

N. SULTANA, 2018



T.R.

NIĞDE ÖMER HALİSDEMİR UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
AGRICULTURAL GENETIC ENGINEERING DEPARTMENT

DOCTOR OF PHILOSOPHY THESIS

BIOINFORMATICS AND MOLECULAR CHARACTERIZATION OF TANDEMLY
ORGANIZED REPETITIVE DNA FAMILY IN Highbush BLUEBERRY
(*VACCINIUM CORYMBOSUM* L.) CULTIVAR 'JUBILEE' GENOME

GRADUATE SCHOOL OF NATURAL AND APPLIED
SCIENCES OF NIĞDE ÖMER HALİSDEMİR UNIVERSITY

NUSRAT SULTANA

July 2018

T.R.
NİĞDE ÖMER HALİSDEMİR UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
AGRICULTURAL GENETIC ENGINEERING DEPARTMENT

BIOINFORMATICS AND MOLECULAR CHARACTERIZATION OF TANDEMLY
ORGANIZED REPETITIVE DNA FAMILY IN Highbush BLUEBERRY
(*VACCINIUM CORYMBOSUM* L.) CULTIVAR 'JUBILEE' GENOME


NUSRAT SULTANA


Doctor of Philosophy Thesis


Supervisor
Prof. Dr. Sedat SERÇE

July 2018

The study titled “**Bioinformatics and Molecular Characterization of Tandemly Organized Repetitive DNA Family in Highbush Blueberry (*Vaccinium corymbosum* L.) Cultivar ‘Jubilee’ Genome**” and presented by Nusrat SULTANA with the help of supervisor **Prof. Dr. Sedat SERÇE**, has been accepted as Doctoral thesis by the jury at the **Department of Agricultural Genetic Engineering** of Niğde Ömer Halisdemir University Graduate School of Natural and Applied Sciences.

Head : Prof. Dr. Sedat SERÇE 
Niğde Ömer Halisdemir University

Member : Prof. Dr. Mehmet ARSLAN 
Erciyes University

Member : Dr. Ali Fuat GÖKÇE 
Niğde Ömer Halisdemir University

Member : Assoc. Prof. Dr. Cem Ömer EGESEL 
Çanakkale Onsekiz Mart University

Member : Dr. Emre AKSOY 
Niğde Ömer Halisdemir University

CONFIRMATION:

This thesis has been found appropriate at the date of/...../..... by the jury mentioned above who have been designated by Board of Directors of Graduate School of Natural and Applied Sciences and has been confirmed with the resolution of Board of Directors dated/...../..... and numbered.....

...../...../20...

Assoc. Prof. Dr. Murat BARUT
DIRECTOR

DECLARATION OF THESIS

I declare that all the information in the thesis is presented maintaining the framework of scientific and academic rules. I also announce that this document is prepared in accordance with the thesis writing rules. The sources of all kinds of expressions and information not original to this dissertation are fully cited.

Nusrat Sultana



ÖZET

'JUBILEE' YÜKSEKÇALI MAVİYEMİŞ (*VACCINIUM CORYMBOSUM* L.) GENOMUNDA ARDIŞIK ORGANİZE TEKRARLANAN DNA AİLESİNİN BİYOİNFORMATİK VE MOLEKÜLER KARAKTERİZASYONU

SULTANA, Nusrat
Niğde Ömer Halisdemir Üniversitesi
Fen Bilimleri Enstitüsü
Tarımsal Genetik Mühendisliği Anabilim Dalı

Danışman : Prof. Dr. Sedat SERÇE

Temmuz 2018, 164 Sayfa

Vaccinium, taksonomik olarak geniş çeşitlilik içeren bir cinistir. Altcinsler ve türler çoğunlukla tür içi ve türler arası düzeylerde geniş poliploidizasyona eğilimlidir. Bu çalışmada, *Vaccinium* genom analizi, ploidi seviye tahmini ve RepeatExplorer yazılımı kullanılarak, tekrarlayan DNA analizi ile gerçekleştirilmiştir. Tekrarlanan dizilerin toplam yüzdesinin, maviyemiş (*Vaccinium corymbosum* L.) ve turnayemişi genomunda (*Vaccinium macrocarpon* Ait.) sırasıyla %80 ve %90 olduğu bulunmuştur. Tekrarların dizilerin büyük çoğunluğu transpoze edilebilir elementlere, uydu DNA'ya ve rDNA dizilerine aittir. *Vaccinium* genomundan toplam altı farklı uydu DNA, 28 tam uzunlukta Ty3/gypsy ve 26 Ty1/copia LTR elementi tanımlanmıştır. Hem uydu DNA hem de LTR elemanlarının iç yapısal özellikleri karakterize edilmiştir. *Vaccinium* cinsinden beş farklı altcins (*Cyanococcus*, *Oxycoccus*, *Myrtillus*, *Hemimyrtillus* ve *Vaccinium*) alınan uydu DNA klonlanması ve moleküler karakterizasyonu, VaccSat1, 5 ve 6'nın bazı türlerde özel olduğunu gösterirken, VaccSat2, 3 ve 7'nin çalışılan tüm türlerde ortak uydu aileleri olduğu belirlenmiştir. Bu araştırmadan elde edilen sonuçlar VaccSat1, 5 ve 6, tür tanımlaması için bir işaretçi olarak önerilebileceğini göstermektedir. Bu tez kapsamında elde edilen sonuçlar *Vaccinium* genom yapısı ve evrimi konusunda önemli katkılar sunmaktadır.

Anahtar Sözcükler: *Vaccinium*, repeatexplorer, LTR retrotransposon, satellit DNA, filogenetik.

SUMMARY

BIOINFORMATICS AND MOLECULAR CHARACTERIZATION OF TANDEMLY ORGANIZED REPETITIVE DNA FAMILY IN Highbush Blueberry (*Vaccinium corymbosum* L.) Cultivar ‘Jubilee’ Genome

SULTANA, Nusrat

Niğde Ömer Halisdemir University

Graduate School of Natural and Applied Sciences

Agricultural Genetic Engineering Department

Supervisor : Prof. Dr. Sedat SERÇE

July 2018, 164 pages

Vaccinium is a taxonomically diverse genus. Subgenus and species are mostly prone to extensive polyploidization within intra and interspecies level. In this study, *Vaccinium* genome analysis has been performed through ploidy level estimation and repetitive DNA exploration using RepeatExplorer software. It is found that total percentage of repetitive sequences is 80 and 90% in blueberry (*Vaccinium corymbosum* L.) and cranberry genome (*Vaccinium macrocarpon* Ait.), respectively. The main portion of which belong to transposable elements, satellite repeats and rDNA sequences. A total of six different satellite families, 28 full-length Ty3/gypsy and 26 Ty1/copia LTR-retrotransposon are identified from the *Vaccinium* genome. Internal structural features of both satellite repeats and LTR retrotransposons are characterized. Cloning and molecular characterization of satellite repeats from five different sections (*Cyanococcus*, *Oxycoccus*, *Myrtillus*, *Hemimyrtilus* and *Vaccinium*) of the genus *Vaccinium* revealed that VaccSat1, 5 and 6 have some levels of species specificity, while VaccSat2, 3 and 7 are common satellite families in all studied species. The results from this investigation suggest VaccSat1, 5 and 6 can be used as markers for species identification. Overall, the results from this research expand the limited knowledge about *Vaccinium* genome structure and evolution.

Keywords: *Vaccinium*, repeatexplorer, LTR retrotransposon, satellite dna, phylogenetics

ACKNOWLEDGEMENT

Life is a journey of learning and gaining experience. Some of the memories are full of happiness but some are painful. All together this journey is all about symbolic graduation for the next phase of life.

First of all, I would like to express my immense gratitude toward my supervisor Prof. Dr. Sedat SERÇE, Niğde Ömer Halisdemir University for his continuous support, guidance and help throughout the time I have been working with him, especially for the moment when I was about to lose through the process. I am also thankful to other faculty members of my our Department for helping me through this Journey. I am very much thankful to Prof. Dr. Thomas Schmidt, Technische Universität Dresden, for accepting me as Erasmus research fellow for short time in his lab where I started to learn the basic tools of bioinformatics characterization of repetitive DNA sequence. I am also thankful to Dr. Gerhard Menzel and Dr. Tony Heitkam of Technische Universität Dresden from whom I learn the skill to perform experiment and data analysis. I am also thankful to all the lab member of Prof. Dr. Thomas Schmidt lab for guiding me and supporting me all the time I needed.

I declare my gratefulness for financial support from the Scientific and Technological Research Council of Turkey (TUBITAK) for TUBITAK-2215 Graduate Scholarship Program and Scientific Research Projects Unit (BAP) of Niğde Ömer Halisdemir University for BAP provided FEB 2017/18 DOKTEP project for conducting research work.

I am also thankful to Prof. Dr. Çiğdem Ulubaş SERÇE for her friendly support while working in her lab. I am thankful to my friend Mehtap Vural and all the well-wishers from Niğde (a small city of Turkey) with their kind word, behaviors, and making my time here like my home country Bangladesh.

I am deeply grateful to my mother and father for their patience and support from the day I born. I am also thankful to my brother who was always beside me while I need financial support during a hard time. I am also thankful for my two little sisters whose unconditional love makes my world colorful.

TABLE OF CONTENT

ÖZET	iv
SUMMARY	v
ACKNOWLEDGEMENT	vi
TABLE OF CONTENT	vii
LIST OF TABLES	x
LIST OF FIGURES	xii
ABBREVIATIONS	xvi
CHAPTER I INTRODUCTION.....	1
1.1 Genome, Genome Size and Repetitive Sequence	1
1.2 Types, Distribution, Functional Role and Regulation of Repeat Sequence.....	3
1.2.1 Tandem repeats	4
1.2.2 Dispersed repeats	7
1.2.3 Regulation of repeat sequence in the genome.....	9
1.3 Origin and Evolution of Repeat Sequence.....	11
1.3.1 Origin and evolution of satellite repeats	11
1.3.2 Origin and evolution of transposable elements.....	12
1.4 Thesis Intent.....	13
1.5 Thesis Outline	14
CHAPTER II LITERATURE REVIEW	16
2.1 Taxonomy of the Genus <i>Vaccinium</i>	16
2.2 Habitat and Geographical Distribution	17
2.3 Economic Importance	18
2.4 Evolutionary Species Dynamics of the Genus <i>Vaccinium</i>	19
2.5 Cultivation of <i>Vaccinium</i>	21
2.6 Overall Progress in Genetics and Genomics of <i>Vaccinium</i>	22
CHAPTER III PLOIDY LEVEL ESTIMATION OF CULTIVATED AND WILD <i>VACCINIUM</i> SPECIES.....	25

3.1 Background Information.....	25
3.2 Materials and Methods.....	26
3.2.1 Collection and identification of Turkish wild <i>Vaccinium</i> species	26
3.2.2. Chromosome preparation and chromosome counting	32
3.2.3 Handling and performing flow cytometry	33
3.3 Results.....	34
CHAPTER IV BIOINFORMATICS ANALYSES AND	
CHARACTERIZATION OF REPETITIVE ELEMENTS OF <i>VACCINIUM</i>	
GENOME	
	39
4.1 Background Information.....	39
4.2 Materials and Methods.....	40
4.2.1 Publicly available genomic database of <i>Vaccinium</i>	41
4.2.2 Illumina paired-end sequence read preparation	42
4.2.3 Clustering and analysis of clusters generated through RepeatExplorer	
pipeline.....	42
4.2.4 Identification and phylogenetic analysis of satellite clusters.....	44
4.2.5 Phylogenetic and heterogeneity analysis of reverse transcriptase protein	
domain sequences.....	45
4.3 Results.....	46
4.3.1 Genomic proportion of different types of repetitive DNA	46
4.3.2 Characteristics features of satellite DNA.....	52
4.3.4 Transposable elements	54
4.3.5 Comparative phylogenetic and heterogeneity analysis of	
retrotransposon elements from blueberry and cranberry	56
4.3.6 Comparative clustering	59
CHAPTER V CLONING AND MOLECULAR CHARACTERIZATION OF	
MAJOR SATELLITE REPEATS FROM CULTIVATED AND TURKISH	
WILD <i>VACCINIUM</i> SPECIES.....	
	62
5.1 Background Information.....	62
5.2 Materials and Methods.....	63

5.2.1 Collection of plant material and genomic DNA extraction	63
5.2.2 Repeat specific primer designing	64
5.2.3 PCR amplification and gel electrophoresis	67
5.2.4 Ligation of PCR product with pGEMT-easy vector system	67
5.2.5 Preparation of XL1-blue <i>E. coli</i> electrocompetent cell.....	67
5.2.6 Transformation of XL1-blue electrocompetent bacterial cells with ligation product	68
5.2.7 Plasmid extraction and positive clone selection for sequencing.....	68
5.2.8 Sequencing of plasmid	70
5.2.9 Computerized sequence analysis.....	71
5.3 Results.....	74
5.3.1 DNA fingerprinting of cultivated and wild <i>Vaccinium</i> species with seven satellite specific primer	74
5.3.2 Diversity and evolution of satellite repeats among <i>Vaccinium</i> species	77
CHAPTER VI OVERVIEW OF MAJOR LTR-RETROTRANSPOSON	
LANDSCAPE IN <i>V. MACROCARPON</i> GENOME.....	112
6.1 Background Information.....	112
6.2 Materials and Methods.....	113
6.3 Results.....	115
CHAPTER VII DISCUSSION	134
CHAPTER VIII CONCLUSION	142
REFERENCES	143
CURRICULUM VITAE.....	163

LIST OF TABLES

Table 1.1. Differences in genome sizes, number of annotated genes, and percentage of repetitive sequences in closely related cultivated plant species of Poaceae family.	2
Table 1.2. Classification of abundant TEs in plants according to Wicker et al. (2007).. .	8
Table 2.1. DNA ratio estimated for some important <i>Vaccinium</i> species compared to <i>Vinca minor</i> L. and their origin by Hummer et al. (2015). (x) measurement not available.	22
Table 3.1. List of plant materials used in this study.	26
Table 3.2. Name of the identified wild <i>Vaccinium</i> species and their GPS coordinates..	27
Table 3.3. Chromosome number, FL1 UV LED mean value and Coefficient of Variation (CV%) of flow cytometry reading of <i>Vaccinium</i> species. (x) measurement failed	38
Table 4.1. Available public genome (DNA) database for <i>V. corymbosum</i> and <i>V. macrocarpon</i> (The genome database for <i>Vaccinium</i> (GDV), Source://www.vaccinium.org)	41
Table 4.2. Repeat Explorer based estimation of different repeat types in <i>V. corymbosum</i> and <i>V. macrocarpon</i> . (Standard deviation is calculated from three different individual RepeatExplorer based estimation)	47
Table 4.3. Characteristics features of the satellite repeats annotated in blueberry and cranberry genome	53
Table 4.4. Retrotransposon protein domain identified from the assembled genome and the sequence similarity of <i>V. corymbosum</i> and <i>V. macrocarpon</i>	55
Table 4.5. Calculation of heterogeneity among identified reverse transcriptase protein domain sequences in <i>V. corymbosum</i> and <i>V. macrocarpon</i>	58
Table 5.1. Features of primer pairs for repeat specific satellite amplification.	65
Table 5.2. Short name for each plant species	72
Table 5.3. Plasmid clone characteristics of <i>Vaccinium</i> satellite repeats.....	79
Table 5.4. Summary of satellite repeat structure in the genome of <i>Vaccinium</i>	96

Table 5.5. Genetic distance among different <i>Vaccinium</i> species based on the consensus monomer sequence of VaccSat1-7. The lowest values were colored with yellow and the highest values were with cyan.....	108
Table 5.6. Overview of satellite repeat diversity in <i>Vaccinium</i> genome.	111
Table 6.1. Cluster, supercluster, genome proportion, structural protein domain, PBS and PPT features identified in Class_I LTR Ty1/ <i>copia</i> retrotransposon in <i>V. macrocarpon</i> (GAG = <i>Gag</i> domain, PROT = Protease, RT = Reverse transcriptase domain, RH = RNase H domain, INT = Integrase domain).	117
Table 6.2. Pairwise similarity distance matrix of LTR regions of identified full length Ty1/ <i>copia</i> element in <i>V. macrocarpon</i> genome. The lowest values were colored with cyan and the highest values were with yellow.....	121
Table 6.3. Pairwise similarity distance matrix of RT regions of identified full length Ty1/ <i>copia</i> element in <i>V. macrocarpon</i> genome. The lowest values were colored with cyan and the highest values were with yellow.....	122
Table 6.4. Cluster, supercluster, genome proportion, structural protein domain feature, sequence information of PBS and PPT of Ty3/ <i>gypsy</i> retrotransposon (GAG = <i>Gag</i> domain, PROT = Protease, RT = Reverse transcriptase domain, RH = RNase H domain, a RH = addition RNase H domain, INT = Integrase domain, CHDII = Chromodomain II).....	125
Table 6.5. Pairwise similarity distance matrix of LTR regions of identified full length Ty3/ <i>gypsy</i> element in <i>V. macrocarpon</i> genome. The lowest values were colored with cyan and the highest values were with yellow.....	130
Table 6.6. Pairwise similarity distance matrix of RT regions of identified full length Ty3/ <i>gypsy</i> element in <i>V. macrocarpon</i> genome. The lowest values were colored with cyan and the highest values were with yellow.....	131
Table 6.7. General structural features and repeat compositions of full length LTR-retrotransposons	133

