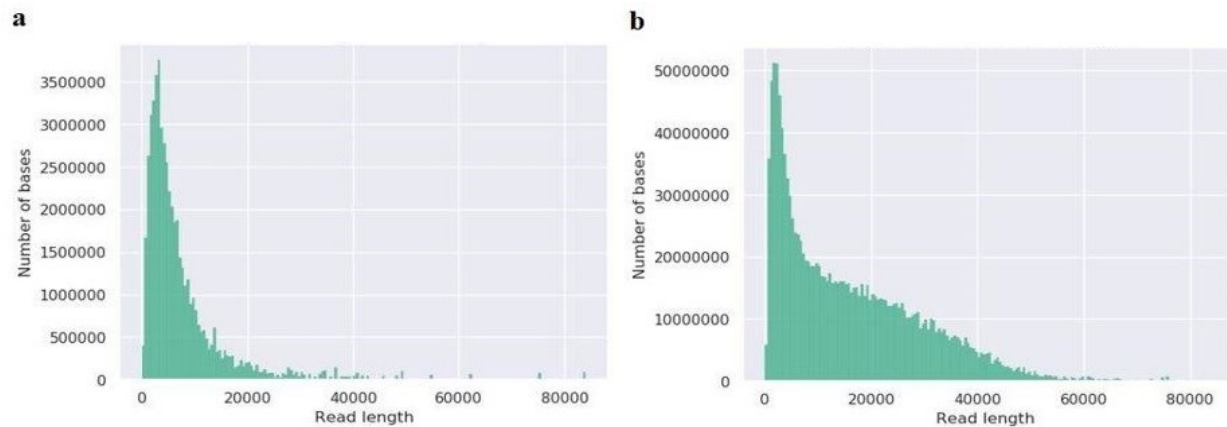
**Figure 2** Relationship between read length and average read quality of ONT sequence data generated. (a) the original and with (b) the modified library preparation protocols. Read length is plotted on X axis and average read quality is plotted on Y axis.



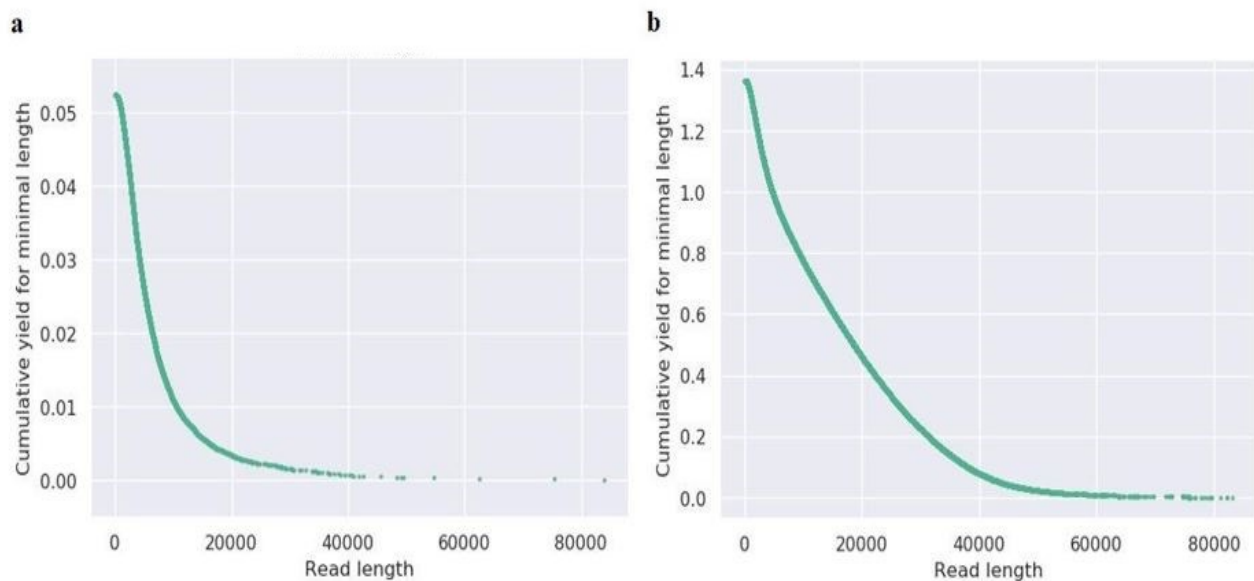
**Figure 3** Relationship among phred quality, read length and number of reads of raw ONT sequence data generated. The graphs depicted the three different variables. (a) the original ONT protocol data and with (b) the modified library preparation protocol on BAC 81D11. The reads were filtered by read quality score of 5.



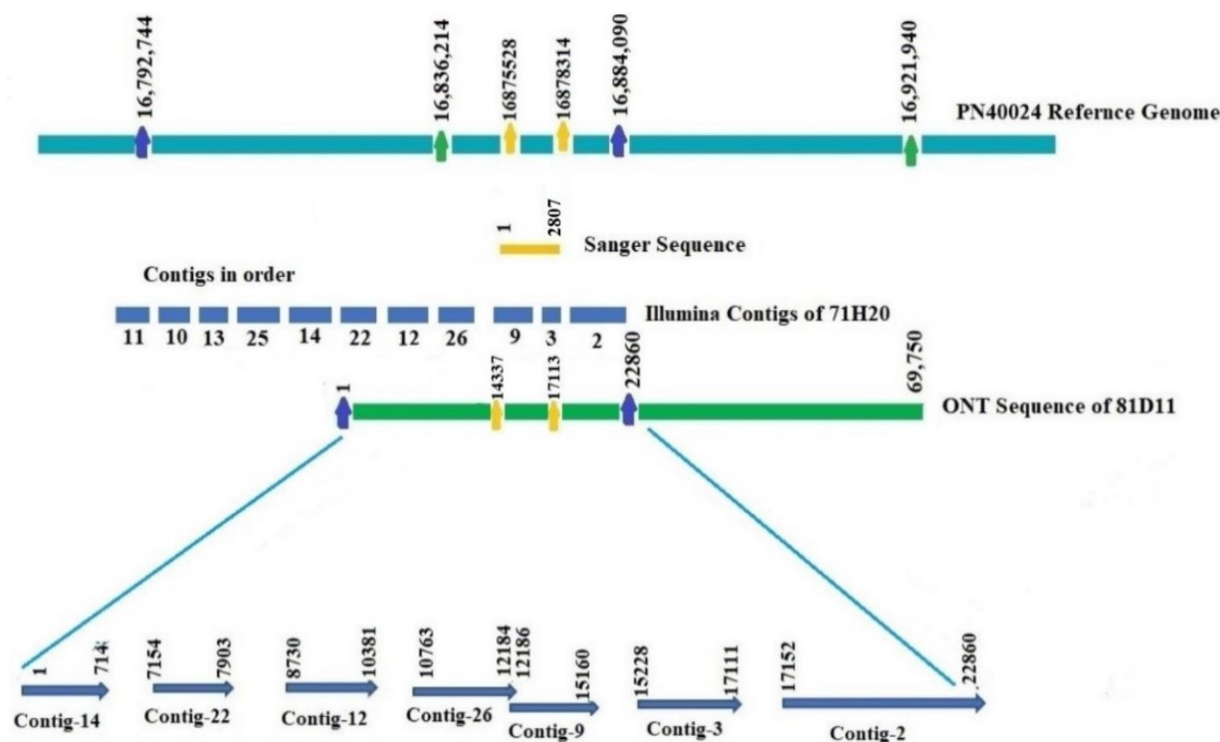
**Figure 4** Relationship between read length and number of reads of the untrimmed base called data generated. (a) the original ONT protocol data and (b) with the modified library preparation protocol on BAC 81D11. Read length is plotted on X axis and number of reads is plotted on Y axis.



**Figure 5** Relationship between the number of bases sequenced and read length in data generated. (a) the original and (b) with the modified library preparation protocols. Each bar represents the number of bases plotted in read-length bins of 500-nucleotide increments.



**Figure 6** Cumulative sequencing yield as a function of read length in Gb in data generated. (a) the original and with (b) the modified library preparation protocols. Read length is plotted on X axis and cumulative yield per minimal length is plotted on Y axis.



**Figure 7** Schematic diagram of the relative position of the Sanger-, Illumina- and ONT-sequenced DNA. Numbers refer to the corresponding chromosome-13 coordinates in the reference genome of PN40024. Gold, blue and green arrowheads represent Sanger, Illumina and ONT sequence coordinates, respectively.