LIST OF FIGURES

- Figure 1.1. A model of a typical plant chromosome showing the characteristic genomic distributions of different classes of repetitive DNA (different color represent different types of repetitive sequence where white represents the genic sequence) (Figure adapted from [Garrido-Ramos, 2015](#)) 4
- Figure 1.2. Interaction between transposable elements of genome and environment (Model is adapted and modified from [Biémont and Vieira, 2006](#))..... 9
- Figure 1.3. The role of small RNAs on regulation of transposable element activity in the plant genome mainly by producing heterochromatic siRNAs (hc-siRNA) (The model is modified and adapted from [Wendel et al., 2016; Ito, 2012](#)) 10
- Figure 1.4. General model of satellite repeat evolution in plant. (Yellow, orange and green are three mutated monomer unit arise from the ancestral monomer unit) (Figure redrawn from [Garrido-Ramos, 2015](#))..... 12
- Figure 1.5. Genetic relationship among major classes of transposable element based on conserved domains or modules (ITRs = inverted terminal repeats; Trp = transposase domain; GAG = group specific antigen protein; LTRs = Long terminal repeats; PROT = protease; RT = reverse transcriptase domain; INT = integrase domain; RH = RNase H; env = envelope protein) (Modified from [Capy and Maisonhaute, 2002; Biémont and Vieira, 2006; Ustyantsev et al., 2017](#)). 13
- Figure 2.1. Phylogenetic relationship between five well-known and economically important *Vaccinium* species from concatenated sequence data of chloroplast MaturaseK gene (*matK* gene) and nuclear ribosomal internal transcribed spacer gene (nrITS gene) using maximum parsimony PAUP* method ([Swofford, 2002](#)). (Phylogenetic dendrogram was recalculated using the sequence data and parameters provided in [Powell and Kron, \(2002\)](#)). 20

Figure 3.1. Morphological diversity of five different <i>Vaccinium</i> species at the fruiting stage. Green house grown cultivated species (a) and wild plants from Kaçkar Mountains (b, c, d, e). [Photographs are copyright September, 2017, by Nusrat Sultana (a) and by Prof. Dr. SEDAT SERÇE (b, c, d, e)].	30
Figure 3.2. Difference in leaf, fruit and stem morphology of the studied <i>Vaccinium</i> species	31
Figure 3.3. Estimation of chromosome number of the studied <i>Vaccinium</i> species. <i>V. corymbosum</i> (cultivar ‘Jubilee’) ($2n = 4x = 48$) (a), <i>V. arctostaphylos</i> (round fruit) ($2n = 4x = 48$) (b), <i>V. arctostaphylos</i> (elongated fruit) ($2n = 4x = 48$) (c), <i>V. myrtillus</i> ($2n = 2x = 24$) (d), <i>V. uliginosum</i> ($2n = 2x = 24$) (e).	36
Figure 3.4. Estimation of ploidy level among studied <i>Vaccinium</i> species using flow cytometry. Peak of <i>V. corymbosum</i> (cultivar ‘Jubilee’) (a), <i>V. arctostaphylos</i> (elongated fruit) (b), <i>V. arctostaphylos</i> (round fruit) (c), <i>V. myrtillus</i> (d), <i>V. uliginosum</i> (e), a-e (f)	37
Figure 4.1. Genome coverage vs repeat proportion identified in the <i>V. corymbosum</i> genome.	43
Figure 4.2. Quantification of different types of repeat sequence from <i>Vaccinium</i> genome	50
Figure 4.3. Supercluster of different types of repeat sequence identified in <i>Vaccinium</i> genome.	51
Figure 4.4. Star typical satellite graph identified in <i>Vaccinium</i> genome.	52
Figure 4.5. Phylogenetic relationship among seven different satellite repeats identified in <i>V. corymbosum</i> and <i>V. macrocarpon</i> genome	54
Figure 4.6. Phylogenetic relationship among reverse transcriptase domain sequence of different lineage of Ty3/gypsy (a), Ty1/Copia (b), LINE (c) in <i>V. corymbosum</i> and <i>V. macrocarpon</i> genome.	57
Figure 4.7. Differences in the heterogeneity of different lineage of reverse transcriptase protein domain sequences in <i>V. corymbosum</i> and <i>V. macrocarpon</i>	59
Figure 4.8. Comparative clustering of repeat sequence of <i>V. corymbosum</i> and <i>V. macrocarpon</i> showing the genome enriched repeat sequence in <i>Vaccinium</i> genome. (Color red, yellow and blue represent the satellite repeat VaccSat 1, VaccSat5 and VaccSat6, respectively)	61
Figure 5.1. Extracted genomic DNA from the <i>Vaccinium</i> species. Lane M: 1 kb DNA Marker (ThermoScientific); Lane 1: <i>V. corymbosum</i> cultivar ‘Jubilee’; Lane	

2: <i>V. corymbosum</i> , cultivar ‘Misty’; Lane 3: <i>V. arctostaphylos</i> - elongated fruit; Lane 4: <i>V. arctostaphylos</i> - round shaped fruit; Lane 5: <i>V. myrtillos</i> ; Lane 6: <i>V. uliginosum</i>	64
Figure 5.2. Satellite repeat sequence and primer designing site (black arrow).	66
Figure 5.3. Representative photo for plasmid extraction and positive clone selection for VaccSat1, VaccSat2 and VaccSat3 from <i>V. corymbosum</i> cultivar ‘Jubilee’. Lane 1 and 22: 1 kb DNA Marker (ThermoScientific), Lane 2-11 VaccSat1, Lane 12-21: VaccSat 2 and Lane 23-32: VaccSat3.	69
Figure 5.4. Representative photo for PCR amplification from extracted plasmid using satellite specific primer of VaccSat5 in three different <i>Vaccinium</i> species (expected band size was 91 bp for trimer and 199 bp for hexamer). Lane 1 and 26: 1 kb DNA Marker (ThermoScientific), Lane 2-17: <i>V. arctostapylos</i> , Lane 18-34 (except 26): <i>V. myrtillos</i> and Lane 35-50: <i>V. uliginosum</i>	70
Figure 5.5. DNA fingerprinting of <i>Vaccinium</i> species with identified satellite repeats. 76	
Figure 5.6. Phylogenetic dendrogram from 149 monomer sequences from seven different satellite families of six studied <i>Vaccinium</i> species with maximum likelihood algorithm using PhyML V3.1 (Guindon et al., 2010). (For monomer name Table 5.3. was referred).	84
Figure 5.7. Dotplot analysis of multimer of <i>Vaccinium</i> satellite repeat clone. VaccSat1 (a), VaccSat2 (b), VaccSat5 (c) and VaccSat6 (d). Monomer unit and subunit are showed as parallel line on the dotplot diagram (For clone number Table 5.3. referred)	91
Figure 5.8. Subunit and higher order repeat unit structure of multimer clone <i>Vaccinium</i> satellite. VaccSat1 clone VC-3030 (a), VaccSat2 clone VM-15 (b), VaccSat5 clone VA-46 (c) VaccSat6 clone VA-31 (d). (For clone number Table 5.3. referred)	95
Figure 5.9. Phylogenetic dendrogram of extracted monomer sequences from 6 different <i>Vaccinium</i> satellite using FastTree 2.1.5 (Price et al., 2010). VaccSat1- VaccSat7 were depicted in (a)-(f). (Referred Table 5.3. for monomer name) Major clusters are colored and minor clusters are given in black	97
Figure 5.10. Multiple sequence alignment of consensus monomer sequence using MAFT multiple sequence alignment tool from six <i>Vaccinium</i> satellite. VaccSat1- VaccSat7 are depicted in (a)-(f). (For consensus monomer sequence name,	

Table 5.3. is referred). Homogeneous region of the multiple alignments is shaded in black and polymorphic site is given with no shading.....	107
Figure 5.11. Heatmap analysis of monomer unit of six <i>Vaccinium</i> satellite. (Pairwise similarity distance heatmap of extracted monomer unit. VaccSat1-VaccSat7 are depicted in (a)-(f) (Scale bar represent the calculated distance matrix value)	110
Figure 6.1. In depth structural diversity analysis of Ty1/ <i>copia</i> retrotransposon in <i>V. macrocarpon</i>	119
Figure 6.2. In depth structural diversity analysis of Ty3/ <i>gypsy</i> retrotransposon in <i>V. macrocarpon</i>	127



ABBREVIATIONS

Abbreviations	Description
BLAST	Basic local alignment search tool
bp	Base pair
CENH3	Centromeric histone H3
CL	Cluster
CV	Coefficient of variation
DAPI	4',6-Diamidine-2-phenylindole
ddNTP	di-deoxynucleoside triphosphate
DNA	Deoxyribonucleic acid
dNTP	deoxynucleoside triphosphate
dsRNA	Double-stranded RNA
dUTP	Deoxyuridine triphosphate
EDTA	Ethylendiaminetetra acetic acid
GDV	The genome database for <i>Vaccinium</i> (GDV)
GPS	The Global Positioning System
h	Hour
hc-siRNA	Heterochromatic small interfering RNA
IPTG	Isopropyl- β -D-thiogalactopyranoside
kbp	Kilobase pair
LB	Luria-Bertani medium
LTR	Long terminal repeat
Mbp	Megabase pair
min	Minute
mM	Milli molar
NCBI	The national center for biotechnology information
nrITS	nuclear ribosomal internal transcribed spacer
OD	Optical density
PAUP	Phylogenetic analysis using parsimony

PCR	Polymerase chain reaction
pM	Picomolar
PVP	Polyvinyl pyrrollidone
rDNA	Ribosomal DNA
RDR	RNA-dependent RNA polymerase (RDR)
RFLP	Restriction fragment length polymorphysm
RNA	Ribonucleic acid
RNase I	Ribonuclease I
Rpm	Rounds per minute
RT	Room temperature
RT domain	Reverse transcriptase domain
SatDNA	Satellite DNA
SCL	Supercluster
siRNA	Small interfering ribonucleic acid
SSRs	Simple sequence repeats
TE	Transposable element
TE buffer	Tris EDTA buffer
Tris	Tris (hydroxymethyl) aminomethane
VaccSat	<i>Vaccinium</i> satellite repeats
X-Gal	5-Bromo-4-chloro-3-indolyl- β -D-Galactopyranoside

CHAPTER I

INTRODUCTION

1.1 Genome, Genome Size and Repetitive Sequence

Genome is the complete set of genetic information that organisms require to function. For higher organisms, this information is stored in the long DNA sequences (Bennetzen, 1996). The main component of the genome is actually thousands of genes that code for functional proteins and some repetitive sequences or noncoding sequences that act as the regulatory regions or buffering regions of the genes. Although the genic region is well-studied region of the genome, knowledge about the repetitive sequence is not comprehensively studied for most of the organisms and even not available for some of the species. Therefore, the repetitive sequence has several names like junk DNA or dark matters of a genome (Vicient and Casacuberta, 2017; Van de Peer et al., 2017)

Although genome size is a constant value representing for a particular species, it does not always consistent across the phylogenetic taxa which is also termed as C-value enigma (Table 1.1.). It is well hypothesized that species evolution and genome size variation in plant kingdom is the product of repeated cyclical episodic duplication and lineages specific expansion or loss of repetitive sequences (Wendel et al., 2016). Therefore, several biological topics like evolution and molecular taxonomy are solely based on the analysis of the diversification, abundance and differential organization of repetitive sequence (Bennett and Leitch, 2005; Besse, 2014).

Table 1.1. Differences in genome sizes, number of annotated genes, and percentage of repetitive sequences in closely related cultivated plant species of Poaceae family.

Species name	Common name	Genome size (Mb)	Number of annotated gene	Percentage of transposon/ repeats	Reference
<i>Oryza sativa</i> L.	Rice	389	37,544	35	International Rice Genome Sequencing Project, 2005
<i>Oryza glaberrima</i> Steud.	African rice	358	33,164	34.3	Wang et al., 2014
<i>Hordeum vulgare</i> L.	Barley	5100	26,159	84	International Barley Genome Sequencing Consortium, 2012
<i>Triticum aestivum</i> L.	Wheat	17,000	124,201	76.6	International Wheat Genome Sequencing Consortium, 2014
<i>Zea mays</i> L.	Maize	2500	32,540	85	Schnable et al., 2009
<i>Sorghum bicolor</i> L.	Sorghum	730	34,496	61	Paterson et al., 2009
<i>Setaria italica</i> L.	Foxtail millet	490	38,801	46	Zhang et al., 2012

1.2 Types, Distribution, Functional Role and Regulation of Repeat Sequence

Repetitive DNA sequence families, the major components of plant genomes can account for up to 85% of the nuclear DNA (Schmidt and Heslop-Harrison, 1998). Repeats are mostly sequence motifs ranged from dinucleotide to several thousand of base pairs, repeated with variable copy number occurred between several hundred or hundreds of thousands of times in the genome. Depending on their mode of distribution throughout the genome, they can be divided into tandemly arranged or dispersed sequences grossly responsible for the genome organization (Heslop-Harrison, 1991; Jagannathan et al., 2018) (Figure 1.1.).

Tandemly organized repeats are repeated motives that accumulate in a certain region of the chromosome. In contrast, dispersed repetitive DNA sequences are scattered throughout the genome or interspersed within other sequences. Both types of repetitive sequences are prone to rapid changes over an evolutionary time period and thought to be a prime reason of speciation (Heslop-Harrison and Schwarzacher, 2011). Therefore, comparative studies of plant repetitive sequences are efficient tools to investigate the evolutionary relationships among closely related species.

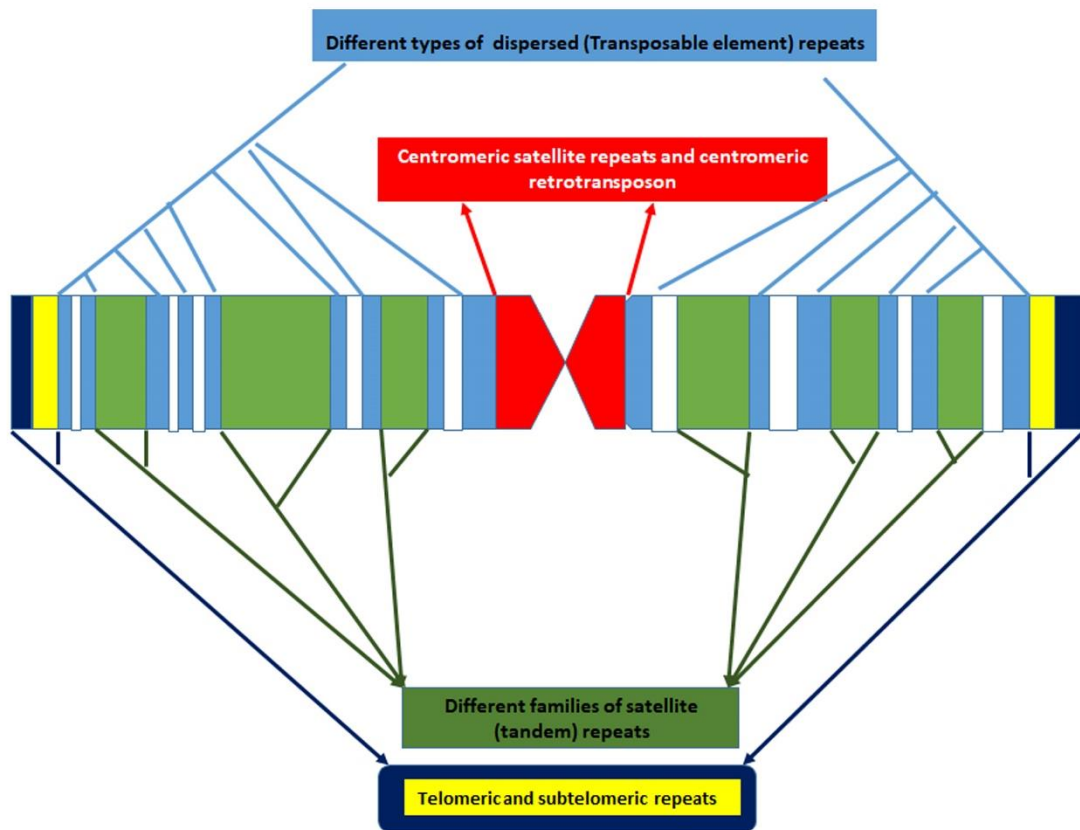


Figure 1.1. A model of a typical plant chromosome showing the characteristic genomic distributions of different classes of repetitive DNAs (different colors represent different types of repetitive sequences where white represents the genic sequences) (Figure is adapted from Garrido-Ramos, 2015)

1.2.1 Tandem repeats

The main tandem repeats groups are categorized as satellite repeats, microsatellites, minisatellites, rDNA sequences and telomeric repeat sequences.

Satellites, microsatellites and minisatellites

Satellites, microsatellites and minisatellites are different based on the size of the repeating motif. While microsatellites or SSRs (simple sequence repeat sequences) motifs are ranged between 2-6 nucleotides, minisatellites motifs ranged from 10-40 nucleotides and satellite repeats are those motifs that are longer than minisatellites (Schmidt and Heslop-Harrison, 1998; Robledillo et al., 2018). Some of these repeats could localize in certain

regions of the chromosome like centromere, pericentromere, subtelomere, intercalary region while others could be dispersed. For instance, most of the simple sequence repeats are often found evenly distributed throughout the chromosome except the rDNA site and centromeric region (Kubis et al., 1998; Kirov et al., 2017). While the satellite repeats are mostly found in the functional region of the chromosome.

Although the monomer unit length of satellite repeats are highly variable, functional and preferential centromere motif sizes range from 150-180 bp to 320-360 bp. These specific size range are essential for the alignment with the size of single nucleosome and shown to interact with centromere-specific histone H3 variants (Iwata et al., 2013; Zhang et al., 2014; He et al., 2015). Although the mechanism involved in the interaction of centromeric satDNA and centromere-specific histone H3 variants is still a matter of investigation, evidence suggest that this is often epigenetically regulated. Moreover, it is found that centromeric satDNA is often transcribed and helps in the formation of the CENH3 nucleosome, RNA directed heterochromatin, kinetochore assembly, chromosome stability and segregation (Heslop-Harrison and Schwarzacher, 2013). However, satellite repeats can evolve and diversify quite quickly in the evolutionary time scale, they often show species-specific or even chromosome specific variants (Heitkam et al., 2015).

rDNA sequences

rDNA sequences are functional and highly repeated sequences, responsible for the production of rRNA or ribosomal RNA. Although the rDNA gene includes 5S rDNA repeats and 18S-5.8S-25S rDNA repeat genes present as a separate cluster in the genome, exception also found when they localize closely (Garcia et al., 2009; Garcia and Kovařík, 2013; Rosato et al., 2016). Both of the rDNA regions include intergenic spacer region. 18S-5.8S-25S rDNA altogether can encode as 45S rDNA and can occupy over 10kb long region (Xue and Barna, 2012).

Telomeric and subtelomeric repeat sequences

Telomeric repeat sequence is a highly conserved sequence across the eukaryotic organism, mainly the G-rich repeat and 5-15 kilobases long in human genome. Although most common telomere repeating motif in the eukaryotic organism is 7 bp repeat “TTTAGGG”, the array length could be highly variable even in the same cell and it is grossly responsible for the end protection. Telomeric repeats are not synthesized in a semiconservative mode of DNA replication but by a different enzyme called telomerase (Nakamura and Cech, 1998). The evolutionary phylogenetic analysis reveals that telomerase is a reverse transcriptase enzyme, more closely related with non-LTR retrotransposon, and uses an RNA template to synthesize DNA (Fajkus and Zentgraf, 2002; De Lange, 2004).

Telomeric repeats are often closely associated with the subtelomeric repeat sequences (Churikov and Price, 2008). Subtelomeric repeats are mainly tandem repeats in most of the organisms and act as a buffer zone between the telomere and the internal zone of the chromosome (Mizuno et al., 2006). The possible postulated role of the subtelomeric repeat is the faithful chromosome replication, genome stability, and transcriptional regulation of subtelomeric gene and stabilization of chromosome end in the absence of telomeric repeats (Torres et al., 2011).

Other important functional roles of tandem repeats in the genome

Satellite repeats can also be important for the gene regulation by serving as promoter elements, transcription start site, and the binding site for transcription factors, thereby influencing the gene expression pattern directly. Satellite repeats are also found to be accumulated in the sex chromosomes and B chromosome, and thought to be linked with their evolution in plants and animals (Plohl, 2010; Garrido-Ramos, 2017). Evidence showed that certain portion of satellite repeats could transcribe and the transcription of satellite repeats are directly related with different environmental and developmental progression (Shapiro and Von Sternberg, 2005; Ugarkovic, 2005; Steflova et al., 2013; Puterova et al., 2017).

1.2.2 Dispersed repeats

Most of the dispersed repeats in the genome belongs to transposable elements (TEs), the most abundant repetitive sequences. Some of the features of TEs are associated with giant motif size that could reach up to 20 kb, enormous diversity and ability to multiply in the number using the host machinery. Therefore, TEs is often called as “jumping gene” or “selfish DNA” (Biémont and Vieira, 2006; González et al., 2017).

Classifications of TEs based on transposition mechanism and structural features

There are two main classes of TEs found in the higher plants; DNA transposons and retrotransposons (Table 1.2.). While DNA transposons use the host cell machinery to jump from one location of the genome to another, retrotransposons act through an RNA intermediate, synthesized with the help of an enzyme called reverse transcriptase (Wicker et al., 2007; Piégu et al., 2015).

Class-DNA transposon

The protein responsible for transposition of the autonomous DNA transposon through “cut and paste” mechanism are called *transposase* (Biémont and Vieira, 2006; Pray, 2008). Target site for transposase protein are often specific with some exception (Muñoz-López and García-Pérez, 2010). Common types of DNA transposon recorded in angiospermic plants are hAT, CACTA, Mutator, PIF/Harbinger, Tc1/mariner (Feschotte et al., 2002; Piégu et al., 2015).

Class-retrotransposon

Retrotransposons act through “copy and paste” mechanism and the prime significant protein domain signature is reverse transcriptase domain (Biémont and Vieira, 2006; Ustyantsev et al., 2017). Retrotransposon could further be subdivided based on the presence or absence of long terminal repeat (LTR). Those that have a directly repeated LTR region at the two opposite poles of the sequence are termed as LTR-retrotransposon; in contrary those without that specialized structure are called as non-LTR retrotransposon (Feschotte et al., 2002; Piégu et al., 2015). Two common LTR-retrotransposons in higher

plants are Ty3-*gypsy* and Ty1-*copia* (Kamm et al., 1995; Galindo-González et al., 2017). The differences between these two groups are mainly based on the organization of their protein domains. In addition to that, some other retrotransposons often associated with plant genome are known as pararetroviruses. Within non-LTR retrotransposon group, long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs) are well-studied ones. All these sequences have an immense effect on the evolution, adaptation and speciation on not only the land plants but also on other organisms (Schmidt, 1999; Pray, 2008; Galindo-González et al., 2017). Like DNA transposons, retrotransposons also show target site specificity in plants with some exception. For instance, evidence has been found for target site integration for the elements of Chromoviruses, Errantiviruses/Athila, Sireviruses types of Ty3-*gypsy* LTR-retrotransposon in plant (Bousios et al., 2010; Neumann et al., 2011; Wollrab et al., 2012; Weber et al., 2013)

Table 1.2. Classification of abundant TEs in plants according to Wicker et al. (2007).

Class	Order	Superfamily
Class1 (retrotransposons)		
	LTR retrotransposon	<i>Copia, Gypsy</i>
	SINE	<i>tRNA, 7SL, 5S</i>
	LINE	<i>L1</i>
Class 2 (DNA transposons)		
	TIR	<i>hAT, CACTA, Mutator, PIF/Harbinger, Tc1/mariner</i>
	Helitron	<i>Helitron</i>

Transposable elements and evolution of plant genome

Although most of the TEs remain silent in the genome, a certain portion could be activated based on the environmental conditions. This phenomenon called “TE-Thrust” (Oliver et al., 2013). Some of the effects of TE-thrust could be summarized as gene modification, disruption and altered gene expression. TEs could also serve as a regulatory sequence, creating novel genes and thus have a great potential impact on environmental adaptation and evolution (Chuong et al., 2017) (Figure 1.2.)

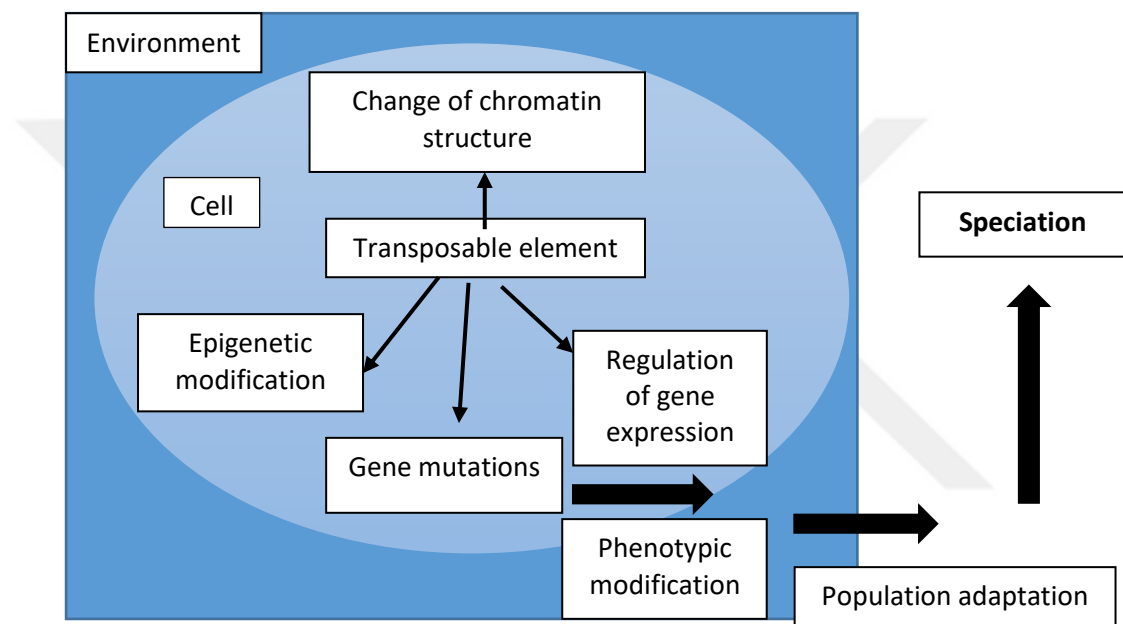


Figure 1.2. Interaction between transposable elements of genome and environment (Model is adapted and modified from Biémont and Vieira, 2006).

1.2.3 Regulation of repeat sequence in the genome

Repeat sequences, especially TEs, could have both beneficial and adverse effect on the genome, thus having a significant influence on genome integrity due to their great power of rapid proliferation and horizontal transposition. Therefore, to control the TEs-activity and to protect the genome from viruses or other exogenous sequences, RNA interference (RNAi) machinery is thought to be involved (Figure 1.3.). Small RNAs act as an important modulator for duplicated, repeated and TE-driven genome expansion. Small RNA sequences, which are highly diversified, could silence or reduce the effect of the

gene or segment of DNA from which it is originated or region of genome with which it has significant similarity with the function in trans (Davidson and Britten, 1979; Wendel et al., 2016).

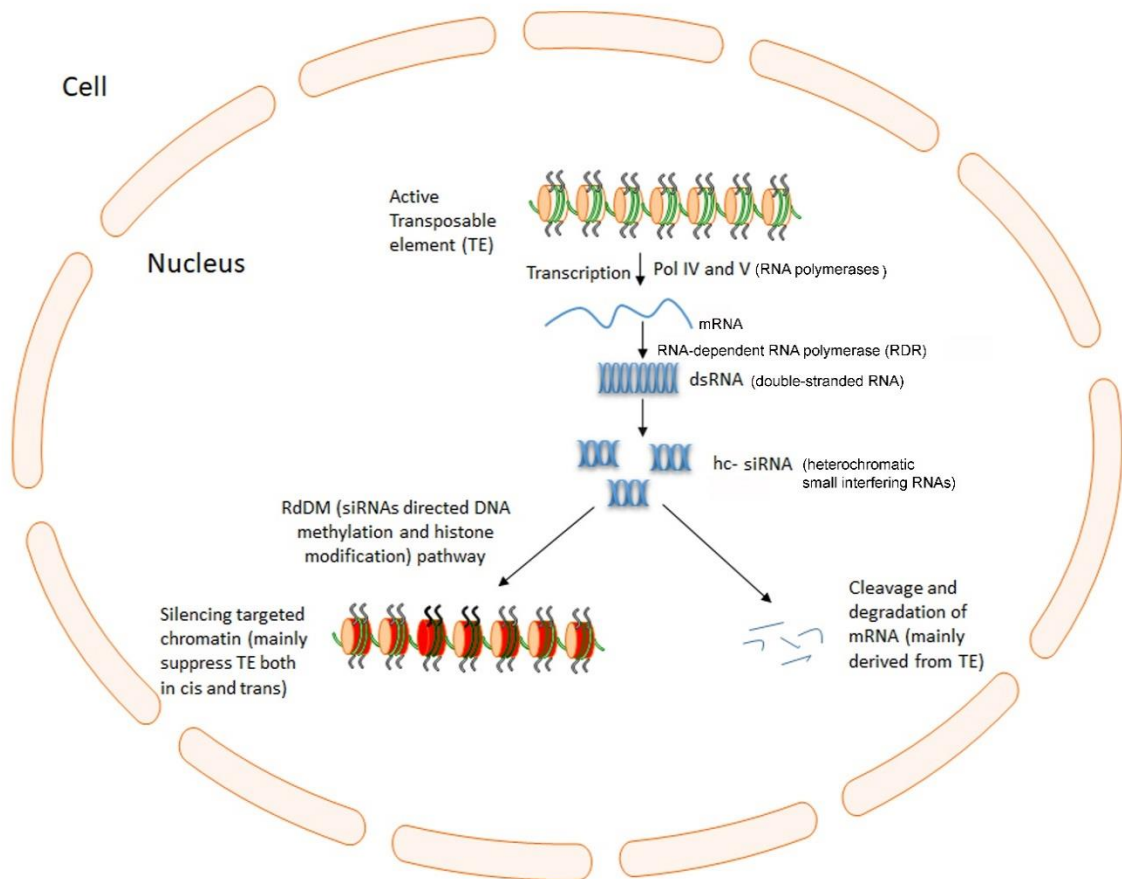


Figure 1.3. The role of small RNAs in regulation of transposable element activity in the plant genome mainly by producing heterochromatic siRNAs (hc-siRNA) (The model is modified and adapted from Ito, 2012; Wendel et al., 2016)

1.3 Origin and Evolution of Repeat Sequences

1.3.1 Origin and evolution of satellite repeats

The origin, amplification, spreading and diversity of satellite repeat within the genome of an organism are still a matter of investigation. Nevertheless, the sequence similarity between the satellite monomer motif and some specific regions of the chromosome (the intergenic spacer of rDNA, transposable element or SSR) suggest the probable origin of the satellite repeats from the respective sequences. The mechanisms involved are thought to be unequal crossing over, strand slippage, rolling circle-based replication of external sequence, transposition, insertion of particular DNA segment and duplication (Mehrotra and Goyal, 2014; Garrido-Ramos, 2017).

“Satellite repeats library” is the total satellite repeats present in an organism and can serve as the resources of satellite repeats for the future generation. For instance, a particular satellite repeat can amplify in some related species but not all. However, it was often explored that common satellite repeat family found in the different taxa are highly homogeneous in intraspecies level and show high interspecific divergence (Garrido-Ramos, 2017). This is also called concerted evolution, where the members of a population have the similar satellite sequence motif. The typical orders followed for this evolution are; first: origin of mutation in a particular satellite motif, second: homogenization of particular types of mutated form in one satellite family of the organism and third: finally fixation. All these steps can be influenced by reproductive mechanism, natural selection and time (Figure 1.4.). For example, species showing reproductive barrier allow gene flow among different taxa. Consequently, evolutionary pattern followed by those species is also termed reticulate evolution when species are not always homogeneous within the population or intraspecific level (Garrido-Ramos, 2015).