## Discussion

In plant genomes, resistance (R) genes are commonly located in regions of highly repetitive DNA, which is also a characteristic of the *REN1* region. Therefore, successful sequencing and assembly of this complex region demand long reads and are generated by third-generation sequencing techniques. In this thesis project, I used an ONT MinION sequencer to generate long reads of the *REN1* region. To overcome the problem of the high error rate associated with ONT technology, I performed deep sequencing of the region. Sanger sequencing was performed of the 2.8 kb sub-region of BAC81D11 to assess the accuracy of the ONT-generated data. I also made a comparison between Sanger and ONT sequences. In my first attempt, I ONT-sequenced seven BACs from both contigs 2 and 3 of the *REN1* region (Fig. 1),

but this resulted in a shallow sequencing depth. To increase depth, I followed up with an alternative method and applied it to the sequencing of a single BAC clone, BAC81D11. This second attempt has led to a substantially deeper sequence and improved data quality.

Sequence assembly was performed using an assembler tool, named Canu, which was developed by Koren et al. (2016). Canu does correction, trimming and assembling of raw ONT base-called reads generated by the MinKNOW software. Canu is able to assemble entire genomes, requires relatively low sequence coverage (~10X), and it functions faster than any other assemblers. It takes only ~20,000 CPU hours to assemble a human genome with this software (Koren et al. 2016). However, my first sequencing attempt of seven BACs from both contigs 2 and 3 of the *RENI* region produced sequences of low quality with a sequence depth of 2X. The read lengths were short and the number of reads were low (Table 2), therefore, the sequence assembly did not produce overlapping contigs. To improve sequence quality, I developed a new ONT sequencing library preparation method. The original idea to improve raw sequence length was obtained from Jain et al. (2018), who successfully sequenced the centromeric region of the human Y chromosome with increased length for each read. They accomplished this by linearizing the BAC clones using a transposase enzyme, which cut into a clone at a single location. Sequencing these full-length linearized clones enabled them to solve the structure of the highly repetitive centromeric region (Jain et al. 2018). This idea prompted me to eliminate the mechanical shearing step and replace it with the linearization of BAC81D11 with a restriction enzyme that cuts the clone at a single location.

The BAC clone 81D11 had not been fully sequenced previously with the Illumina method, therefore, it served not only as a test DNA for the new method, but it generated novel information in the *RENI* region. The restriction enzyme that cut BAC81D11 at a single site was



*FseI*, and its restriction site was identified by a fellow lab member Dr. Courtney Coleman. To ensure that the *FseI* enzyme was completely eliminated following the restriction reaction, the linearized BAC DNA was subjected to a phenol/chloroform purification. This may also have increased the quality of the DNA and consequently of the read data. This alternative method dramatically increased the number of raw reads, the read length (Figure 2), the sequencing depth, the number of nucleotides sequenced, the read quality (Figure 2), and the N50 and L50 indices. The median read length was likely decreased in the modified method because of the large number of short (< 10,000 bases) outliers. However, removing these short outliers from the dataset considerably increased the median read length (Table 2). The raw reads were assembled into a single contig of 69,750 nucleotides.

While comparing both old and modified library preparation methods, the total number of nucleotides sequenced is 2.58 times higher with the modified method than with the ONT-provided protocol. The total number of reads produced is 17.2 times higher in the new method.

While extended overlaps of raw reads certainly improved sequence accuracy, a comparison between ONT- and Sanger-sequenced data reveals that ONT still produces data with considerable inaccuracy. Firstly, there are 31 indels between ONT and Sanger sequences of the same DNA fragment. The ONT method produced a 2,777-nucleotide long sequence, but Sanger produced a 2,807-nucleotide long sequence (Figure 7). These errors are attributed to the indel-specific events during nanopore sequencing. The multiple deletions and the single insertion in the ONT sequence relative to the Sanger sequence are explained as “skips” and “stay”, respectively. A skip occurs when a k-mer, of length 5 nucleotides DNA stretch, is erroneously read as only a single nucleotide. A stay is due to reading the same k-mer twice by falsely splitting an event into two events (de Lannoy et al. 2017). Skips and stays are common in the

ONT sequence because of the occasional aberrant functioning the motor protein and noise in the electrical signal due to voltage change. Currently, there is no reliable computational method to correct skips and stays.

While ONT managed to generate a sequence that covers the entire BAC81D11 insert, and is therefore superior at sequencing complex genomic regions, it still has a serious disadvantage, which is high cost. To obtain the sequence quality reported for the second ONT attempt, I had to use one MinION flowcell for a single BAC clone. The cost of one MinION flowcell is currently \$900, which makes the sequencing all 13 BAC clones prohibitively expensive for small laboratories. This is the reason why I was unable to sequence all of the BAC clones and fill gaps 1 and 2 in the *REN1* region physical map.

Aligning of the newly obtained sequence of BAC81D11 with existing BAC sequencing revealed that BAC81D11 overlapped 22,860 nucleotides with BAC71H20, a BAC clone that has been previously sequenced using Illumina technology. Fortuitously, the stretch of DNA corresponding to the 2,807-nucleotide long Sanger-sequenced region is part of this overlap. Therefore, I had Sanger, Illumina, and ONT data for this 2,777-nucleotide-long stretch of DNA, and this allowed me to compare and draw conclusions about the accuracy of all three sequencing methods. In these comparisons, I considered the Sanger sequence as the standard reference, as it is the method producing reads of the highest accuracy. Within the 2,777-nucleotide-long stretch of DNA shared by all three data types, Illumina has an insertion (an extra T) relative to Sanger data. This extra T nucleotide is likely be due to the homopolymer related error, which often occurs in the Illumina sequencing (Konstantinidis et al. 2012). Further supporting the erroneous nature of this insertion is that the ONT and the Sanger data agree in having only nine T nucleotides at this location. Moreover, this 'T' falls at the end of an Illumina read where there is

a higher probability of producing errors (Hillier et al. 2008, Van Tassel et al. 2008, Dohm et al. 2008). Additional information that sheds light on the Illumina read quality comes from my observation that the overlap in Illumina sequences does not perfectly correspond to the 22,860 nucleotides of the ONT sequence but fall into two non-overlapping contigs with a 460-nucleotide long gap between the contigs.

## References

- Alkan C, Coe BP and Eichler EE. 2011. Genome structural variation discovery and genotyping. *Nat Rev Genet* 12:363–376.
- Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* 215:403-410.
- Ari S and Arikan M. 2016. Next-Generation Sequencing: Advantages, Disadvantages, and Future. *In Plant Omics: Trends and Applications* (1), pp 109-135. Springer International publishing.
- Coleman C, Copetti D, Cipriani G, Hoffman S, Kozma P, Kovacs L, Morgante M, Testolin R and Di Gaspero G. 2009. The powdery mildew resistance gene *REN1* co-segregates with an NBS-LRR gene cluster in two Central Asian grapevines. *BMC Genet* 10:89–109.
- Coleman C. 2016. Positional Cloning and Functional analysis of *REN1* analysis in Kishmish vatkana. Dissertation, University of Missouri, Columbia.
- Costello et al. 2018. Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. *BMC Genomics* 19.
- de Coster, Sverre D'Hert, Darrin T Schultz, Marc Cruts and Christine Van Broeckhoven. 2018. NanoPack: visualizing and processing long-read sequencing data, *Bioinformatics* 34:2666–2669.
- de Lannoy C, de Ridder D and Risse J. 2017. A sequencer coming of age: *De novo* genome assembly using MinION reads. *F1000Res*. 6.
- Dohm JC, Lottaz C, Borodina T and Himmelbauer H. 2008. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res* 36:e105.
- Heather JM and Chain B. 2016. The sequence of sequencers: The history of sequencing DNA, *Genomics* 107:1–8.