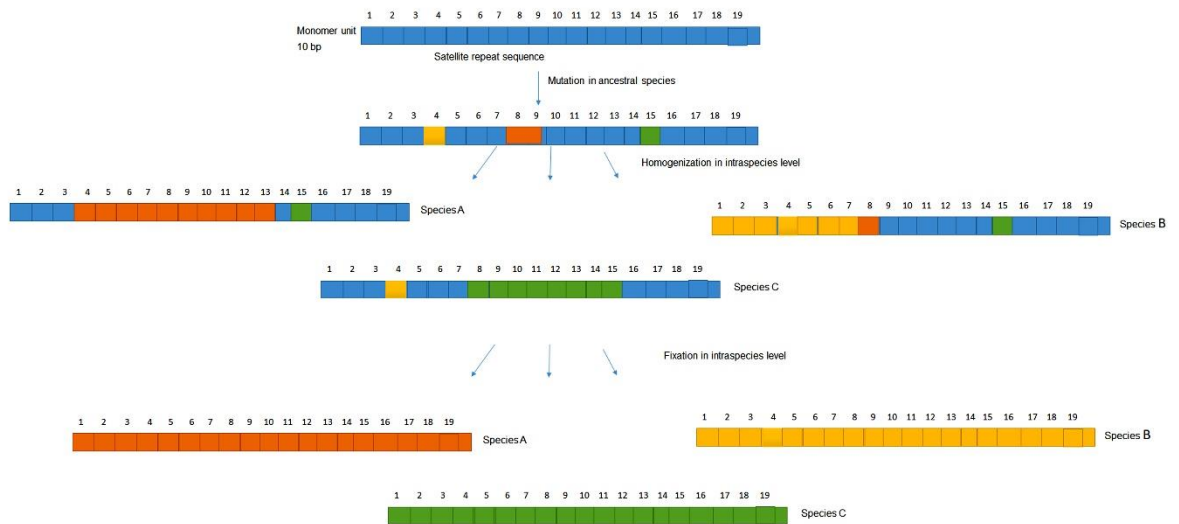


Figure 1.4. General model of satellite repeat evolution in plant. (Yellow, orange and green are three mutated monomer unit arise from the ancestral monomer unit) (Figure redrawn from Garrido-Ramos, 2015)

1.3.2 Origin and evolution of transposable elements

Supporting evidence describing the origin of TEs is quite intriguing and actually does not show a clear direction (Ustyantsev et al., 2017). Nonetheless, as TEs is shared by both prokaryotic and eukaryotic organisms it is often postulated that their origin is from the simplest organism. Moreover, they are also opportunistic sequences and evolve through acquiring or losing different modules (Figure 1.5.). Due to their ability of horizontal transfer (transpose), these sequences showed tremendous fitness ability with different environmental challenges (Bennetzen, 1996; Biémont and Vieira, 2006; Ustyantsev et al., 2017)

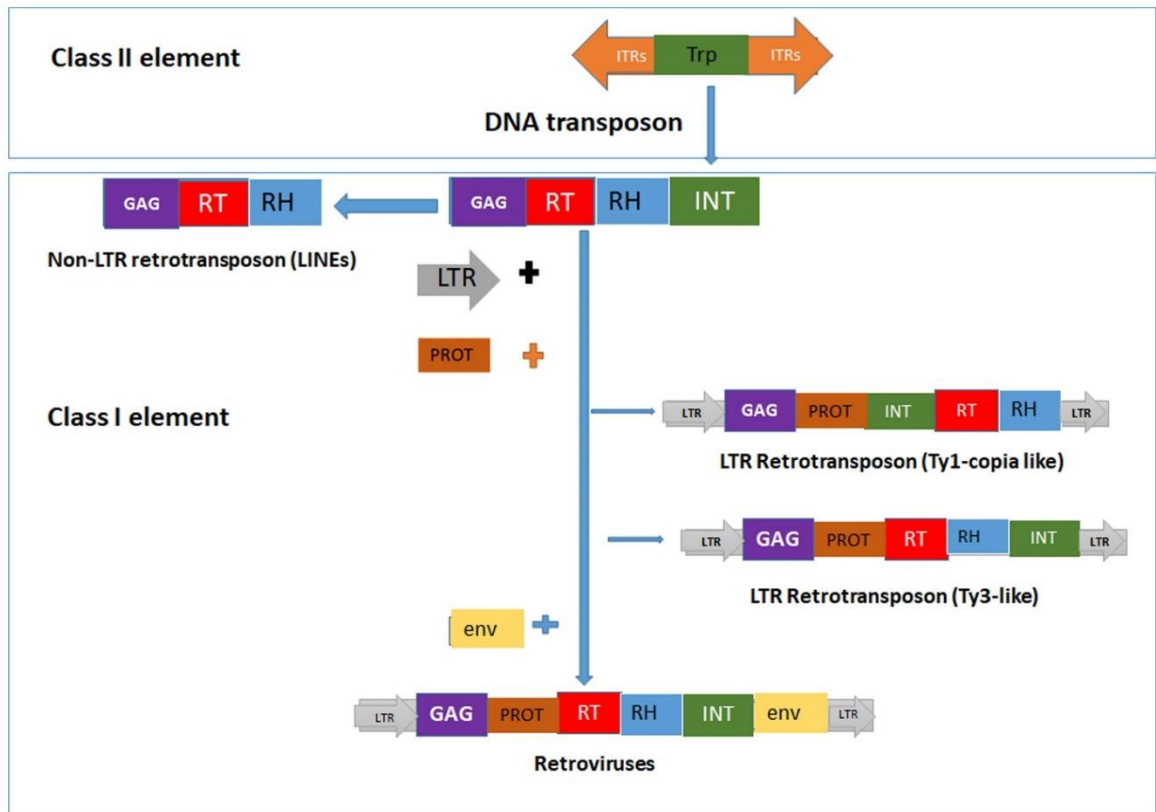


Figure 1.5. Genetic relationship among major classes of transposable element based on conserved domains or modules (ITRs = inverted terminal repeats; Trp = transposase domain; GAG = group specific antigen protein; LTRs = Long terminal repeats; PROT = protease; RT = reverse transcriptase domain; INT = integrase domain; RH = RNase H; env = envelope protein) (Modified from Capy and Maisonhaute, 2002; Biémont and Vieira, 2006; Ustyantsev et al., 2017).

1.4 Thesis Intent

Vaccinium is an economically important small berry fruit species, which has drawn media's attention recently due to its nutritional and medicinal importance. Although genetic and genomic studies on this recently evolved berry fruit have rapidly progressed in the last decades, identification and characterization of blueberry plant species is still a difficult task due to extensive intra and interspecific hybridization. Identification and characterization of the repetitive sequences is an excellent tool to solve this issue because of the diversity of repetitive sequences which is directly correlated with the evolutionary diversification of species complex. Besides, repetitive sequences are currently being used

as an advanced biotechnological tool for artificial gene transformation, an artificial construction of centromere and chromosome, epigenetic studies, genome assembly and so on. However, information about the structure and diversity of repetitive sequences for the genus *Vaccinium* is rather scarce. Therefore, in this study, a serious attempt has been taken to in-depth characterization of repetitive sequences of the genus *Vaccinium*.

Therefore, the main objectives of this study could be summarized as:

1. To understand the ploidy level differences in the Turkish wild *Vaccinium* species,
2. To understand the structure and evolution of different types of repeat sequences and diversity of repeat sequences in the *Vaccinium* species (*V. corymbosum* L. and *V. macrocarpon* Ait.),
3. To understand how satellite repeats interplay with *Vaccinium* species evolution and taxonomy, including cultivated (*V. corymbosum* and *V. macrocarpon*) and Turkish wild species (*V. arctostaphylos* L., *V. myrtillus* L. and *V. uliginosum* L.).

1.5 Thesis Outline

Chapter I presents the overall introduction about types, distribution, abundance, origin, evolution and functional role of different types of repetitive sequences that have been already characterized in different organisms. Introductory chapter also describes why repetitive sequence studies are important for taxonomic identification of the different plant species. Chapter II summarizes the different research studies focusing on taxonomy, habitat and geographical distribution, ploidy level complexity, and species evolution conducted on the genus *Vaccinium*. Chapter II also describes the recent progress on studying genetics and genomics of *Vaccinium*. Chapter III explains the ploidy level of selected *Vaccinium* species like *V. corymbosum*, *V. arctostaphylos*, *V. myrtillus* and *V. uliginosum* through chromosome counting and flow cytometry. Chapter IV deals with the bioinformatics analysis to quantify and identify different types of repetitive sequence for the genus *Vaccinium*, as well as comparative clustering to detect the species-specific repetitive sequences. Chapter V explains the characterization and evolution of different satellite repeats of the genus *Vaccinium* and interplay between satellites repeats diversity

with *Vaccinium* species. Chapter VI deals with the LTR-retrotransposon landscape in cranberry (*V. macrocarpon*) genome.



CHAPTER II

LITERATURE REVIEW

2.1 Taxonomy of the Genus *Vaccinium*

Vaccinium is an economically important genus belonging to Ericaceae family containing about 30 different sections and 450 species (Trehane, 2004; Kloet and Dickinson, 2009). Ericaceae is one of the families of the order Ericales. Other economically important plants of Ericales are tea (*Camellia sinensis* L.) and kiwi (*Actinidia deliciosa* Planchon). Although some species of the genus *Vaccinium* are native to certain regions of the world, others are distributed in the different continents of the world with wider geographical distribution. Origin of this genus is thought to be South America; however, species are distributed in the different regions of North America, Europe, and Asia (Vander Kloet, 1988).

Both morphological and molecular data support that *Vaccinium* is a monophyletic genus. This means having a common single ancestor and separated into sections are based on the well-supported bootstrap value of more than 75% (Kron et al., 2002; Vander Kloet and Dickinson, 2009). Nonetheless, placement of different species in different sections is not always consistent due to inefficient nuclear and chloroplast data used to study the molecular taxonomy of the genus. Therefore, identification and classification are still based on available morphological and phenotypic data which sometimes creates the species paradox in the genus *Vaccinium* (Kloet, 1980). The main difficulty describing the phenotypic data is due to the complex ploidy level of the different species of this genus which is ranged from diploid to hexaploid (2x, 3x, 4x, 5x, 6x). However, species of the genus have a common basic chromosome number which is always $x = 12$ (Zdepski et al., 2011; Rawland et al., 2012).

Although most of the species of this genus are cross-pollinating, self-pollinating species also exist. For instance, *V. macrocarpon* Aiton (cranberry) and *V. oxycoccus* L. (small cranberry) belong to the section *Oxycoccus*, having different mode of pollinations. Eventhough cranberry is self-fertile and has narrow genetic background, occasionally it

can fertilize with *V. oxycoccus* and produce fertile offsprings (Hancock et al., 2008; Česonienė et al., 2013).

Another example could be the section *Cyanococcus* which includes one of the economically important and well-studied species called blueberry. Blueberry is a taxonomically diverse group containing genotypes mostly from different species like *V. corymbosum* L. (highbush blueberry), *V. angustifolium* Aiton. (lowbush blueberry) and *V. virgatum* Aiton. (rabbit-eye blueberry). Nonetheless, species boundary of the section *Cyanococcus* is not well resolved meaning that a significant interspecies hybridization is frequent in the natural condition (Qu and Hancock, 1995). The commercially important cultivars of blueberry are mainly a hybrid (the mixture of different species) in origin (Qu and Vorsa, 1999; Brevis et al., 2008).

Other economically important but non-cultivated plant species of this genus are bilberry (*V. myrtillus* L., 2x, Section: *Myrtillus*), cowberry or ligonberry (*V. vitis-idaea* L., 2x Section: *Vitis-idaea*), Caucasian whortleberry (*V. arctostaphylos* L., 4x, section: *Hemimyrtillus*) and bog bilberry (*V. uliginosum* L., 2x, 4x, 6x, Section: *Vaccinium*). Geographical distribution of these sections is thought to be circumboreal and species are both self and cross-pollinated (Retamales and Hancock, 2012).

2.2 Habitat and Geographical Distribution

Normally species from the genus *Vaccinium* are “acid-loving” (normal pH range for the growth of this species is 5.8 or less) shrubs or woody bushes which may also play role in the natural distribution of the species of this genus. Species habitats are diverse mainly based on section (Vander Kloet, 1988).

Most of the blueberry species are native to North America and have chilling requirement. Nevertheless, this requirement was minimized through hybridization with the southern blueberry species. Although species from blueberry can grow in nutrient-poor soil, the soil needs good drainage (Retamales and Hancock, 2012). In contrast, cranberry and small cranberry grow in bog habitats (*Sphagnum* peat bog), swamps and temperate climate. This kind of habitats have very poor water drainage site but high water level. Species of cranberry and small cranberry can survive acidity range from pH 3-4.5. Even though

cranberry is native to North America, small cranberry is circumboreal and grows in cooler regions of Europe and North America (Česonienė et al., 2013).

Bilberry normally grows in pine and spruce heath forests, plateau area near mountain and old peat bogs area in Europe, North America, the Greenland and northern part of Asia (Vander Kloet, 1988). Bilberry can propagate asexually through rhizome and sexually by the seeds (Zoratti et al., 2015). *V. uliginosum* normally grows in cooler parts of northern hemisphere, similar environments to bilberry or bog environment (Wang et al., 2014). *V. arctostaphylos* is also called Caucasian whortleberry and mainly distributed in Iran, Turkey, Caucasus, Bulgaria and near this area. Whortleberry prefers forests with fagus, firs, pine or with *Rhododendron* with the high elevated areas with enough rain (Nickavar and Amin, 2004).

2.3 Economic Importance

Small berries from different species of the genus *Vaccinium* had been picked from nature by the local people from all different areas of the world until domestication started during 1900. The berries are mainly used to be eaten as fresh fruits or used for the preparation of jam, juice, pies, jelly and wine (Çelik, 2012).

Recently this fruit has been claimed as superfood in the media due to various medicinal and nutritional properties and its consumption has increased much more in the last 15 years (Mudd et al., 2013). Fruit of blueberry is mainly rich in many beneficial secondary plant metabolites like anthocyanin, carotenoids, flavonoids, polyphenols, galactosides, glucosides, the lower amount of ascorbic acid and so on (Moyer et al., 2002; Nickavar and Amin, 2004). Evidence shows that some of these natural components have high beneficial impacts on the human body including, anticancer, antioxidant and antidiabetic activity (Trehane, 2004; Mudd et al., 2013). Moreover, those compounds have been successfully used in the reduction of cardiovascular disease, improving brain function and cognitive ability in replacement of artificial drugs or medicines and research on this area is still ongoing (Hou, 2003; Mudd et al., 2013).

The United States is the largest (half of total world production) blueberry and cranberry producing country in the world and this has a great impact on its economy (FAOSTAT,

2017). Most of the commercially important cultivars of blueberry are developed and patented by this country. Moreover, worldwide production of the berry fruit has also increased and some other countries of South America and Europe are also coming forward to produce berry fruit (Lobos and Hancock, 2015).

2.4 Evolutionary Species Dynamics of the Genus *Vaccinium*

Several biological phenomena need to be considered when describing the phylogenetic relationship among different species of the *Vaccinium* according to Camp, 1942;

1. Genus *Vaccinium* has a same basic set of chromosomes and is prone to produce unreduced gamete (Qu and Hancock, 1995; Qu and Vorsa, 1999).
2. Species have reduced reproductive barrier and can hybridize interspecies or even in some cases at intersectional level as long as they are homoploid (same ploidy level) (Chavez and Lyrene, 2010).
3. Most of the species of this genus are self-sterile and prime fertilizations are by means of cross-pollination (Qu and Vorsa, 1999).
4. The genus shows a high incidence of higher ploidy level (both autopolyploidy and allopolyploidy) due to self sterility (Camp, 1942).
5. Species of this genus are perennial and can propagate asexually. Once a new genome evolved in nature through this complex hybridization procedure they can have more chance to survive and give rise to a new population once having the favorable environment (Chavez and Lyrene, 2010).

Two main factors (the abiotic factor and time factor) are thought to be linked with the *Vaccinium* species evolution. The abiotic factor is that *Vaccinium* is an “acid-loving” plant and cannot survive naturally when the soil pH rises so much towards the alkalinity. Therefore, the distribution could be limited in different geographical regions and it can also be affected by genetic changes (Camp, 1942). The other one is the time factor. As a group, *Vaccinium* is quite an old genus thought to be evolved in Cretaceous era.

Therefore, it is very common to find morphologically distinct but genomically homoploid species in this genus (Camp, 1942).

Therefore, even in the modern era taxonomy of the genus *Vaccinium* is still an elusive one. Powell and Kron, (2002) studied the taxonomic position and phylogenetic divergence in different sections of this genus through chloroplast and mitochondrial DNA sequence analyses (Figure 2.1.). GeneBank ID for *matK* are AF382810, U61316, AF419706, AF419717, AF419702 and for *nrITS* are AF382732, AF382730, AF419778, AF419788, AF419774 for *V. myrtillus*, *V. macrocarpon*, *V. corymbosum*, *V. uliginosum* and *V. arctostaphylos*, respectively.

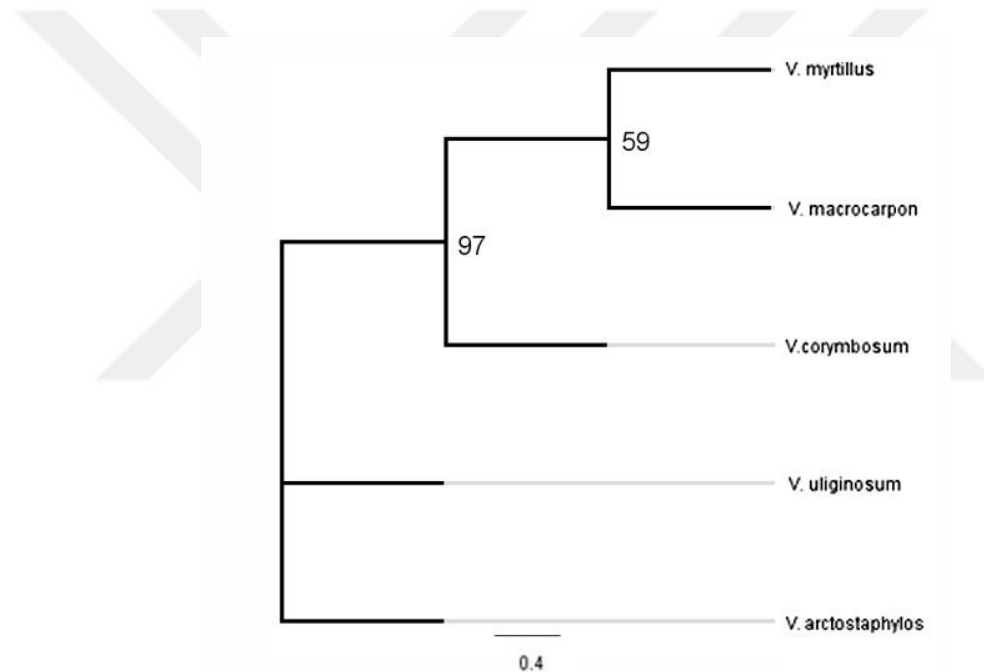


Figure 2.1. Phylogenetic relationship between five well-known and economically important *Vaccinium* species from concatenated sequence data of chloroplast MaturaseK gene (*matK* gene) and nuclear ribosomal internal transcribed spacer gene (*nrITS* gene) using maximum parsimony PAUP* method (Swofford, 2002). (Phylogenetic dendrogram was recalculated using the sequence data and parameters provided in Powell and Kron, (2002).

2.5 Cultivation of *Vaccinium*

Today commercial cultivation of *Vaccinium* species is only limited to the two major crop plants blueberry and cranberry. Blueberry was originally first domesticated by the US Department of Agriculture (USDA) in 1908 from the wild bushes belonging to the species *Vaccinium corymbosum* L. 4x Section: *Cyanococcus* (highbush blueberry) (Lyrene, 2008). Nevertheless, there are two southern highbush cultivars, northern and southern type. Southern types have less chilling requirement and winter hardiness as well as excellent adaptability. First southern highbush blueberry cultivar was developed from northern type by hybridization with different native evergreen species like *V. darrowi* Camp., 2n = 2x Section: *Cyanococcus* and was initiated in the 1950s in Florida (Sharp and Darrow, 1959). Since then northern highbush cultivars have been crossed with many different types of wild *Vaccinium* species to integrate beneficial gene from different backgrounds and to increase their adaptability (Darnell and Williamson, 1997). Many commercially successful southern highbush varieties since then have been released by United States of America (USA) (Spiers et al., 1996). On the other hand, the earliest recorded cranberry (section *Oxycoccus*) cultivation was far back to 1816 in Cape Cod Massachusetts, where cranberry was first domesticated from the wild species (Trehane, 2004). Since then over 100 different clones from the wild species has been selected for different properties like disease and pest resistance (Hancock et al., 2008).

Besides the species of sections, *Cyanococcus* and *Oxycoccus*, representing two important crop blueberry and cranberry, other species belonging to different important sections like *Myrtillus*, *Vaccinium*, *Hemimyrtilus* are also attracting research interest (Wang et al., 2014). Collection and preservation from these groups of species are also ongoing in different countries of Europe besides the United States. For instance, due to several important adaptive traits belonging to whortleberry (*V. arctostaphylos* Section *Hemimirtillus*) and the ability to produce fertile offsprings from hybridization with the Section *Cyanococcus*, section *Hemimirtillus* is now considered as the tertiary gene pool for blueberry crop improvement (Ballington, 2001; Ehlenfeldt and Ballington, 2012).

2.6 Overall Progress in Genetics and Genomics of *Vaccinium*

It is well established that cultivated blueberry is tetraploid ($2n = 4x = 48$) but cranberry is diploid ($2n = 2x = 24$). However, chromosome number and ploidy level estimation are lacking for some of the wild but important *Vaccinium* species. Hummer et al. (2015) takes an initiative to estimate ploidy level of the wild germplasm accessions collected in National Clonal Germplasm Repository (NCGR), The U.S. Department of Agriculture, Agricultural Research Service, of United States. A total of 58 taxa representing 20 sections were evaluated in that study to estimate the ploidy levels through flow cytometry which revealed the ploidy level diversity in the genus *Vaccinium*. Table 2.1. shows estimated ploidy level of some selected plant species.

Table 2.1. DNA ratio estimated for some important *Vaccinium* species compared to *Vinca minor* L. and their origin by Hummer et al. (2015). (x) measurement not available.

Section	<i>Vaccinium</i> species	Accession number/cultivar	DNA ratio/ Mean ratio \pm SD	Ploidy	Origin
<i>Cyanococcus</i>	<i>V. corymbosum</i> L.	20 cultivars	1.513 \pm 0.03	4x	US
<i>Cyanococcus</i>	<i>V. corymbosum</i> L. (complex hybrid cultivar)	5 replicates	2.05 \pm 0.15	6x	US
<i>Hemimyrtillus</i>	<i>V. arctostaphylos</i> L.	<i>V. arctostaphylos</i> GE-2004-098	1.46 \pm x	4x	Republic of Georgia
<i>Myrtillus</i>	<i>V. myrtillus</i> L.	Three Creek Trail	0.74 \pm x	2x	US, Oregon
<i>Oxycoccus</i>	<i>V. macrocarpon</i> Aiton	Cultivar-Hamilton	0.65 \pm x	2x	US

The haploid genome size of the blueberry (*V. corymbosum*) ranges from 500 to 600 million bases (Mb) depending on the accession, used which is still about five times the size of model plant *Arabidopsis thaliana* L. (Costich et al., 1992). In addition, the size of the haploid genome of cranberry is estimated to be 470 Mb (Folta and Kole, 2016).

To understand the natural genetic diversity among the *Vaccinium* species, the first blueberry marker used was restriction fragment length polymorphisms (RFLPs) to deduce the segregation level in mitochondrial DNA and to distinguish diverse high-bush blueberry cultivars (Haghighi and Hancock, 1992). Later on other PCR-based molecular markers have been also employed in blueberry, such as random amplified polymorphic DNA (RAPD), simple sequence repeat (SSR) and express sequence tag-polymerase chain reaction (EST-PCR) markers (Levi and Rowland, 1997; Dhanaraj et al., 2004; Rowland et al., 2010). Due to the advancement of sequencing technologies developed throughout the last decade, several thousand expressed sequence tags (ESTs) and a few hundred SSRs marker derived transcriptome sequences have been made available for the breeders and researchers (Dhanaraj et al., 2007; Bian et al., 2014). The main goal of these markers study was to construct the genetic linkage map of important gene like cold hardiness, disease resistance, pest resistance, tolerance to high pH condition and other commercially important breeding traits (Dhanaraj et al., 2004; Rowland et al., 2012; Bian et al., 2014). Schlautman et al. (2017) studied the construction of a composite map from a high density multigene pedigree linkage mapping through genotyping by sequencing.

The most important steps in *Vaccinium* genome research were the release of publicly available whole genome sequence (Illumina reads) data (Brown et al., 2011) and transcriptomic data (Gupta et al., 2015) for *V. corymbosum* accession ‘W8520’. Moreover, genomic and transcriptomic data was also made available for *V. macrocarpon*, cultivar ‘Ben Lear’ (Polashock et al., 2014) as well as whole chloroplast genome sequence data (Fajardo et al., 2013). In addition to that an integrated web-based database called “The genome database for *Vaccinium* (GDV)” was established in 2011 (<https://www.vaccinium.org>). The main goal for this database development was to integrate genetics, genomics and breeding data from economically important gene pools of *Vaccinium* and to make them available for the scientist from different backgrounds and countries. Now, GDV database has downloadable genome assembly, transcriptome assembly, raw whole genome sequence data, annotated transcript, maps and markers

developed so far from different sources. The database also has some bioinformatics tools like Blast+, JBrowse and Mapviewer to annotate and analyze important regions of the genome. Moreover, database has been constantly being updated from many different new sources (<https://www.vaccinium.org>).

Although the advancement of *Vaccinium* research is quite satisfactory for the past years, information about the repetitive sequences in the *Vaccinium* genome and how these sequences affect the evolution of *Vaccinium* genome is quite limited (Die and Rowland, 2013; Mudd et al., 2013). Therefore, in this study, an attempt has been taken to find out the interplay between repeat sequences and *Vaccinium* genome evolution.



CHAPTER III

PLOIDY LEVEL ESTIMATION OF CULTIVATED AND WILD *VACCINIUM* SPECIES

3.1 Background Information

Total nuclei content of an organism is called genome and genome size is normally a constant value representing a particular species (Bennett and Leitch, 2005). Although genome size of a species is closely related to chromosome number and ploidy level of that species, it does not always follow the rules when considering organismic complexity or phylogenetic relationship. The anomaly of the genome size in intraspecific level is termed as the C-Value paradox. Due to hybridization among related species and high level of polyploidy, differences are higher in the plants than animals (Costich et al., 1993). The main differences between ploidy level and genome size are thought to be linked to the differences in repetitive DNA sequences (satellite, Transposable element (TE), rDNA repeats, telomere repeats and so on) but not in genic sequences. Evolution of repeat sequences, however, is related with adaptive forces in different environmental and developmental regulation in different species throughout the evolutionary time scale (Macas et al., 2015). Therefore, nowadays studying genome size variation among inter-intraspecies level, estimation of ploidy level, and analysis of repetitive DNA is a very common framework to reveal in-depth phylogenetic relationships among closely related species (Pellicer et al., 2018).

Vaccinium is one of the complex, widely distributed and intrigued genera, capable to produce unreduced gamete and hybridization among inter or intra-specific level are very common in natural condition (Retamales and Hancock, 2012). Nonetheless, basic knowledge about chromosome number is an important cytological character which can solve the taxonomic status of different species of this genus originating from the different geographical regions of the world. In this study, an attempt was taken to determine chromosome number and ploidy level variation of cultivated and wild *Vaccinium* species.

3.2 Materials and Methods

The plant materials investigated in the study are listed in the (Table 3.1.)

Table 3.1. List of plant materials used in this study.

Species	Section (Vander Kloet and Dickinson, 2009)	General information		Reference
		Chromosome number	Origin	
<i>Vaccinium corymbosum</i> L. (Cultivar ‘Jubilee’)	<i>Cyanococcus</i>	48	USA	Trehane, 2004
<i>Vaccinium arctostaphylos</i> L. (wild)	<i>Hemimyrtillus</i>	48	Turkey	Çelik, 2012
<i>Vaccinium myrtillus</i> L.- (wild)	<i>Myrtillus</i>	24	Turkey	Çelik, 2012
<i>Vaccinium uliginosum</i> L.- (wild)	<i>Vaccinium</i>	24	Turkey	Çelik, 2012

3.2.1 Collection and identification of Turkish wild *Vaccinium* species

Çelik, (2012) recorded that wild native *Vaccinium* species was grown in natural condition near the northeastern part of Turkey (Black Sea Region). Therefore, field expedition was carried out several rounds throughout these regions during the years of 2015-2017. Three different *Vaccinium* species with potential horticultural importance were identified and localized near the Kackar Mountains National Park of Rize, Turkey through GPS system. Recorded names of identified species and GPS coordinate are shown in Table 3.2.

Table 3.2. Name of the identified wild *Vaccinium* species and their GPS coordinates

Species	Location		Elevation
	Latitude	Longitude	
<i>Vaccinium arctostaphylos</i> (round shaped fruit)	40.8402982	41.10292690000006	2641 m
<i>Vaccinium arctostaphylos</i> (elongated shaped fruit)	40.84379129298008	41.08144321960765	2779 m
<i>Vaccinium myrtillus</i>	40.9524618	41.101181999999994	1251 m
<i>Vaccinium uliginosum</i>	40.952461	41.101181999999994	1251 m

Species were further identified in the laboratory of Agricultural Genetic Engineering Department of the Niğde Ömer Halisdemir University, based on several morphological properties of plant-like plant height, stem diameter, leaf, flower and fruit according to Çelik, (2012) (Figure 3.1. and 3.2.). Turkish wild *Vaccinium* species were grown and maintained in greenhouse condition, Ayhan Şahenk Faculty of Agricultural Sciences and Technologies of the Niğde Ömer Halisdemir University, Niğde, together with the cutting of cultivar ‘Jubilee’ (a highbush blueberry cultivar (*V. corymbosum* L.)) brought from United States about three years ago. Morphology of both wild and cultivated species were studied comparatively to deduct morphological differences among the species. Morphology analysis were done using a millimeters scale (Figure 3.2.). To identify each species about 10 individual plant was carefully analyzed and monitored for about one year to understand their morphological differences. Morphological diversity of wild and cultivated *Vaccinium* species are shown in Figure 3.1. and 3.2.

General morphological differences identified in cultivated and wild *Vaccinium* species;

V. corymbosum (cultivar ‘Jubilee’)

Plant was about 3-4 meter long and bushy type. Leaf width × length ranged between 20×45 mm to 30×65 mm and leaf margin was entire. Leaf color was bright green but at the end of the season, the leaf color turned to red. Each leaf had leaf bud of about 4-5 mm at the bottom of each petiole. Stem diameter ranged between 3 mm-10 mm. Fruit diameter was 10-12 mm, color dark blue, round, fruit flesh yellowish, have a wide mouth opening (Kloet, 1980) (Figure 3.1.).

V. arctostaphylos-round fruit

Plant was about 3-4 meter long, bushy type, grows comparatively lower plateau in the Kaçkar Dağları Milli Parkı. Leaf width × length ranged between 20×45 mm to 30×85 mm. Leaf was elongated and leaf margin was entire, color was reddish green with leaf bud of 2-5 mm. Stem diameter was 5-12 mm. Fruit was round shiny blackish color with a small opening. Fruit flesh was yellowish. Fruits were found around this area from July to October (Nickavar and Amin, 2004; Çelik, 2012). Plants were growing in individual cluster than other three species of *Vaccinium* in the natural environment (Figure 3.2.).

V. arctostaphylos-elongated fruit

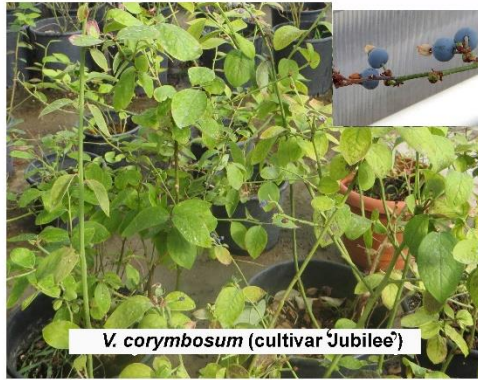
Plant size was similar to that of the round shaped fruit, but leaf color was more greenish compared to *V. arctostaphylos*-round shaped fruit. Fruit was elongated with shiny blackish color, flesh yellowish, and grows in the cluster. Plants found to be grown under pine forest in shady areas and higher plateau compared to *V. arctostaphylos*-round shaped fruit (Çelik, 2012) (Figure 3.2.).

V. myrtillus

Plant was woody about 1-2 meter long and low bush type. Leaf was greenish with the reddish cluster, 10×20 mm long with leaf bud under petiole. Plants grew in the open lower plateau with constant water flow with rocky area. Fruit was 10-12 mm long, round, blue color, almost like blueberry, but flesh color was dark reddish, mouth wide open, mostly single. Stem diameter was 2-5 mm. Fruits was collected similar times like *V. acrostaphylos*. *V. myrtillus* and *V. uliginosum* were found to grow close o each other in the natural environment (Çelik, 2012) (Figure 3.2.).