- Hillier et al. 2008. Whole genome sequencing and variant discovery in *C. elegans*. *Nat Meth* 5:183-188.
- Jaillon O et al. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449:463–467.
- Jain et al. 2018. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol* 36:338–345.
- Konstantinidis KT, Read T, Kyrpides N, Tsementzi D and Luo C. 2012. Direct Comparisons of Illumina vs. Roche 454 Sequencing Technologies on the Same Microbial Community DNA Sample. *Plos ONE* 7.
- Koren S, Walenz BP, Berlin K, Miller JR and Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* 27:722–736.
- Lander et al. 2001. Initial sequencing and analysis of the human genome. *Nature* 409:860–921.
- Levene MJ, Korlach J, Turner SW, Foquet M, Craighead HG and Webb WW. 2003. Zero-mode waveguides for single-molecule analysis at high concentrations. *Science* 31:682-686.
- Liu Q, Guo Y, Li J, Zhang B and Shyr Y. 2012. Steps to ensure accuracy in genotype and SNP calling from Illumina sequencing data. *BMC Genomics* 13:1-8.
- Loman NJ, Quick J and Simpson JT. 2012. Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol* 30:434-439.
- Loman NJ and Quinlan AR. 2014. Poretools: a toolkit for analyzing nanopore sequence data. *Bioinformatics* 30: 3399-4401.
- Metzker ML. 2010. Sequencing technologies—the next generation. *Nat Rev Genet* 11:31-46.
- Rhoads A and Au KF. 2015. PacBio sequencing and its applications *Genomics Proteomics Bioinformatics* 13:278-289.
- Sambrook J and Russell DW. 2006. *The Condensed Protocols from Molecular Cloning: a Laboratory Manual*, 1st ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
- Sanger F, Nicklen S and Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA* 74:5463-5467.
- Schadt EE, Turner S and Kasarskis A. 2010. A window into third-generation sequencing. *Hum Mol Gen* 19:227–240.

- Treangen TJ and Salzberg SL. 2013. Repetitive DNA and next-generation sequencing: computational challenges and solutions *Nat Rev Genet* 13:36-46.
- Van Tassell CP, Smith TP, Matukumalli LK, Taylor JF, Schnabel RD, Lawley CT, Haudenschild CD, Moore SS, Warren WC and Sonstegard TS. 2008. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nat Methods* 5:247–252.
- Venter et al. 2001. The sequence of the human genome. *Science* 291:1304–1351.
- Voelkerding KV, Dames SA and Durtschi JD. 2009. Next-generation sequencing: from basic research to diagnostics. *Clin Chem* 55:641–658.
- Whiteford N, Skelly T, Curtis C, Ritchie ME, Lohr A, Zaranek AW, Abnizova I and Brown C. 2009. Swift: primary data analysis for the Illumina Solexa sequencing platform. *Bioinformatics*. 25:2194–2199.

## CHAPTER 2

### SEQUENCE ANALYSIS AND HAPLOTYPING OF PART OF THE *REN1* REGION

#### Introduction

Grapevine is a highly heterozygous species, which means that the nucleotide sequence of each of its chromosomes is different from the corresponding homolog (Fang et al. 2010). Most of the heterozygosity in grapes is due to single nucleotide polymorphisms, or SNPs, which may occur in both coding and non-coding regions of the DNA. Lijavetzky et al. (2007) found that, on average, there is one SNP in every 64 nucleotides of grape DNA, with a distribution of one SNP in every 47 nucleotides in non-coding regions and one SNP in every 69 nucleotides in coding regions. Depending on their location in a chromosome, SNPs can have potentially important functional consequences. Those outside of protein-coding regions can influence the binding of regulatory proteins, which, in turn, can alter gene expression levels. SNPs inside protein-coding regions can cause missense mutations, non-sense mutations or silent mutations. Missense mutations lead to different sense codons, whereas non-sense mutations convert codons to translation termination sites and lead to truncated proteins. These altered variants of the wild-type proteins may cause a dramatically different phenotype in the plant. Heterozygosity can also be caused by polymorphism at repetitive sequences in grapevines. Repetitive sequences can be tandem repeats or dispersed repeats, and depending on genomic location, either of these may have an impact on the phenotype of the plant. If a repetitive sequence is distributed randomly along chromosomes, then it is termed a dispersed repeat. Dispersed repeats constitute transposons as well as complete and truncated copies of retroviruses and gene paralogues, and found in all chromosomes of a plant. Tandem repeats are most commonly constituted by

microsatellites and minisatellites (Richard et al. 2008), but tRNA genes and repeat sequences at the telomeric region are also categorized as tandem repeats (Sharma et al. 2005). Microsatellites have a 2-5 nucleotide motif that is repeated 5-50 times (Richard et al. 2008, Gulcher, 2012), whereas minisatellites have a 6-100 nucleotide repeat size, with an average of a ~15 nucleotide motif. Both microsatellites and minisatellites together are termed as variable number of tandem repeats (VNTRs).

As both SNPs and transposons are the results of random mutations and transposition events, the majority of them occur in regions of the genome where they have no phenotypic impact at all. Variation in the number of repeated elements also rarely affect phenotype. Regardless of phenotype effect, SNPs and microsatellites are used as markers of chromosomal locations. If a polymorphic SNP or microsatellite is identified and an assay is developed for its routine detection, then it is referred to as a molecular marker, and its variants are considered marker alleles. The abundance of SNPs and microsatellites in grapevine facilitated the development of dense linkage maps, which in turn, enabled geneticists to map the position of important genes in the grape genome. In addition, the use of molecular markers helped identify grape genotypes and reveal important information about diversity among cultivated vines and wild populations (Myles et al. 2011, de Andres et al. 2012).

One phenomenon in which the role of high heterozygosity in grapevine is well understood is immunity against pathogens. Grapevines, as most other woody perennial plant species, maintain a high-level of polymorphism in resistance genes, and the resulting heterozygote advantage helps protect wild populations of grapes from rapidly evolving pathogens. In contrast to animals, plants depend exclusively on a cell-autonomous defense system against disease. Consequently, plants cells evolved to have a sophisticated innate defense

against pathogens, which is efficient at recognizing and killing pathogens (Jones et al. 2006). Innate immunity in plants has two layers: (1) pathogen- or microbe-associated molecular pattern (PAMP or MAMP) triggered immunity (PTI) and (2) effector-triggered immunity (ETI). The primary immunity is PTI and is activated immediately after the pathogen comes into contact with the host cell. The ETI may be initiated when the pathogen is successful at subverting PTI by secreting its effector proteins into the plant cell.

PTI is initiated when the plant cell recognizes PAMPs/MAMPs or endogenous danger-associated molecular patterns (DAMPs) at the host cell plasma membrane (Macho et al. 2014). These PAMPs are detected as foreign molecular entities by plant transmembrane proteins termed pattern recognition receptors (PRRs). PAMPs include flagellin, lipopolysaccharide, and chitin, cellular components that are essential for the pathogen's survival and reproduction (Chishlom et al. 2006). PAMPs are highly conserved among microbes and are absent in the host, which enables them to be specifically recognized as “foreign” by the host PRRs. Among the best-understood examples of PRRs are FLAGELLIN-SENSITIVE 2 (FLS2) and HOST ELONGATION FACTOR RECEPTOR (EFR) in *Arabidopsis thaliana*, which recognize the bacterial proteins flagellin and elongation factor EF-Tu, respectively (Kunze et al. 2004, Zipfel et al. 2004). PTI provides immunity to plants against adapted and non-adapted pathogens and so it is a basal line of non-specific immunity or primary immunity (Spoel et al. 2012). The PRR at the plasma membrane is complexed with several co-receptors and kinases to enable PRRs to initiate a signaling cascade (Macho et al. 2014). The signal from the receptor is transferred to the MAP kinase signaling pathway and then to WRKY transcription factors. Transcription of PTI related-genes by WRKY factors will then sets the counterattack against the pathogen into motion.



Pathogens evolved to inhibit PTI by either preventing the plant cell from recognizing PAMPS or, more commonly, by blocking the ability of the plant to transfer the intracellular signal from the PRRs to the nucleus. Both bacterial and fungal pathogens block signaling by secreting effector proteins into the plant cell. If the plant cell is unable to detect the presence or activity of the effector, then PTI fails to be activated and the pathogen will kill the plant cell or draw its resources unimpeded. Plants, however, evolved to express resistance (R) proteins, which have the ability to recognize pathogen effectors. If this recognition of an effector is successful, the plant cell initiates the second layer of defense, ETI (Shao et al. 2003, Chisholme et al. 2006).

The key players of ETI are the R proteins. R proteins are large polypeptides that detect the presence or activity of an effector in the plant cell cytoplasm and initiate the transduction of a signal to the nucleus. Most R proteins do not interact directly with their specific cognizant effector proteins. Instead of directly interacting with the R protein, another protein, named the guard protein may interact with the effector. This interaction is then detected by the R protein and activates resistance. In this case, the target of the effector protein is the guard protein (Der Biezen et al. 1998). The strongest experimental support for this hypothesis comes from the interaction of the R protein Rps2, which detects the ubiquitination of the RIN4 guard protein by the bacterial effector Rpt2 in *A. thaliana* (Mackey et al. 2003). Most R proteins have a conserved structure, which contains a nucleotide-binding leucine-rich repeat domain (NB-LRR) domain. NB-LRR proteins fall in one of two categories based on the N-terminal domains present in the protein. The first NB-LRR class is the Toll Interleukin-1 receptor (TIR)-NB-LRR, which contains the TIR domain at the amino terminal end of the protein. The second class is the coiled coil (CC)-NB-LRR protein with a CC domain at its amino terminal end. The coiled coil or the TIR region present at the amino terminal end of the NB domain is responsible for binding to

other cellular proteins, which causes downstream signaling (Tao et al. 2002, Xu et al. 2000). The NB domain is specific for binding with ATP or ADP (Reidl et al. 2005, Yan et al. 2005). The ADP-bound state is inactive and the ATP-bound state is the active signaling state of the NB-LRR protein.