V. uliginosum

Plant was woody about 0.9-1.5 meters long, low bush type. Leaf was dark bluish green. Stem diameter was 1.5-2 mm. Leaf was small about 5 mm long and round shape with leaf bud of about 0.5-0.8 mm. Fruit was dark blue, small size of about 3-5 mm long, round shape, flesh was white color. Species of *V. uliginosum* were found to grow with *V. myrtillus* (Wang et al., 2014; Çelik, 2012) (Figure 3.2.).



a



b



c



d

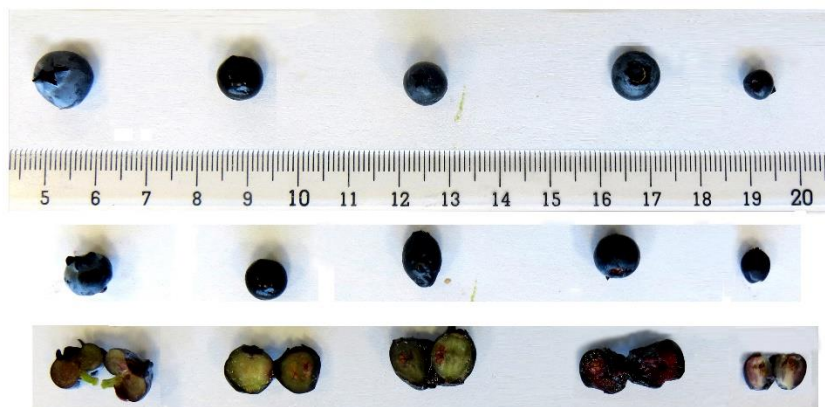


e

Figure 3.1. Morphological diversity of five different *Vaccinium* species at the fruiting stage. Green house grown cultivated species (a) and wild plants from Kaçkar Mountains (b, c, d, e). [Photographs are copyright September, 2017, by Nusrat Sultana (a) and by Prof. Dr. SEDAT SERÇE (b, c, d, e)].



Cultivar Jubilee⁷ *V. arctostaphylos* (round fruit) *V. arctostaphylos* (elongated fruit) *V. myrtillus* *V. uliginosum*



Cultivar Jubilee⁷ *V. arctostaphylos* (round fruit) *V. arctostaphylos* (elongated fruit) *V. myrtillus* *V. uliginosum*

Figure 3.2. Difference in leaf, fruit and stem morphology of the studied *Vaccinium* species

3.2.2. Chromosome preparation and chromosome counting

Chromosomes were prepared according to the protocol described by Heitkam et al. (2015) with some modification. Young leaf buds were collected during the spring and summer season (July to August) for both the cultivated and wild *Vaccinium* species and instantly stored in ice at 4° C and kept for overnight. Fixation of leaf buds was performed in methanol/ glacial acetic acid (3:1) with 1% PVP. The fixative solution was changed several rounds to ensure complete removal of all plant pigment and finally preserved at -20° C in the same solution until slide preparation.

After removing the outer leaf coat under the microscope the only inner white meristematic portion of a single leaf buds was washed in citrate buffer (4 mM citric acid and 6 mM sodium citrate, pH 4.5) several round to remove the existing fixative and digested overnight at room temperature with the enzyme mixture PINE containing 2% (w/v) cellulase *Aspergillus niger* (Sigma C-1184), 4% (w/v) cellulase onozuka R10 (Sigma 16419), 2% (w/v) cytohelicase from *Helix pomatia* (Serva C-8274), 0.5% (w/v) pectolyase from *A. japonicus* and 5% (w/v) pectinase from *A. niger* in citrate buffer (4 mM citric acid and 6 mM sodium citrate, pH 4.5). The material was digested until become soft followed by centrifugation and washing nuclei with citrate buffer for two times and one time in fixative at 4°C at 4500 rpm. Washing step was finished by one final wash in fixative with increased 5000 rpm at 4°C. Pellet of nuclei was dissolved in 500 µl of fixative and mixed by slow pipetting. For each slide 13 µl of mixture was spread on the slide and was instantly washed with fixative three times and air dried.

Air-dried slides were stained with 1% aceto-orcein solution and checked several rounds to ensure perfect staining which is about 30 minutes. After proper staining, the slide was washed with 70% ethanol to remove extra dye and mounted with Canada balsam solution. Slides having well-spread metaphase chromosome and no cytoplasm contamination were selected for image acquisition.

Total 10 well-spread cells, containing metaphase chromosome was counted for each species separately to identify the ploidy level. Cell and chromosomes were analyzed using Leica-Epifluorescence microscope (LEICA, DMIL-LED) equipped with LAS software (Leica) and LEICA high resolution digital camera. Images of well-spread metaphase

chromosomes were captured at 100X magnification and analyzed through the LAS software. After image acquisition, the final picture was processed through Adobe Photoshop 7 software.

3.2.3 Handling and performing flow cytometry

Nuclei of both cultivar and wild *Vaccinium* species were isolated and stained using Partec CyStain PI absolute P reagent kit according to the manufacturer's instructions. First, a specific portion of young leaf material (approximately 1 cm²) was chopped with a sharp razor blade with the 400 µl prescribed extraction buffer (provided) and filtered through a 50 µm Partec cell trices to remove the cell debris. After that 1.6 ml staining buffer including RNAase solution was added. The extraction was incubated in ice with 76 µl Propidium iodide (PI) solution (1 mg/ml) until analysis in Partec Cyflow Space (Partec, Münster, Germany) with 488 nm laser output operated with 200 mW power. Ploidy level was estimated based on the peak position of about 3000 particles in flow cytometer. Histogram of peak position of individual species and all species together were analyzed with the FlowMax software v.2.4 (Partec GmbH).

3.3 Results

Ploidy level estimation of *Vaccinium* species

Ploidy level of the five studied plant species was identified through complemented chromosome counting and flowcytometry analysis and represented in Table 3.3., Figure 3.3. and 3.4.

V. corymbosum (cultivar ‘Jubilee’)

Both from the literature review and our analysis it was found that *V. corymbosum*-cultivar ‘Jubilee’ is a tetraploid species. The total chromosome number for cultivar ‘Jubilee’ was $2n = 4x = 48$ and mean FL1 UV LED value was 95.56 after total counted cell number of 3000 with Coefficient of Variation (CV%) 5.38. Meanwhile, cultivar ‘Jubilee’ was used as a standard to detect the ploidy level of other four unknown or less cited wild species (Figure 3.3a and 3.4a).

V. arctostaphylos (round fruit)

Total chromosome number for this studied species was $2n = 4x = 48$ but FL1 UV LED mean value was 144.83 with Coefficient of Variation (CV%) 2.82. Therefore, our data was contradictory for chromosome counting (tetraploid) and flow cytometry analysis (Hexaploid). As because it is well-known that species with same chromosome number or ploidy level could have different genome size due to the differential accumulation of repeat sequence, therefore estimated ploidy level for this species was tetraploid (Macas et al., 2015) (Figure 3.3b and 3.4b).

V. arctostaphylos (elongated fruit)

Total chromosome number for this species was counted $2n = 4x = 48$ and FL1 UV LED mean value was 98.29 with Coefficient of Variation (CV%) 7.56. Therefore this is a tetraploid species (Figure 3.3c and 3.4c).

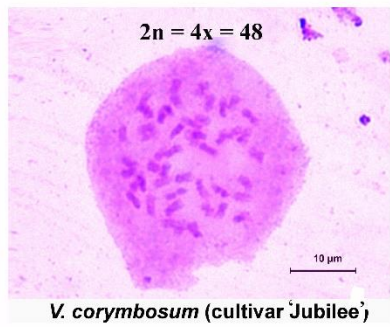
V. myrtilus

Total chromosome number was, $2n = 2x = 24$ and FL1 UV LED mean value was 33.44 with Coefficient of Variation (CV%) 13.25. Therefore estimated ploidy level for this species was diploid (Figure 3.3d and 3.4d).

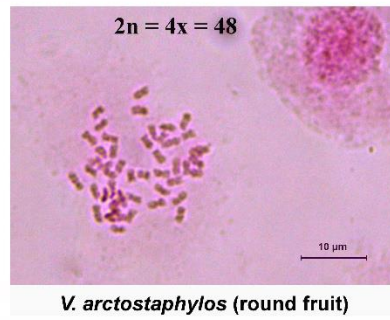
V. uliginosum

Total chromosome number was, $2n = 2x = 24$ and FL1 UV LED mean value was 65. Therefore, estimated ploidy level is diploid (Figure 3.3e and 3.4e).

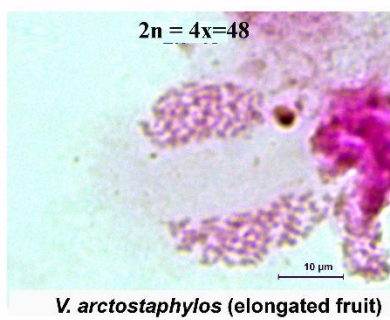




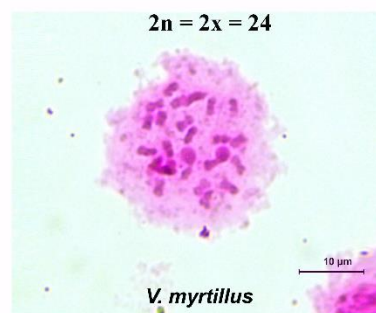
a



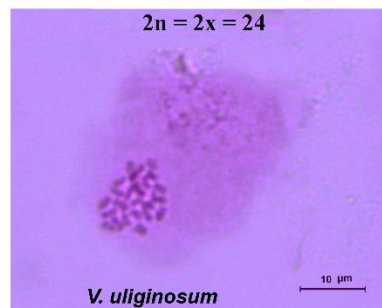
b



c



d



e

Figure 3.3. Estimation of chromosome number of the studied *Vaccinium* species. *V. corymbosum* (cultivar 'Jubilee') ($2n = 4x = 48$) (a), *V. arctostaphylos* (round fruit) ($2n = 4x = 48$) (b), *V. arctostaphylos* (elongated fruit) ($2n = 4x = 48$) (c), *V. myrtillus* ($2n = 2x = 24$) (d), *V. uliginosum* ($2n = 2x = 24$) (e).

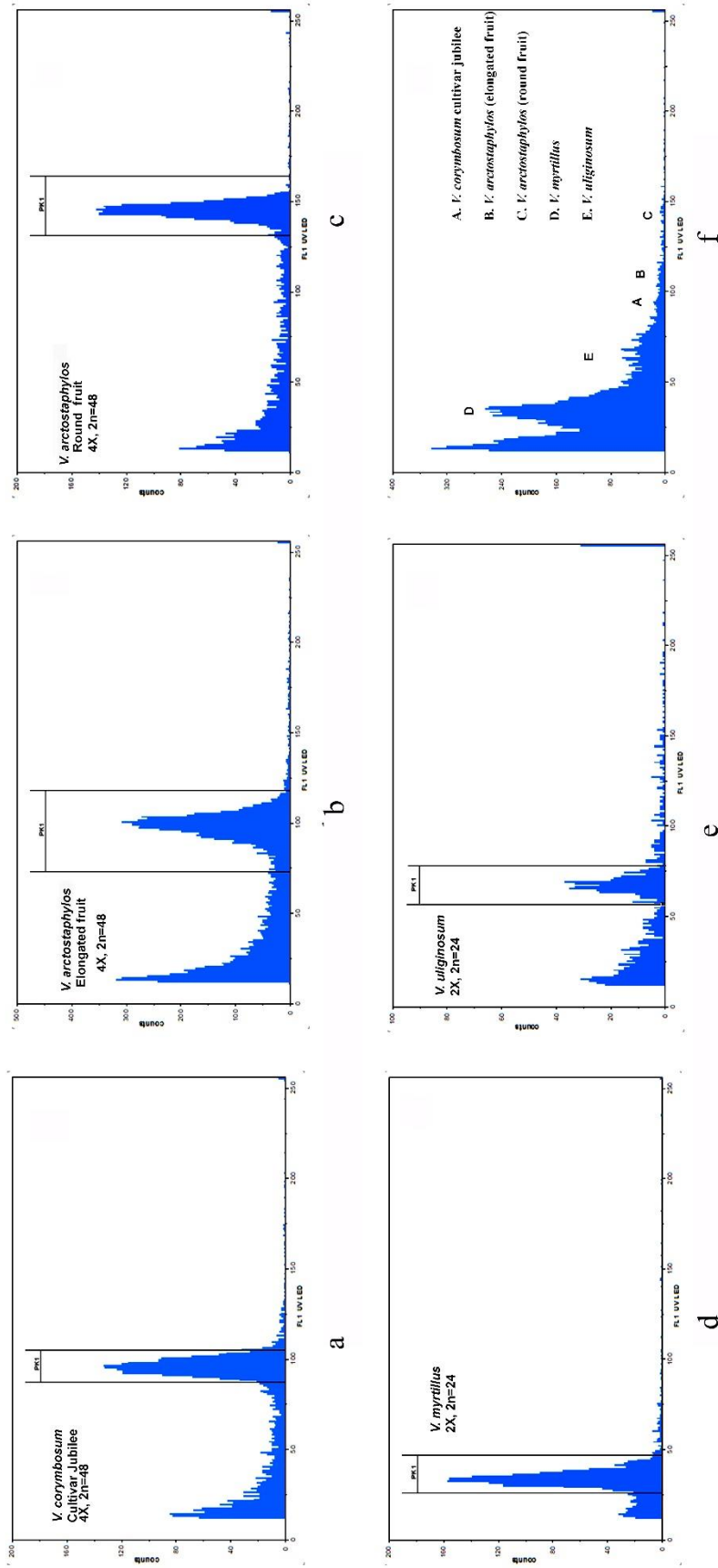


Figure 3.4. Estimation of ploidy level among studied *Vaccinium* species using flow cytometry. Peak of *V. corymbosum* (cultivar 'Jubilee') (a), *V. arctostaphylos* (elongated fruit) (b), *V. arctostaphylos* (round fruit) (c), *V. myrtilles* (d), *V. uliginosum* (e), (a-e) (f)

Table 3.3. Chromosome number, FL1 UV LED mean value and Coefficient of Variation (CV%) of flow cytometry reading of *Vaccinium* species. (x) measurement failed.

Plant material	Chromosome number	Mean value in the FL1 UV LED scale	CV%	Ploidy level (Reference Çelik, 2012)
<i>V. corymbosum</i> -cultivar 'Jubilee'	2n = 4x = 48	95.56	5.38	Tetraploid
<i>V. arctostaphylos</i> (elongated fruit)	2n = 4x = 48	98.29	7.56	Tetraploid
<i>V. arctostaphylos</i> (round fruit)	2n = 2x = 48	144.83	2.82	Tetraploid
<i>V. myrtillus</i>	2n = 2x = 24	33.44	13.25	Diploid
<i>V. uliginosum</i>	2n = 2x = 24	65	x	Diploid

CHAPTER IV

BIOINFORMATICS ANALYSES AND CHARACTERIZATION OF REPETITIVE ELEMENTS OF *VACCINIUM* GENOME

4.1 Background Information

The first isolation method for repetitive DNA was achieved from the density-gradient ultracentrifugation in cesium chloride solution. The protocol confirmed that if the genomic DNA is kept under ultracentrifugation in the presence of cesium chloride, the most repetitive portion of the DNA produces a distinct band in the tube and it is possible to isolate and manipulate them for further studies (Waring and Britten, 1966). Later, scientists developed the Cot analysis to isolate repeat sequences based on the principle of DNA renaturation kinetics when repeat sequence re-associate much faster than non-repeat sequence after denaturation (Britten and Kohne, 1968). Ultracentrifugation and Cot analysis based methods were substituted with one of the popular techniques invented based on restriction digestion of genomic DNA with an appropriate restriction enzyme, and the method is supplemented with the cloning, sequencing and dot blot analysis technique. Even though efficient, the technique has the limitation of deducing the vast amount of sequence heterogeneity present in a group of satellite repeats (Singer, 1982). Therefore, implementation of PCR amplification was often carried out through designing repeat specific primers (Garrido-Ramos, 2017).

Recent breakthrough of technological advancement of sequence data opens a new door for repeat sequence analysis. Due to dramatic reduction in cost and presence of diverse sequencing platforms, many economically and agronomically important organisms' genomes have already been sequenced, and genomic information is publicly available in the online-based platforms so that scientists can use them for their own purposes. In addition to database development, several user-friendly data analysis programs and bioinformatics tools have also been developed and constantly being updated (Mehrotra and Goyal, 2014; Garrido-Ramos, 2017). One of the efficient bioinformatics pipelines called RepeatExplorer has recently been installed in Galaxy based platform, reducing the technological barrier to scientists (Novák et al., 2013).

RepeatExplorer is a pipeline of de novo identification of repeat sequence from the species lacking the reference genome (Novák et al., 2013). RepeatExplorer follows a similarity based read clustering. The clustering follows by generating a graph from the read similarities. Different types of repetitive DNA can be identified and classified based on the size and shape of the graph. Besides RepeatExplorer, there are some other softwares like satMiner and TAREAN. These additional softwares are also important to identify some special types of repeats such as the repeats that present in very small quantities or the repeats that have low complexity and can skip from RepeatExplorer based analysis (Mehrotra and Goyal, 2014; Novák et al., 2017). The above-mentioned tools have been extensively used in different plant and animal species to identify and characterize the repetitive DNA sequences (Garrido-Ramos, 2017). The main objectives of this chapter was to identify and quantify different types of repeat sequences from the blueberry and cranberry genome.

4.2 Materials and Methods

Here, publicly available next-generation sequence data was used from the genomic database for *Vaccinium* (GDV). Different bioinformatics tools were used for processing the next generation sequence reads and identification of repetitive sequence in blueberry and cranberry genome (Ruiz-Ruano et al., 2016; Novák et al., 2017).

Softwares used for these analyses were as follows; 1. RepeatExplorer galaxy instance (Novák et al., 2017), 2. TAREAN (Novák et al., 2017), 3. Trimmomatic tools (Bolger et al., 2014), 4. FASTX-Toolkit, seqtk (github.com/lh3/seqtk) tool, 5. Geneious bioinformatics platform (<http://www.geneious.com/>; Kears e et al., 2012)

4.2.1 Publicly available genomic database of *Vaccinium*

Paired-end Illumina genomic next generation sequence data and genome assembly were downloaded through the link provided by the Genome Database for *Vaccinium* (GDV) (Table 4.1.).

Table 4.1. Available public genome (DNA) database for *V. corymbosum* and *V. macrocarpon* (The genome database for *Vaccinium* (GDV),
Source://www.vaccinium.org)

Types of data		<i>V. corymbosum</i>	<i>V. macrocarpon</i>
General information	Name of the studied organism, Chromosome number, genome size and Reference	<i>V. corymbosum</i> L. strain: 'W8520', common name: Blueberry, 2n = 2x = 24, genome size 500 Mb, (Reference; Brown et al., 2011; Bian et al., 2014; Gupta et al., 2015)	<i>V. macrocarpon</i> Ait. cultivar: 'Ben Lear' (CNJ99-125-1 inbred clone), common name: Cranberry, 2n = 2x = 24, genome size 470 Mb, (Reference; Polashok et al., 2014)
Whole Genome Sequencing (NCBI-Sra data)	Study accession	PRJNA170639	PRJNA245813
	Instrument	Illumina HiSeq 1000	Illumina Genome Analyzer IIX
	Library layout	Paired	Paired
	Sra accession	SRA053499	SRA161994
Assembled genome			
	Accession Name	V_corymbosum_Aug_2015	ASM77533v1
	Assembly accession or available from	IGBQuickLoadsite http://www.igbquickload.org/ blueberry	Assembly: GCA_000775335.1 (NCBI)

4.2.2 Illumina paired-end sequence read preparation

Paired-end reads of whole genome sequence data were subjected to quality filtration, adapter trimming, processing read length to range from 76-100 bp. Trimmomatic tool was used to remove the existing Illumina Truseq adapters sequence (Bolger et al., 2014) with the parameters ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10. Fastx_trimmer from FASTX-Toolkit (hannonlab.cshl.edu/fastx_toolkit) were used to trim all reads to either 70bp or 100 bp length and subsequent removal of the shorter sequence using seqtk (github.com/lh3/seqtk). The protocol follows to interlace the paired-end reads and selection of 1M, 5M, and 10M reads representing different genome coverage ranged from 0.02X to 2.04X using FASTX-Toolkit (hannonlab.cshl.edu/fastx_toolkit).

4.2.3 Clustering and analysis of clusters generated through RepeatExplorer pipeline

RepeatExplorer pipeline (Novak et al., 2013) with parameters (minimal overlap of clustering) -l 39 for and (minimum overlap for assembly) -o 30 was used for clustering, considering different genome coverage each time ranged from 0.02X to 2.04X. Different genome coverage was tested to identify the best genome coverage leading to the identification and characterization of the maximum amount of repetitive sequence despite the limitation of handling a large amount of data of the pipeline (Figure 4.1.)

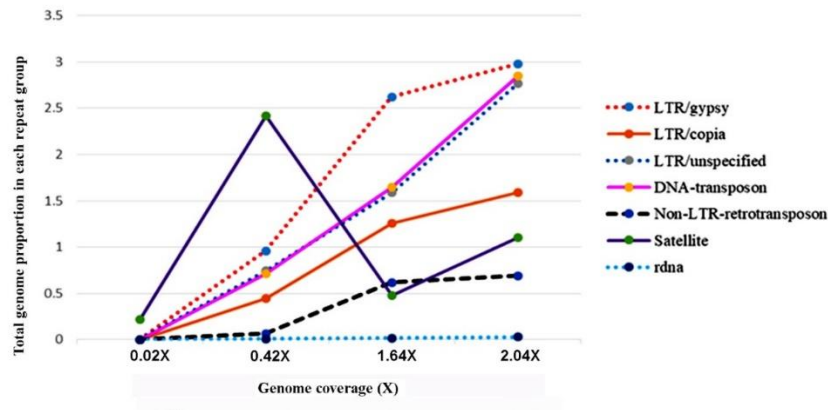


Figure 4.1. Genome coverage vs repeat proportion identified in the *V. corymbosum* genome.

It was found from the analysis that genome coverage of 2.04X identifies and characterizes the maximum amount of repeat sequence. Therefore, individual read clustering was performed separately for both *V. corymbosum* and *V. macrocarpon* with the respective genome coverage near 2.04X.

RepeatExplorer pipeline is a computational pipeline, for producing read clustering from the provided reads sequence based on the similarity hit of sequence reads which is further characterized through RepeatMasker software (<http://www.repeatmasker.org>; Smit, et al., 2013-2015). Clusters generated from this pipeline for both *V. corymbosum* and *V. macrocarpon* were used for further analysis to identify each different types of repeat sequence based on their cluster graph shape, protein domain hits and repeat masker hits. Each cluster was then assigned to supercluster based on their annotation, similarity hits with connected read cluster and other information generated through the RepeatExplorer pipeline. After complete annotation of the each characterized cluster and supercluster, approximate quantification of the different types of repetitive DNA was performed through Microsoft Excel.

To understand the level of differences and similarity of *Vaccinium* genome, comparative sequence clustering was performed for blueberry and cranberry. Comparative sequence clustering is slightly different than individual clustering. The main difference is the preparation of input sequence data for comparative clustering when the sequence reads from both of the species were tagged and a combined single dataset were produced containing sequence reads represent the equal genome proportion from both of the

species. After that dataset was used for Repeat Explorer based clustering analysis. Here, we used equal genome coverage of 0.3X for both of the species.

Comparative cluster analysis aided to detect species-specific cluster through scatter plot analysis of the genome proportion of each different types of the repeat sequence. It was found that most of the species-specific clusters belong to satellite repeats having a star or circular graph-like structure. Therefore, those particular clusters were analyzed in detail through Geneious software version 6.8.1 (<http://www.geneious.com/>; Kearse et al., 2012). The protocol begins with the importing assembled contig for the specific cluster from RepeatExplorer run to the Geneious. After that, sorting all contigs based on their size, mapping all contig sequence against the longest contig sequence as a reference with Geneious mapper tool with fine tuning interaction parameter 10, extracting the consensus sequence from the mapped reads and finally using the identified consensus sequence to detect the monomer unit of the particular satellite repeats.

4.2.4 Identification and phylogenetic analysis of satellite clusters

Dotplot analysis was performed on the seven identified longest satellite contigs sequences using EMBOSS 6.5.7 dotmatcher tool from Geneious window (Rice et al., 2000). Graph showed the monomer unit as a parallel dot on the matrix and a putative monomer unit was extracted. Eventually, an artificial tetramer from the putative monomer sequence was constructed by adding four monomer units subsequently and used as a reference sequence for mapping with the all contigs once more time to construct a more robust consensus monomer sequence using Geneious mapper tools with default parameters (Kearse et al., 2012). Monomer unit generated through this way was subjected to blastn search in the publicly available assembled database enriched with different sequencing platform (BAC-sequencing, Illumina sequencing, Pyrosequencing) in Geneious window. The parameters for blastn search in Geneious window were Maximum hits 300, Match mismatch 2-3, Gap cost 5 2, Max E-value 1e-1, word size 11, Maximum target sequence 300. The consensus satellite monomer sequences identified from the assembled contig of blueberry and cranberry were subjected to phylogenetic dendrogram preparation to identify the relationship among different satellite repeats.

The protocol followed for phylogenetic dendrogram preparation was begun with multiple sequence alignments of consensus satellite sequences for both blueberry and cranberry. Geneious software plugin tools MAFFT, a tool to identify their pairwise similarity and creating multiple sequence alignment was used for this purposes (Kato and Standley, 2013). The parameters for MAFFT multiple sequence alignment tool was default in Geneious and described as follows: Algorithm = Auto (selects an appropriate strategy from L-INS-I, FFT-NS-i and FFT-NS-2 according to data size), scoring matrix = 200PAM/K = 2, Gap open penalty = 1.53, offsetvalue = 0.123. After that, the phylogenetic dendrogram was produced using the FastTree plugin tools of Geneious software platforms with default parameters (Price et al., 2010). For tree construction with default parameters FastTree use minimum-evolution subtree-pruning-regrafting (SPRs) and maximum-likelihood (NNIs) methods for tree construction with the “CAT” approximation. In addition, rate categories site 20 was used for the tree construction. Constructed tree was viewed in Geneious tree viewer tools. The dendrogram was viewed in rooted tree layout and node was colored according to species specific monomer. Scale bar (below the tree) represent the substitutions per site.

4.2.5 Phylogenetic and heterogeneity analysis of reverse transcriptase protein domain sequences

In order to detect main types of repetitive sequences such as retrotransposable elements, reverse transcriptase protein domain sequences were extracted from the publicly available assembled database for *V. corymbosum* and *V. macrocarpon* using the protein domain finder tools implemented in RepeatExplorer Galaxy instance (Novak et al., 2013). The extracted sequences were quality filtered (minimum identity: 0.3 minimum similarity: 0.4, minimum alignment length 0.8, Interruptions: 3). After filtering, the remaining sequences were subjected to clustering with 90% similarity hit with Cd-hit (Li and Godzik, 2006). Clustered core reverse transcriptase (RT) protein domain sequences were aligned with multiple sequence alignment tool, MAFFT a plugin implemented in Geneious with the default parameters (Kato and Standley, 2013). Resulting multiple sequence alignment was used for the creation of phylogenetic dendrogram using Maximum likelihood methods using the FastTree tools of Geneious with default parameters (Price et al., 2010).

To detect the comparative sequence heterogeneity among the identified families and lineages of the transposable element between blueberry and cranberry genome, pairwise sequence similarity value was recorded from MAFFT multiple sequence alignment. Average pairwise similarity and ratio of pairwise sequence similarity between the two species for each specific lineages of the transposable element were calculated using Microsoft Excel software. Data derived from such analysis were depicted in a scatter plot analysis through Microsoft Excel to reveal the levels of heterogeneity among the identified reverse transcriptase protein domain sequence between the two species.

4.3 Results

4.3.1 Genomic proportion of different types of repetitive DNA

In order to gain insight information about genome proportion of repetitive sequences, publicly available whole genome sequence data were used to perform graph-based sequence clustering for blueberry and cranberry. The results show that quantification of different types of repetitive DNA in *Vaccinium* are directly related with initial input read sequence and genome coverage used. Therefore, in order to get a consistent quantification of repeat sequence, three individual RepeatExplorer run with constant genome coverage 2.04X were performed for both blueberry and cranberry. The final quantification of total repeat sequences and their lineages was considered the average of three individual run. Eventually, standard deviation was calculated from the three individual RepeatExplorer run (Table 4.2.).

Table 4.2. Repeat Explorer based estimation of different repeat types in *V. corymbosum* and *V. macrocarpon*. (Standard deviation is calculated from three different individual RepeatExplorer based estimation)

Major repeat types	Family/lineages	Genome proportion (%)	
		<i>V. corymbosum</i>	<i>V. macrocarpon</i>
LTR-retrotransposon			
	ltr/copia/unspecified	0.18 ± 0.11	0.17 ± 0.03
	ltr/copia/maximus	0.06 ± 0.02	0.21 ± 0.00
	ltr/copia/tork	0.07 ± 0.02	0.71 ± 0.28
	ltr/copia/angela	0.34 ± 0.00	0.35 ± 0.02
	ltr/copia/bianca	0.11 ± 0.00	0.14 ± 0.02
	ltr/copia/tar	0.17 ± 0.01	0.72 ± 0.27
	ltr/copia/ivana	0.13 ± 0.01	0.11 ± 0.00
	ltr/copia/alei	0.13 ± 0.18	
	ltr/copia/aleii	0.42 ± 0.00	1.04 ± 0.39
	ltr/gypsy/unspecified	0.66 ± 0.11	1.01 ± 0.04
	ltr/gypsy/chromo	0.21 ± 0.03	0.89 ± 0.40
	ltr/gypsy/ogre_tat	1.64 ± 0.06	4.25 ± 1.89
	ltr/gypsy/athila	0.28 ± 0.05	1.17 ± 1.04
	ltr/unspecified	2.18 ± 0.83	5.07 ± 1.29
Total		6.62 ± 1.43	15.85 ± 5.67
Non-LTR-retrotransposon			
	line	0.56 ± 0.06	2.085 ± 0.52
	para	0.09 ± 0.01	0.131 ± 0.05
Total		0.66 ± 0.07	2.216 ± 0.57

Table 4.2. (Continued) Repeat Explorer based estimation of different repeat types in *V. corymbosum* and *V. macrocarpon*. (Standard deviation is calculated from three different individual RepeatExplorer based estimation)

Major repeat types	Family/lineages	Genome proportion (%)	
		<i>V. corymbosum</i>	<i>V. macrocarpon</i>
DNA-transposon			
	dna	1.69 ± 0.91	2.84 ± 0.35
	dna/cacta	0.13 ± 0.12	0.26 ±
	dna/hat	0.28 ± 0.04	0.48 ± 0.01
	dna/mutator	0.10 ± 0.01	0.62 ± 0.19
	dna/pif_harbinger	0.05 ± 0.01	0.34 ± 0.15
	helitron		0.09 ± 0.03
Total		2.25 ± 1.08	4.53 ± 0.73
rdna	45S	0.01 ± 0.00	0.86 ± 1.01
	5S	0.02 ± 0.00	
Total		0.03 ± 0.00	0.086 ± 1.01
Satellite		0.79 ± 0.44	0.033 ± 0.00
plastid		0.88 ± 0.10	4.38 ± 0.83
mitochondrion		0.08 ± 0.031	1.366 ± 1.05
Unknown		16.9 ± 0.85	8.835 ± 0.23
singlets		23.8 ± 5.091	8.83 ± 0.30
uncharacterized_clusters		48.25 ± 1.34	53.35 ± 9.98

Schematic pie chart developed from the average value of this analysis reveals a more realistic view and quantification of repeat sequence in both species. The results explained that a large proportion of blueberry and cranberry genome (79% and 91.4%) are actually repeat sequence or derived from the repetitive DNA while rest of them are mainly singlets (genic) or nonrepetitive sequence in blueberry and cranberry, respectively (Figure 4.2.). Among them 24% and 49% were LTR-retrotransposon, 2% and 6% belongs to non-LTR retrotransposon, 8% and 11% DNA transposon, 3% and 0.06% were satellite for blueberry and cranberry, respectively. Other annotated reads belong to rDNA, mitochondrial DNA and plastid DNA. Nonetheless, the overall characterized repeat sequence is different for two species. While characterized cluster in blueberry is 37%, for cranberry it is about 49%, possibly due to the quality and length of the repeat sequence of publicly available whole genome sequence data (Figure 4.2.).

Bar graph generated through the annotated and calculated superclusters and their genome proportion reveal the differences in the genome proportion of different types of repetitive DNA. In Figure 4.3., each bar represents each individual types of the repeat sequence, where the height of bar represents the total number of sequence read and the width represent their genome proportion. From this analysis, it is clear that the most abundant supercluster in blueberry belongs to tandem repeats group while in cranberry it belongs to L1-type, LINEs, and a non-LTR retrotransposon. This also indicates that abundant types of repeat sequences are the tandem repeats and non-LTR retrotransposons for blueberry and cranberry, respectively.