In my thesis research, I sequenced a cloned DNA fragment of the powdery mildew-resistant *Vitis vinifera* grape variety ‘Kishmish Vatkana’ (KV) and took advantage of the high heterozygosity of the species to determine that the cloned fragment is derived from the resistance homologue. I also searched the nucleotide sequence of the clone and identified an NB-LRR gene, which may be a candidate to confer resistance to this grape variety against powdery mildew.

## **Materials and Methods**

**Sequence Analysis.** The ONT sequence of BAC 81D11 was aligned with the *V. vinifera* PN40024 genome sequence to identify the coordinates that delimit the 81D11-homologous region in the reference genome sequence (Jaillon et al. 2007), and to compare the two haplotypes. Sequence alignments were performed using the Basic Local Alignment Search Tool (BLAST) with default parameters (Altschul et al. 1990; accessible at [https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE\\_TYPE=BlastSearch&LINK\\_LOC=blasthome](https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome)). The gene content of the 81D11 ONT sequence was analyzed using an *ab initio* gene identification software FGENESH (Salamov et al. 2000; accessible at <http://www.softberry.com/berry.phtml?topic=fgenes&group=programs&subgroup=gfind>). The predicted genes and their predicted mRNA and amino acid sequences were analyzed by BLASTN and BLASTP software. The predicted polypeptides encoded by the hypothetical genes were characterized based on amino acid sequence similarity searches using the Interproscan

tool (Zdobnov et al. 2001; accessible at <http://www.ebi.ac.uk/interpro/search/sequence-search>). This tool aligns the amino acid sequences against protein signature databases from InterPro (Mulder et al. 2005). Gene ontology (GO) information about all predicted proteins was obtained by searching the Arabidopsis GO database; accessible at <https://www.arabidopsis.org/>. Predicted proteins were also characterized by comparing them with their Arabidopsis orthologs.

**Haplotyping of 81D11.** To determine which haplotype of KV chromosome 13 was represented by the 81D11 BAC clone, a 1,044-nucleotide-long internal fragment within the 2,777 nucleotide fragment from both haplotypes of the susceptible parent ‘Thompson Seedless’ (TS) was compared to the corresponding 81D11 fragment. This was accomplished by cloning and Sanger-sequencing several PCR-amplified copies of this fragment. PCR amplification was accomplished using unique primers and cloning was performed using the T/A cloning kit (NEB BioLabs, Ipswich MA, USA). Sanger sequencing and alignment was performed to identify indels and SNPs that could differentiate the resistance-associated KV haplotype from the two different TS haplotypes. The primers from both ends were designed using Primer 3 software accessible at <http://primer3.ut.ee/>.

To clone the 2,777-nucleotide long fragment, a high-fidelity PCR reaction was carried out with 81D11, KV, ‘Vassarga Tcherniaia’ (VT) and TS template DNA. High-fidelity PCR reactions contained buffer (1X), MgCl (0.425mM), dNTP (0.2 mM), HiFidelity Taq polymerase (0.02 U), forward and reverse primers (1 uM each) and about 10 ng of DNA. High-fidelity PCR reactions for TS included buffer (1X), MgCl (0.875mM), dNTP (0.2 mM), enzyme (0.02 U), primers (1 uM each) and about 4 ng of DNA. Agarose gel (1%) electrophoresis was performed in order to identify the PCR products. The ‘TS’ PCR amplicons were fractionated using agarose gel electrophoresis in TAE buffer, and the band of interest was extracted from the gel. Another

round of PCR was performed using 4 internal primer sets within the 2,777-nucleotide region of the separated DNA band. The 2.7 kb PCR amplified products from 81D11, KV and VT were ligated to the pMiniT vector and transformed in *E. coli* 10 beta cells according to the manufacturer's protocol (New England PCR Cloning Kit, New England BioLabs, Ipswich MA, USA). The 1,044 nucleotide fragment within the 2.7 kb fragment was cloned to the TOPO® vector and transformed in One Shot® *E. coli* cells according to the manufacturer's protocol using NEB PCR Cloning Kit (NEB BioLabs, Ipswich, MA, USA). A colony PCR was performed on colonies selected from all samples to identify the banding pattern produced, in an assumption that all colonies that produced the same banding patterns likely contained the same haplotype. Two haplotypes were selected for sequencing from each individual (KV, VT, and TS). Since 81D11 is a single haplotype, one of the 81D11 colonies was selected randomly from all of the colonies. Plasmid DNA was extracted from all seven samples using the GenElute™ Plasmid Miniprep Kit (Sigma Aldrich, St. Louis, MO, USA) according to the manufacturer's protocol (81D11, KV-1, KV-2, VT-1, VT-2, TS-1, and TS-2), and Sanger sequencing was performed by Nevada Sequencing Center, Reno, NV.

## Results

**Sequencing and haplotyping of BAC81D11.** BAC81D11 is homologous to the corresponding *V. vinifera* PN40024 reference genome sequence between coordinates 16,836,214 and 16,921,889 on chromosome 13. There are four regions in the BAC81D11 ONT sequence that are absent from the corresponding region in PN40024. These four regions fall between coordinates 17,290 and 17,471, between 23,383 and 23,631, between 49,378 and between 64,112, and 69,614 and 69,750 in the BAC81D11 insert, respectively. There are six regions in

the corresponding region of the reference sequence that are not present in the BAC 81D11 ONT sequence. The length of the PN40024 reference sequence corresponding to the 69,750 nucleotide-long 81D11 ONT sequence is 85,676 nucleotides. Two different haplotypes from TS, KV, and VT were separated by ligating each haplotype fragment into a separate plasmid vector. Both haplotypes from TS, VT, and KV are Sanger sequenced and were differentiated into TS-1, TS-2, VT-1 and VT-2, and KV-1 and KV-2. The Sanger sequences of each haplotype from the susceptible parent, TS, were compared with the BAC81D11 Sanger sequence. Genotyping based on Sanger sequencing of selected fragments was performed with a 304-nucleotide fragment from both haplotypes of TS and 81D11. There were four SNPs and one indel between the two TS haplotypes in the 304-nucleotide fragment from the REN1-homologous region. Three SNPs were observed between one haplotype (TS-1) of TS and BAC81D11 and eight SNPs, including one indel, were observed between the other haplotype (TS-2) of TS and the BAC81D11 fragment (Figure 8). This is further supported by my results that the Sanger-sequenced haplotype of the 2.7 kb fragment from both KV and VT, named KV-1 and VT-1, were identical to the inset of BAC81D11.

**Gene Content of the BAC81D11 insert.** The coding sequences (CDS) of three structural genes were predicted in the 69,750 nucleotide-long BAC 81D11 insert using the FGENESH gene finding software (Figure 9). In addition to CDS predictions, this analysis provided the exact coordinates of the of the transcription start site and intron-exon boundaries, as well as the mRNA and the amino acid sequences of the predicted proteins. Coordinates and mRNA information of each predicted gene are listed in Table 4. The hypothetical genes predicted encode a TCP-like transcription factor (TCP TF), a cinnamoyl CoA reductase-1(CCR-1) like protein and a recognition of *Peronospora parasitica* 13 (RPP13)-like disease resistance protein. In addition, a

retrotransposon region with a reverse transcriptase gene was also predicted in the 81D11 BAC insert (Figure 9).

**Predicted Gene Sequence Analysis.** Orthologous genes for all the hypothetical CDS were identified and are listed in Table 5, and superfamily and domain information of the hypothetical protein sequences are shown in Table 6. The Gene Ontology information of all the proteins are listed in Table 7. The predicted CDS, protein domain and Gene Ontology information are all in agreement with the FGENSH prediction that the BAC81D11 insert contains genes for a transcription factor, a CCR-1 like enzyme, and an NBS-LRR protein. Comparison of the amino acid sequence of the TCP transcription factor gene with the reference genome PN040024 showed that the BAC81D11 and the reference genome sequence of this gene are identical (Table 8). Further comparison with the KV-1 and VT-1 haplotype amino acid sequences revealed that the 81D11, VT-1, KV-1 sequences are 100% identical, but the transcription factor gene differs at six SNPs in the TS-1 haplotype (Figure 10). The polypeptide chain analysis of the TS-1 and 81D11 revealed that among the six SNPs, three are missense mutations and the remaining three are silent mutations (Figure 11).

The CCR-1 gene is not present at the corresponding location in the reference genome sequence. Paralogous sequences of this hypothetical protein share sequence homology with multiple sites on chromosome 13 and other chromosomes in the reference genome sequence. The RPP13 gene itself is not present in the 81D11 coordinate of the reference genome, or present only as a substantially diverged variant. The longest alignment present at the reference genome is 1,871 nucleotides long (14.3% of the gene sequence) which is present at chromosome 13 at the *REN1* region.

**Table 4** Hypothetical coding regions and their coordinates in BAC81D11.

<b>Hypothetical Genes Name</b>	<b>Coordinates</b>		<b>mRNA</b>
	<b>pBAC81D11</b>	<b>Reference Genome</b>	
Transcription Factor	16061- 16451	16877245-16877637	393 nt
CCR-1 Like-Protein	36323 - 49409	16900536-16911637	1,665 nt
Disease Resistant Protein	52019 - 65134	16989175- 16991697	3,123 nt
Retrotransposon Region	66186-69584	16918443-16921852	1854 nt

**Table 5** Ortholog information of genes of BAC81D11.

<b>Gene Name</b>	<b>NCBI Accession Number of Nearest Ortholog</b>	<b>Ortholog Species</b>
Transcription Factor- Gene#1	XM_021793362.1	<i>Hevea brasiliensis</i>
CCR-1 Like -Protein- Gene#2	XM_018984607.1	<i>Juglans regia</i>
Disease Resistant Protein- Gene#3	XM_010927512.1	<i>Elaeis guineensis</i>
Retrotransposon <i>pol</i> gene	XM_024171691.1	<i>Morus notabilis</i>

**Table 6** Hypothetical proteins encoded by genes of BAC81D11.

<b>Predicted Protein</b>	<b>Amino Acid Sequence length</b>	<b>Domains</b>	<b>Domain Coordinates in Amino acid Sequence</b>	<b>Domain Interpro ID</b>	<b>Superfamily</b>
Transcription Factor	130	TCP	68 - 128	IPR017887	None
CCR Like Protein	554	NAD dependent Epimerase/ Dehydratase	58 - 297	IPR001509	NAD(P) – binding Superfamily (2)
Disease Resistant Protein	1040	NB-ARC, Rx, N-terminal	367 – 405, 431 – 551 & 560 - 606	IPR041118, IPR002182	P-loop containing nucleotide triphosphate superfamily, Leucine rich repeat Superfamily (2)
Pol polyprotein of LTR retrotransposon	617	Reverse transcriptase, RNA-dependent DNA polymerase	479 - 549	IPR013103	Ribonuclease H

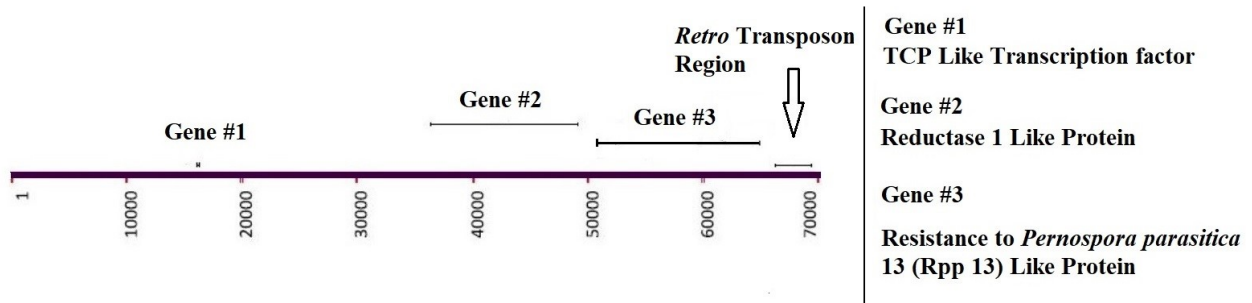
**Table 7** Gene ontology information for genes of BAC81D11.

<b>Gene Name</b>	<b>Accession Number</b>	<b>Molecular Function</b>	<b>Cellular Location</b>	<b>Biological Process</b>
Transcription Factor	AT1G58100	Regulation of transcription	Nucleus	RNA polymerase II proximal promotor sequence-specific DNA binding, Transcription regulatory region DNA binding.
CCR-1 Like Protein	AT1G5950.1	Catalytic activity, Cinnamoyl CoA reductase activity, Coenzyme binding, Oxidoreductase activity, Oxidoreductase activity acting on the CH-OH group of donors, NAD and NADP as acceptor.	Cytoplasm, Cytosol	Lignin biosynthetic process, oxidation-reduction process, Circadian rhythm, Response to cold
RPP 13 Like Protein	AT1G58602.1	ADP binding, ATP binding	Guard cell	Defense Response



			ai.bi.ci	
TS-2	TATCACTAACAAACAACATAAAATAT	TCAAAAAAATTTAAAGTATAAATGTTAAACTA	59	
TS-1	TATCACTAACAAACAACATAAAATAT	TCAAAAAAATTTAAAGTATAAATGTTAAACTA	60	
81D11	TATCACTAACAAACAACATAAAATAT	TCAAAAAAATTTAAAGTATAAATGTTAAACTA	60	
*****				
TS-2	ATCTCTTTTACGTATCCTTTATCGAATAACCATATATGCATGTCATTAAATGGTATGAGTG	119		
TS-1	ATCTCTTTTACGTATCCTTTATCGAATAACCATATATGCATGTCATTAAATGGTATGAGTG	120		
81D11	ATCTCTTTTACGTATCCTTTATCGAATAACCATATATGCATGTCATTAAATGGTATGAGTG	120		
*****				
TS-2	TTATCCCTATCTTCAAATCCTTTTAAATTAATATAGTATTTTGATGTACTAAAAGAAT	179		
TS-1	TTATCCCTATCTTCAAATCCTTTTAAATTAATATAGTATTTTGATGTACTAAAAGAAT	180		
81D11	TTATCCCTATCTTCAAATCCTTTTAAATTAATATAGTATTTTGATGTACTAAAAGAAT	180		
*****				
TS-2	TTTAGTCTCCTTTAAAATGACCAAATGATTTTTTAACATTTAATCCATTCAACTTTGAAT	239		
TS-1	TTTAGTCTCCTTTAAAATGACCAAATGATTTTTTAACATTTAATCCATTCAACTTTGAAT	240		
81D11	TTTAGTCTCCTTTAAAATGACCAAATGATTTTTTAACATTTAATCCATTCAACTTTGAAT	240		
*****				
TS-2	CATGTTTACTAAATTTATATGGTTGTCTTTCTACTTTCTTTATTTTAGATAGAGTTAG	299		
TS-1	CATGTTTACTAAATTTATATGGTTGTCTTTCTACTTTCTTTATTTTAGATAGAGTTAG	300		
81D11	CATGTTTACTAAATTTATATGGTTGTCTTTCTACTTTCTTTATTTTAGATAGAGTTAG	300		
*****				
TS-2	ATGT	303		
TS-1	ATGT	304		
81D11	ATGT	304		
****				

**Figure 8** Alignment of a 304-nucleotide fragment of 81D11 with the corresponding haplotypes from ‘Thompson Seedless’ (TS-1 and TS-2). SNPs and indels are marked in red boxes.



**Figure 9** Map of the BAC81D11 insert showing the position of predicted genes and the retrotransposon fragment.

TS-1	ATGGATGAAGATGTTCTTACTGTGGAACAACCTTAGAGAAGTGATGCCATCGCAAAC	60
81D11	ATGGATGAAGATGTTCTTACTGTGGAACAACCTTAGAGAAGTGATGCCATCGCAAAC	60
VT-1	ATGGATGAAGATGTTCTTACTGTGGAACAACCTTAGAGAAGTGATGCCATCGCAAAC	60
KV-1	ATGGATGAAGATGTTCTTACTGTGGAACAACCTTAGAGAAGTGATGCCATCGCAAAC	60
*****		
TS-1	ATGACCGAAGAAGGAATCCAGTAACACAAAAATTCATGCAATTGAAGAGCTTCCAGCA	120
81D11	ATGACCGAAGAAGGAATCCAGTAACACAAAAATTCATGCAATTGAAGAGCTTCCAGCA	120
VT-1	ATGACCGAAGAAGGAATCCAGTAACACAAAAATTCATGCAATTGAAGAGCTTCCAGCA	120
KV-1	ATGACCGAAGAAGGAATCCAGTAACACAAAAATTCATGCAATTGAAGAGCTTCCAGCA	120
*****		
TS-1	GCACCATTAGGAGCAAAACAAGAATCAAAGGCTTTGAGAAAGGCGACAAAACATAAACCT	180
81D11	GCACCATTAGGAGCAAAACAAGAATCAAAGGCTTTGAGAAAGGCGACAAAACATAAACCT	180
VT-1	GCACCATTAGGAGCAAAACAAGAATCAAAGGCTTTGAGAAAGGCGACAAAACATAAACCT	180
KV-1	GCACCATTAGGAGCAAAACAAGAATCAAAGGCTTTGAGAAAGGCGACAAAACATAAACCT	180
*****		
TS-1	AGGGTTTCGACCACAAAAAACCAACAGATCCATGTGAAGGTGGAGGGACGTGGGCGA	240
81D11	AGGGTTTCGACCACAAAAAACCAACAGATCCATGTGAAGGTGGAGGGACGTGGGCGA	240
VT-1	AGGGTTTCGACCACAAAAAACCAACAGATCCATGTGAAGGTGGAGGGACGTGGGCGA	240
KV-1	AGGGTTTCGACCACAAAAAACCAACAGATCCATGTGAAGGTGGAGGGACGTGGGCGA	240
*****		
TS-1	AGCATTCCGTTACCAAATCATGTGCAAATGAACTCTTTGAGCTCACACGACCGTCAAC	300
81D11	AGCATTCCGTTACCAAATCATGTGCAAATGAACTCTTTGAGCTCACACGACCGTCAAC	300
VT-1	AGCATTCCGTTACCAAATCATGTGCAAATGAACTCTTTGAGCTCACACGACCGTCAAC	300
KV-1	AGCATTCCGTTACCAAATCATGTGCAAATGAACTCTTTGAGCTCACACGACCGTCAAC	300
*****		
TS-1	TACAAGTGGGCTGGCCAAACCATCTGTTGGCTTTTGGAGAATGTCGAACCAGCAATCATC	360
81D11	TACAAGTGGGCTGGCCAAACCATCTGTTGGCTTTTGGAGAATGTCGAACCAGCAATCATC	360
VT-1	TACAAGTGGGCTGGCCAAACCATCTGTTGGCTTTTGGAGAATGTCGAACCAGCAATCATC	360
KV-1	TACAAGTGGGCTGGCCAAACCATCTGTTGGCTTTTGGAGAATGTCGAACCAGCAATCATC	360
*****		
TS-1	AAAGCCACCAGTACAAAGGAGAAGAAGAACTAG	393
81D11	AAAGCCACCAGTACAAAGGAGAAGAAGAACTAG	393
VT-1	AAAGCCACCAGTACAAAGGAGAAGAAGAACTAG	393
KV-1	AAAGCCACCAGTACAAAGGAGAAGAAGAACTAG	393
*****		

**Figure 10** Alignment of transcription factor gene of BAC81D11, and TS-1, VT-1 and KV-1 haplotypes. SNPs are numbered and marked in red boxes.

TS1	MDDEVPTVEQLLEKCMPSQTMTEEGTFRANTKIHAIEELPAAPLGAKQESKVLRKATKLKP	60
81D11	MDDEVPTVEQLLEKCMPSQTMTEEGTFRNTKIHAIEELPAAPLGAKQESKVLRKATKLKP	60
*****		
TS1	RVSTTKKPTDRHVKVEGRGRSIRLPNGCANELFELTRRLNYKWAGQTICWLLLENVEPAII	120
81D11	RVSTTKKPTDRHVKVEGRGRSIRLPNACANELFELTRRLNYKWAGQTICWLLLENVEPAII	120
*****		
TS1	KATSTREKKN	130
81D11	KATSTKEKKN	130
*****		

**Figure 11** Alignment of the transcription factor amino acid sequence of TS-1 and 81D11. Amino acid substitutions and silent mutations are highlighted with and green boxes, respectively.

## Discussion

BAC 81D11 is a heretofore unsequenced BAC clone from the *REN1* region of the resistance haplotype of the *V. vinifera* variety KV. In this work, I haplotyped this clone by taking advantage of its richness in polymorphism. A 1,180-nucleotide fragment containing a transcription factor gene was subcloned from BAC81D11, two haplotypes from KV (KV-1 and KV-2) and two haplotypes from its susceptible parent TS (TS-1 and TS-2) were sequenced, and the sequence data were aligned to one-another. Although the sequencing of the entire 1,180-nucleotide fragment was unsuccessful in one of these subclones, a 304 nucleotide-long sequence could be successfully sequenced and was available in all of these haplotypes. I found that the nucleotide sequence of the BAC81D11 clone was different from either haplotype of the susceptible parent TS, which meant that the BAC81D11 haplotype must have been passed down from the disease-resistant parent VT. This conclusion is further supported by my results that the Sanger-sequenced haplotype of the 2.8 kb fragment from both KV and VT, named KV-1 and VT-1, respectively, were identical to the corresponding fragment in BAC81D11. Taken together, these data provide conclusive evidence that the insert in BAC81D11 is a fragment of the resistance haplotype of KV. Even though the 3040-nucleotide region is short, I have confidence in these data because this stretch of BAC81D11 DNA differs from its TS-1 counterpart in 3 SNPs and from its TS-2 counterpart in seven SNPs plus an indel (Figure 8). Furthermore, I examined the raw sequence reads by directly comparing the chromatogram of the haplotype to one-another in all possible combinations. The direct examination of chromatogram data excluded the possibility of mistaking sequencing errors for polymorphisms. The sequence divergence in this short stretch of non-coding DNA vividly demonstrates the high-level of heterozygosity of *V. vinifera* and other *Vitis* species (Myles et al. 2009).

The *REN1* region had been mapped by positional cloning and its representative BAC clones identified by a combined genetic and physical mapping approach (Coleman 2016). Among the *REN1* BAC clones is BAC81D11, which had been tentatively mapped to the left side of the first gap in the physical map. Because of its tentative position, BAC 81D11 could provide important novel information in its gene content and potentially filling one of the gaps in the *REN1* physical map. Therefore, I sequenced this BAC clone using ONT hardware and software and predicting gene content using FGENESH and Genscan. Both programs identified three putative genes, namely a TCP-like transcription factor, a cinnamoyl-CoA reductase-like protein, and a RPP13-like protein, plus identified sequences characteristic of an LTR retrotransposon containing a partial sequence of a pol polyprotein gene.

**Characterization of the TCP like transcription factor.** This transcription factor (TF) gene product is best studied in *Arabidopsis thaliana*. The TF gene is named TCP TF gene family after its founding members TEOSINTE BRANCHED1 (TB1), CYCLOIDEA (CYC), PROLIFERATING CELL NUCLEAR ANTIGEN FACTOR1 (PCF1) and PCF2 (Li 2015). The polypeptide chain of the TF gene has 130 amino acids. The predicted TF gene is classified into Class I TCP-like TF because of the presence of Arg-16, Cys-21, and Ala-22 in its polypeptide sequence (Aggarwal et al., 2010, Li 2015). Class 1 TCP TFs have a conserved cysteine residue at the 20<sup>th</sup> position in the 60 amino acid-long TCP domain, which differentiates the Class I from the Class II TCP domain where no Cys-20 is present. Also, the Cys-20 is followed by an alanine residue and an arginine residue at the 15<sup>th</sup> position in all the class 1 TCP domains of the TF polypeptide chains (Li 2015). The cloned TF gene has 61 amino acids present in the TCP domain. Therefore, one position downstream for all these conserved TCP class I amino acids in the predicted gene will be due to the extra one amino acid present in the TCP domain, and

therefore, it can be classified as class I TCP TF. The TCP domain is responsible for the DNA-binding property of this TF protein family and the Cys-20 residue plays an important role in its regulation. Interaction of this cysteine residue with oxidants inhibits DNA-binding activity, and this inhibition can be reversed by the presence of reductants (Viola et al. 2011).

Members of the TCP TF protein family have been implicated in ETI and other plant developmental function and therefore it is of interest for this study. A single mutation in three TCP TF paralogs TCP13, TCP14 and TCP19 in *Arabidopsis* leads to considerably higher susceptibility to infection to two different avirulent *Hyaloperonospora arabidopsidis* isolates (Mukhtar et al. 2011). On the other hand, the same study found that a mutation which abolished TCP15 expression exhibits enhanced disease resistance to a virulent *H. arabidopsidis* isolate. Therefore, it was concluded that TCPs can act as negative regulators of ETI also. An indirect but noteworthy piece of evidence for the involvement of TCP in ETI is that several of the TCPs interact with the *SUPPRESSOR OF Rps4- RLD1 (SRFR1)* at the nucleus (Kim et al. 2014). *SRFR1* is a suppressor of ETI. It is also evident that several TCP paralogs in *A. thaliana* cause the expression of genes that antagonistically act on *SRFR1*, thereby balancing the immune response in *Arabidopsis* (Mukhtar et al. 2009). Kim et al. (2014) performed yeast two-hybrid assays to identify whether TCP TFs are involved in interaction with SRFR1 to block their immunity inhibition. Six TCP TF paralogs have been shown to interact with SRFR1 (Kim et al. 2014).

The coding sequence of the TF gene in the *REN1* region is 390 nucleotides long and it contains a single exon. As in other plant species, the predicted polypeptide chain is 130 amino acids long and contains a 60 amino acid-long TCP domain. The TCP coding sequence of BAC81D11 is 100% identical to the VT-1, and KV-1 haplotypes, but different from the TS-1

haplotype of TS. Among the six SNPs between 81D11 and TS-1 TF genes, three of the SNPs are missense mutations and the remaining three are silent mutations. All three silent mutations fall in arginine codons. The three missense mutations in the polypeptide chain are A28T, G87A and R126K). Among the three missense mutations, A28T and R126K are of approximately the same size and have the same hydropathy profile. The G87A amino acid substitution causes the greatest change in size and hydropathy. Two of the three missense mutations and all three silent mutations are in the DNA-binding TCP domain of the TF polypeptide chain (Table 9). Nonetheless, it is difficult to consider that any of these polymorphisms may influence disease resistance in KV, because the powdery mildew susceptible PN40024 genome has the exact same TCP allele.

**Characterization of the Predicted CCR- 1 like protein.** Along with cinnamyl alcohol dehydrogenase (CAD), cinnamoyl-CoA reductase enzymes function in the conversion of cinnamoyl-CoA esters to monolignols. Cinnamoyl-CoA ester is the precursor of monolignol, which is the monomer of lignin (Boerjal et al. 2003). Lignin acts as a physical barrier to almost all pathogens (Moeshbacher et al. 1990), and lignin-like phenolic compounds are deposited and concentrated at the site of pathogen attack (Reimers et al. 1991, Lange et al. 1995, Campbell et al. 1992, Anterola et al. 2002). Lignin is also deposited in the vascular tissues to provide strength to the cell wall (Boerjal et al. 2003), but this developmental lignin has been shown to be different from defense lignin in its monomeric constitution (Lange et al. 1995).

Cinnamoyl CoA reductase 1-like protein is best studied in *Oryza sativa* in which Cinnamoyl-CoA Reductase 1 (OsCCR1) has been shown to play a role in defense against pathogens (Kawasaki et al. 2005). There is strong evidence from previous studies that OsCCR1 functions in the biosynthesis of lignin. The expression of the *OsCCR1* gene is activated by the

OsRac1 protein, which is a GTPase and is involved in the production of reactive oxygen species (ROS). Lignin is synthesized from monolignols at the cell wall by peroxidase and H<sub>2</sub>O<sub>2</sub> (Boerjal et al. 2003). H<sub>2</sub>O<sub>2</sub> may enhance polymerization of monolignol and monolignols themselves have antimicrobial properties (Keen et al. 1979).

**Characterization of the Predicted *RPP13*-Like Protein.** The *RPP13* gene is best studied in *A. thaliana*. *RPP13* in Arabidopsis encodes a coiled-coil-type nucleotide-binding site and leucine rich repeat (CC-NBS-LRR) domain R gene (Liu et al. 2018). *RPP13* confers resistance against the biotrophic-oomycete *Pernospora parasitica*, which causes downy mildew disease in several plant species. The putative RPP13 protein from BAC81D11 was predicted to contain a coiled-coil and a nucleotide-binding site domain, but it was not predicted to have a leucine-rich repeat (LRR) domain. Nonetheless, certain elements of LRR domains are recognizable by direct observation, and the overall predicted structure the RPP13-like gene in BAC81D11 is recognized as an LRR protein by various databases (Table 9). Therefore, the RPP13-like gene from 81D11 can be considered as a CC-NBS-LRR protein (Ade et al. 2007).

## References

- Ade J, DeYoung BJ, Golstein C and Innes RW. 2007 Indirect activation of a plant nucleotide binding site-leucine-rich repeat protein by a bacterial protease. *Proc Natl Acad Sci USA* 104:2531-2536.
- Aggarwal P, Das Gupta M, Joseph AP, Chatterjee N, Srinivasan N and Nath U. 2010. Identification of specific DNA binding residues in the TCP family of transcription factors in Arabidopsis. *Plant Cell* 22:1174-1189.
- Anterola AM and Lewis NG. 2002. Trends in lignin modification: a comprehensive analysis of the effects of genetic manipulations/mutations on lignification and vascular integrity. *Phytochemistry* 61:221-294.
- Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* 215:403-410.

- Bittner-Eddy PD, Crute EB, Holub JL and Beynon JL. 2000. RPP13 is a simple locus in *Arabidopsis thaliana* for alleles that specify downy mildew resistance to different avirulence determinants in *Peronospora parasitica*. *Plant J* 21:177–188.
- Boerjan W, Ralph J and Baucher M. 2003. Lignin biosynthesis. *Annu Rev Plant Biol.* 54:519–546.
- Campbell MM and Ellis BE. 1992. Fungal elicitor-mediated responses in pine cell cultures. *Planta* 186:409–417.
- Chisholm ST, Coaker G, Day B and Staskawicz BJ. 2006. Host-microbe interactions: shaping the evolution of the plant immune response. *Cell* 124:803–814.
- Coleman C, Copetti D, Cipriani G, Hoffman S, Kozma P, Kovacs L, Morgante M, Testolin R and Di Gaspero G. 2009. The powdery mildew resistance gene *REN1* co-segregates with an NBS-LRR gene cluster in two Central Asian grapevines. *BMC Genet* 10: 89–109.
- Crute IR and Pink D. 1996. Genetics and utilization of pathogen resistance in plants. *Plant Cell* 8:1747–1755.
- de Andrés MT, Benito A, Pérez-Rivera G, Ocete R, Lopez MA, Gaforio L, Muñoz G, Cabello F, Martínez-Zapater JM and Arroyo-García R. 2012. Genetic diversity of wild grapevine populations in Spain and their genetic relationships with cultivated grapevines. *Mol Ecol* 21:800–816.
- Der Biezen E and Jones J. 1998. Plant disease-resistance proteins and the gene-for-gene concept. *Trends Biochem Sci* 23:454–456.
- Dodds PN, Lawrence GJ and Ellis JG. 2001. Six amino acid changes confined to the leucine-rich repeat  $\beta$ -strand/ $\beta$ -turn motif determine the difference between the P and P2 rust resistance specificities in flax. *Plant Cell* 13:163–178.
- Dodds PN and Schwechheimer C. 2002. A Breakdown in Defense Signaling. *Plant Cell* 14:5–8.
- Fang JG, Dong HQ, Cao X, Yang G, Yu HP, Nicholas KK and Wang C. 2010. Discovery and characterization of SNPs in *Vitis vinifera* and genetic assessment of some grapevine cultivars. *Sci Hortic* 125:233–238.
- Gulcher J. 2012. Microsatellite markers for linkage and association studies. *Cold Spring Harb Protoc.* 2012:425–432.
- Jaillon et al. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449:463–467.
- Jia Y, McAdams SA, Bryan GT, Hershey HP and Valent B. 2000. Direct interaction of resistance gene and avirulence gene products confers rice blast resistance. *EMBO J* 19:4004–4014.



- Jones J and Dangl JL. 2006. The plant immune system. *Nature* 444:323–329.
- Kobe B and Deisenhofer J. 1994. The leucine-rich repeat: A versatile binding motif. *Trends Biochem Sci* 19:415-421.
- Kawasaki T, Koita H, Nakatsubo T, Hasegawa K, Wakabayashi K, Takahashi H, Umemura K, Umezawa T and Shimamoto K. 2005. Cinnamoyl-CoA reductase, a key enzyme in lignin biosynthesis, is an effector of small GTPase Rac in defense signaling in rice. *Proc Natl Acad Sci USA* 103:230-235.
- Keen NT and Littlefield LJ. 1979. The possible association of phytoalexins with resistance gene expression in flax to *Melampsora lini* *Physiol. Plant Pathol* 14:265–280.
- Kim SH, Son GH, Bhattacharjee S, Kim HJ, Nam JC, Nguyen PD, Hong JC and Gassmann W. 2014. The Arabidopsis immune adaptor SRFR1 interacts with TCP transcription factors that redundantly contribute to effector-triggered immunity. *Plant J* 78:978-989.
- Kunze G, Zipfel C, Robatzek S, Niehaus K, Boller T and Felix G 2004. The N Terminus of Bacterial Elongation Factor Tu Elicits Innate Immunity in Arabidopsis Plants. *The Plant Cell* 16:3496–3507.
- Lange BM, Lapierre C, Sandermann H Jr. 1995. Elicitor-Induced Spruce Stress Lignin (Structural Similarity to Early Developmental Lignins). *Plant Physiol* 108:1277–1287.
- Li 2015. The Arabidopsis thaliana TCP transcription factors: A broadening horizon beyond development. *Plant Signal Behav* 10:1-12.
- Lijavetzky D, Cabezas JA, Ibáñez A, Rodríguez V and Martínez-Zapater JM. 2007. High throughput SNP discovery and genotyping in grapevine (*Vitis vinifera* L.) by combining a re-sequencing approach and SNPlex technology. *BMC Genomics* 8:424.
- Liu Z, Cheng J, Fan H, Li L, Hu B and Liu H. 2018. Genome-wide Identification and Expression Analyses of RPP13-like Genes in Barley. *BioChip journal* 12:102–113.
- Mackey D, Belkhadir Y, Alonso JM, Ecker JR and Dangl JL. 2003. Arabidopsis RIN4 is a target of the type III virulence effector AvrRpt2 and modulates RPS2-mediated resistance. *Cell* 112:379-389.
- Macho AP et al. 2014. Plant PRRs and the activation of innate immune signaling. *Mol Cell* 54:263-272.
- Moershbacher B, Noll U, Gorrichon L and Reisener HJ. 1990. Specific inhibition of lignification breaks hypersensitive resistance of wheat to stem rust. *Plant Physiol* 93:465–470.

- Myles S, Boyko AR, Owens CL, Brown PJ, Grassi F, Aradhya MK, Prins B, Reynolds A, Chia J-M, Ware D, Bustamante CD and Buckler ES. 2011. Genetic structure and domestication history of the grape. *Proc Natl Acad Sci USA* 108:3457–3458.
- Myles S, Chia M, Hurwitz B, Simon C, Zhong GY, Buckler E and Ware D. 2009. Rapid Genomic Characterization of the Genus *Vitis*. *PLoS One* 5.
- Mukhtar MS et al. 2011. Independently evolved virulence effectors converge onto hubs in a plant immune system network. *Science* 333:596-601.
- Mulder KW, Mulder KW, Winkler GS and Timmers HT. 2005. DNA damage and replication stress induced transcription of RNR genes is dependent on the Ccr4-Not complex. *Nucleic Acids Res* 33:6384-92.
- Mukhtar MS et al 2011. Independently evolved virulence effectors converge onto hubs in a plant immune system network. *Science* 333:596-601.
- Reimers PJ and Leach JE. 1991. Race-specific resistance to *Xanthomonas oryzae* pv. *oryzae* conferred by bacterial blight resistance gene Xa-10 in rice *Oryza sativa* involves accumulation of a lignin-like substance in host tissues. *Physiol. Mol Plant Pathol* 38:39–55.
- Richard GF, Kerrest A and Dujon B. 2008. Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. *Microbiol Mol Biol Rev* 72:686-727.
- Riedl SJ, Li W, Chao Y, Schwarzenbacher R and Shi Y. 2005. Structure of the apoptotic protease-activating factor 1 bound to ADP. *Nature* 434:926–933.
- Salamov AA and Solovyev. 2000. Ab initio gene finding in *Drosophila* genomic DNA. *Genome Res* 10:516-522.
- Shao F, Golstein C, Ade J, Stoutemyer M, Dixon JE and Innes RW. 2003. Cleavage of *Arabidopsis* PBS1 by a bacterial type III effector. *Science* 301:1230-1233.
- Sharma S and Raina SN. 2005. Organization and evolution of highly repeated satellite DNA sequences in plant chromosomes. *Cytogenet Genome Res* 109:15-26.
- Spoel SH and Dong X. 2012. How do plants achieve immunity? Defence without specialized immune cells. *Nat Rev Immunol* 12:89–100.
- Sprang SR. 1997. G protein mechanisms: insights from structural analysis. *Annu Rev Biochem* 66:639–678.
- Tao X, Xu Y, Zheng Y, Beg AA and Tong L. 2002. An extensively associated dimer in the structure of the C713S mutant of the TIR domain of human TLR2. *Biochem Biophys Res Commun* 299:216–221.

- Viola IL, Uberti Manassero NG, Ripoll R and Gonzalez DH. 2011. The Arabidopsis class I TCP transcription factor AtTCP11 is a developmental regulator with distinct DNA-binding properties due to the presence of a threonine residue at position 15 of the TCP domain. *Biochem J* 35:143-155.
- Wang XS, Wu WR, Jin GL and Zhu J. 2005. Genome-wide identification of R genes and exploitation of candidate RGA markers in rice. *Chin Sci Bull* 50:1120–1125.
- Xu Y. et al. 2000. Structural basis for signal transduction by the Toll/interleukin-1 receptor domains. *Nature* 408:111–115.
- Yan N. et al. 2005. Structure of the CED-4-CED-9 complex provides insights into programmed cell death in *Caenorhabditis elegans*. *Nature* 437:831–837.
- Zipfel C, Robatzek S, Navarro L, Oakeley EJ, Jones JD, Felix G and Boller T. 2004. Bacterial disease resistance in Arabidopsis through flagellin perception. *Nature*. 428:764-767.
- Zdobnov EM and Apweiler R. 2001. InterProScan--an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* 17:847-848.

## SUMMARY

This work focused on the molecular characterization of the *REN1* locus of the Central Asian *Vitis vinifera* variety ‘Kishmish Vatkana’. The *REN1* locus of this grape confers resistance to the economically important pathogen powdery mildew (*Erysiphe necator*). A DNA fragment, cloned as an insert of a bacterial artificial chromosome (BAC), was sequenced using the third-generation sequencing technique of Oxford Nanopore Technology (ONT). For this, a modified version of the ONT library preparation method was developed, which was based on linearizing the BAC clone with a restriction enzyme that cut the entire plasmid at a single site. Comparative analysis of the sequence data from the modified and the original ONT methods demonstrated that the modified protocol yielded substantially more data, deeper sequence coverage, higher read quality and greater read length than the original protocol. The resulting reads were then assembled into a single 69,750 nucleotide-long contig, which extended the sequence information of the *REN1* region by 46,890 nucleotides. The newly generated DNA sequence revealed that the BAC insert contained three complete open reading frames encoding functional proteins of a TCP-type transcription factor, a CCR-1 enzyme, and a CC-NBS-LRR-type resistance gene, all of which could potentially contribute to the resistance phenotype of ‘Kishmish Vatkana’. An approximately 2.7 kb fragment from within the locus was cloned in two haplotypes of the susceptible parent of ‘Kishmish Vatkana’, both of which differed from the BAC sequence at several single nucleotide polymorphisms, providing conclusive evidence that the insert in the BAC clone represents the resistance haplotype.

A major limitation of these data is that they do not provide conclusive evidence about the contribution of the genes to powdery mildew resistance. Further limitations are that the BAC81D11 insert sequence generated by ONT method is less than 100% accurate and the gaps

between the previously assembled contigs remain to be unfilled. Nonetheless, this thesis opens opportunities for hypothesis-driven experiments. The evidence that the BAC81D11 insert is of the resistance haplotype and the exact demarcation of the coding sequences enable other researchers to clone these genes, transfer them to powdery mildew-susceptible plants, and test the resistance of the resulting transgenic plants. Furthermore, this research contributes to the methodology of third-generation DNA sequencing in that it presents an advanced ONT library preparation protocol which can be used by other researchers sequencing BAC libraries.