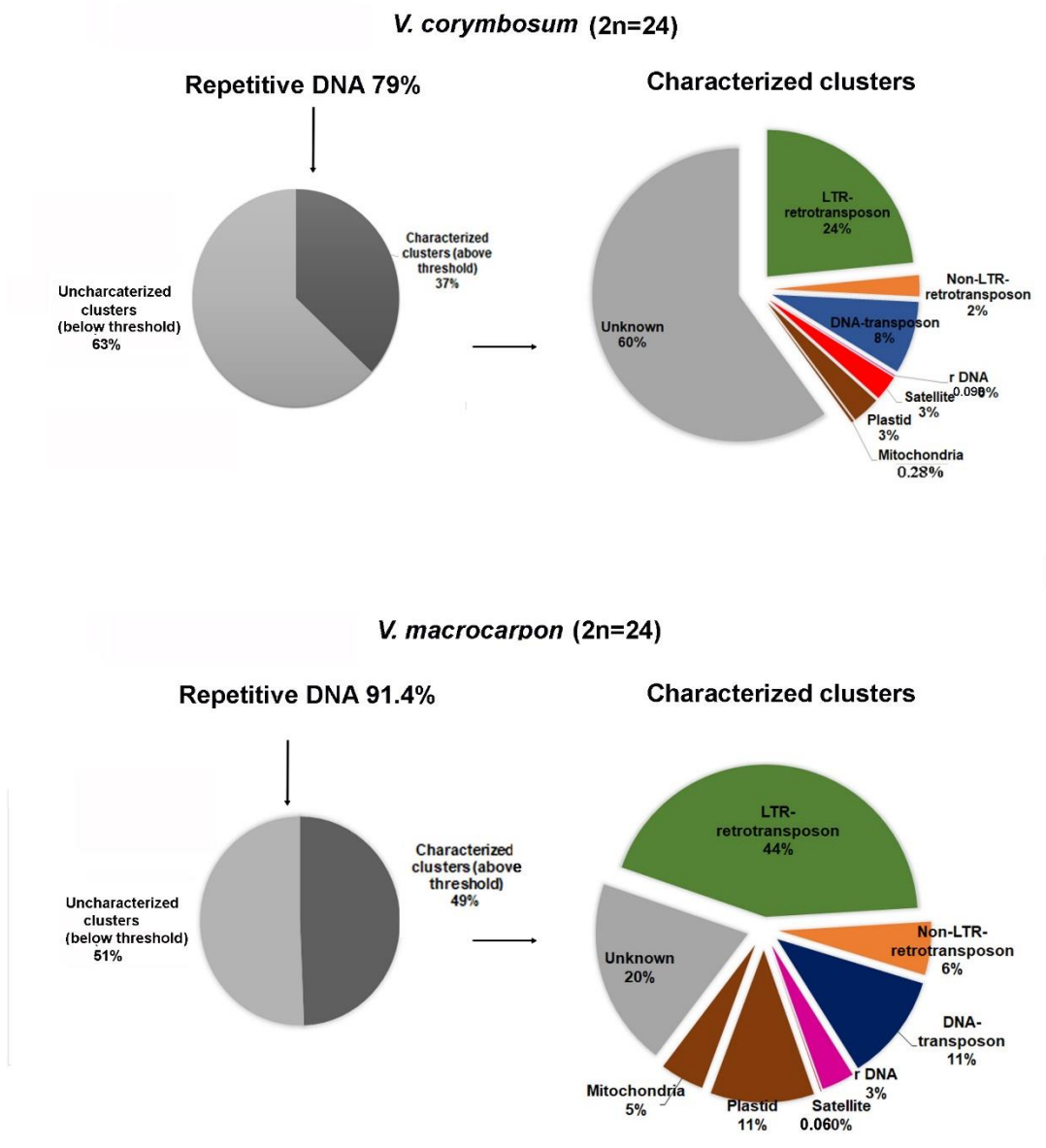
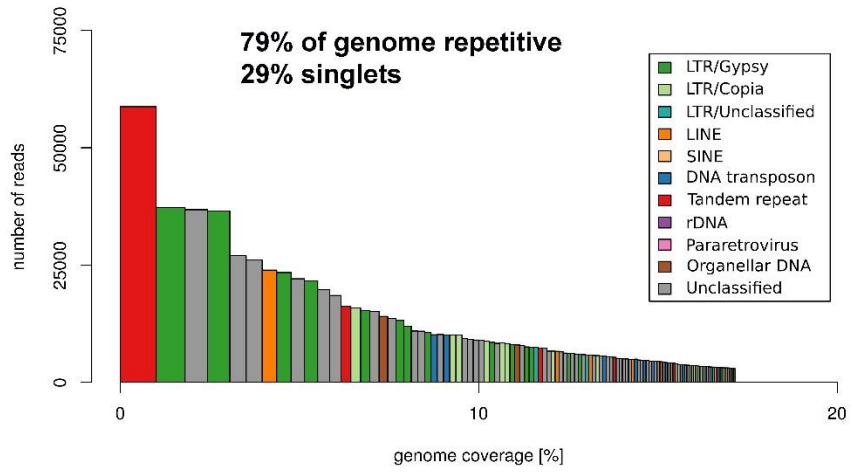


Figure 4.2. Quantification of different types of repeat sequence from *Vaccinium* genome

Repeats in *Vaccinium corymbosum* genome



Repeats in *Vaccinium macrocarpon* genome

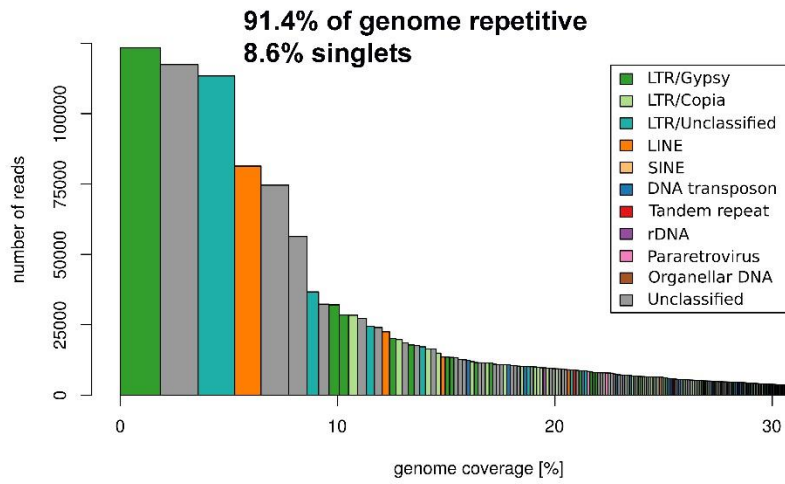


Figure 4.3. Supercluster of different types of repeat sequence identified in *Vaccinium* genome

4.3.2 Characteristics features of satellite DNA

Star typical graphs of RepeatExplorer output is considered for further analysis to identify satellite cluster. A total of seven different satellite repeats was found in *Vaccinium* genome through individual read clustering from blueberry and cranberry genomes. Satellite repeats are named as VaccSat1 to VaccSat7 depending on the sequential time of finding. However, in-depth analysis of individual clustering proves that clusters representing VaccSat1, VaccSat4, VaccSat5, VaccSat6 are only present in the blueberry genome but not in cranberry genome. In contrast, VaccSat2, VaccSat3 and VaccSat7 representing clusters are present in both blueberry and cranberry genomes (Figure 4.4. and Table 4.3.). Both VaccSat1 and VaccSat3 are subdivided into two smaller individual clusters and form one single big supercluster explained the presence of subgroup or subfamily (Figure 4.4. and Table 4.3.).

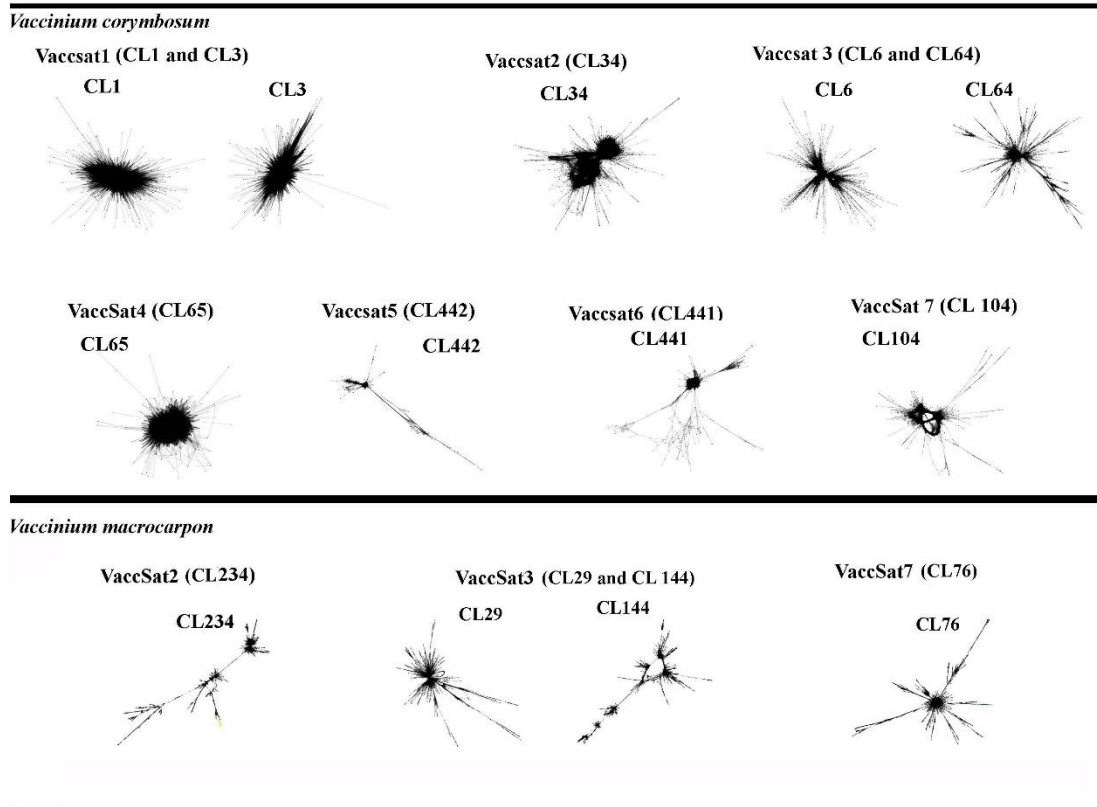


Figure 4.4. Star typical satellite graph identified in *Vaccinium* genome.

Table 4.3. Characteristics features of the satellite repeats annotated in blueberry and cranberry genome

Name of satellite	Monomer size (bp)	Average GC content [%]	<i>V. corymbosum</i>		<i>V. macrocarpon</i>	
			Super-cluster	Genome proportion (%)	Super-cluster	Genome proportion (%)
VaccSat1	146-147	18.4	SCL1	0.999	-	-
VaccSat2	238	40.9	SCL29	0.124	SCL155	0.066
VaccSat3	154	21.1	SCL5	0.36	SCL20	0.284
VaccSat4	101	17.4	SCL1	0.050	-	-
VaccSat5	36-38	19.9	SCL36 7	0.015	-	-
VaccSat6	49	22.4	SCL36 6	0.015	-	-
VaccSat7	49-70	31.4	SCL79	0.069	SCL56	0.133

In order to identify the genetic relationship among the seven identified satellite repeats of blueberry and cranberry, a phylogenetic dendrogram was constructed from the consensus contig sequence from the respective clusters of each species and artificial trimer sequence specific to each satellite repeats (Figure 4.5.). Our results suggest that VaccSat1 and VaccSat4 belong to the same satellite groups and can be considered as a single Satellite family. Therefore, total satellite repeats family in *Vaccinium* genome could be considered as six instead of seven. Even though VaccSat2 and VaccSat7 belong to two different branches there could be a significant sequence similarity between these two satellite families which need closer investigation. Overall, all satellite families are highly heterogeneous in both of the species.

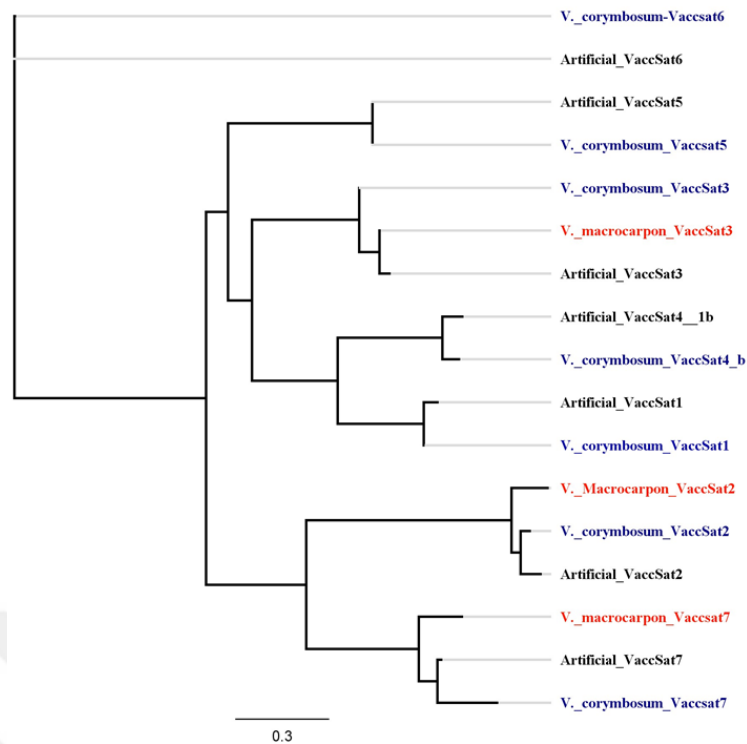


Figure 4.5. Phylogenetic relationship among seven different satellite repeats identified in *V. corymbosum* and *V. macrocarpon* genome

4.3.4 Transposable elements

To gain overall insight information about the transposable elements of *Vaccinium* genome reverse transcriptase protein domain sequences (RT) were extracted and clustered from both blueberry and cranberry assembled genome databases (Table 4.4.).

The maximum number of clustered RT sequences belongs to L1 type LINEs a non-LTR retrotransposon. The number is 249 and 269 for blueberry and cranberry genome respectively. Which suggests that L1 type LINEs sequences are one of the most diversified elements in both blueberry and cranberry genome. Moreover, all known Ty1/*copia* and Ty3/*gypsy* lineages are found in *Vaccinium* genome. The identified lineages and the clustered sequences in Ty3/*gypsy* lineages are as follows; Ogre/Tat (the most diversified) 123 and 183 followed by CRM 12 and 34, Tekay 20 and 28, Athila 12 and 25, Reina 8 and 15 for blueberry and cranberry, subsequently. Ale/retrofit lineages with the clustered sequence number 82 and 109 was the most diversified one within Ty1/*copia* families, followed by TAR 17 and 22, Ivana 10 and 12, Tork 6 and 12, SIRE

5 and 7, Ikeros 3 and 3, Bianca 4 and 3, Alesia 0 and 1 clustered sequences for blueberry and cranberry, respectively (Table 4.4.)

Table 4.4. Retrotransposon protein domain identified from the assembled genome and the sequence similarity of *V. corymbosum* and *V. macrocarpon*

Superfamily/clade	Family/Lineages	<i>V. corymbosum</i>		<i>V. macrocarpon</i>	
		Total RT protein domain sequence	Clustered sequences at 90% similarity threshold after removing invalid sequence	Total RT protein domain sequence	Clustered sequences at 90% similarity threshold after removing invalid sequence
Class_I LTR Ty1 copia					
	Ale	309	82	587	109
	Alesia	0	0	8	1
	Angela	7	0	7	0
	Bianca	15	4	34	3
	Ikeros	24	3	27	3
	Ivana	47	10	86	12
	SIRE	17	5	21	7
	TAR	62	17	105	22
	Tork	49	6	142	12
Class_I LTR Ty3 gypsy					
Chromovirus	CRM	48	12	105	34
	Galadriel	8	2	13	4
	Reina	10	8	25	15
	Tekay	87	20	116	28
non-chromovirus OTA	Athila	34	12	134	25
	Ogre/Tat	407	123	739	183
Class_I LINE		577	269	781	249
Class_I pararetrovirus		16	4	13	4

4.3.5 Comparative phylogenetic and heterogeneity analysis of retrotransposon elements from blueberry and cranberry

Phylogenetic analysis revealed that sublineages of both Ty3/*gypsy* and Ty1/*copia* superfamily produced separate clusters in the phylogenetic tree (Figure 4.6.). In case of Ty3/*gypsy*, Ogre/Tat lineages were the most diversified clade and possible to divide into seven individual clusters. Chromomovirus lineage produced four clusters representing sublineages CRM, Reina, Galadriel and Tekay. The remaining two clusters of Ty3/*gypsy* superfamily belonged to lineages Athila and pararetroviruses (Figure 4.6). On the other hand, Ty1/*copia* produced three major clusters in the phylogenetic tree. The most diversified cluster belonged to Ale/Retrofit lineages. Sublineages Tork, TAR, Ikeros and Bianca produced one single cluster. The remaining cluster was constituted with SIRE and Ivana (Figure 4.6). In addition, to Ty3/*gypsy* and Ty1/*copia*, LINE was one of the diversified elements in the *Vaccinium* genome (Figure 4.6)

Comparative heterogeneity analysis of reverse transcriptase (RT) domains showed that the extent of diversification among the clustered sequences were different in blueberry and cranberry genome (Table 4.5.). For instance, pararetroviruses were more diversified in blueberry compared to cranberry genome. On the other hand, lineages that were more heterogeneous in cranberry compared to blueberry were LINEs, Ikeros, Reina, Sire and Ogre/Tat (Table 4.5. and Figure 4.7.). However, lineages Tekay, CRM, Athila, Tar, Ivana, Tork, Ale, Bianca and Galadriel were almost equally heterogeneous in both of the species (Table 4.5. and Figure 4.7.).

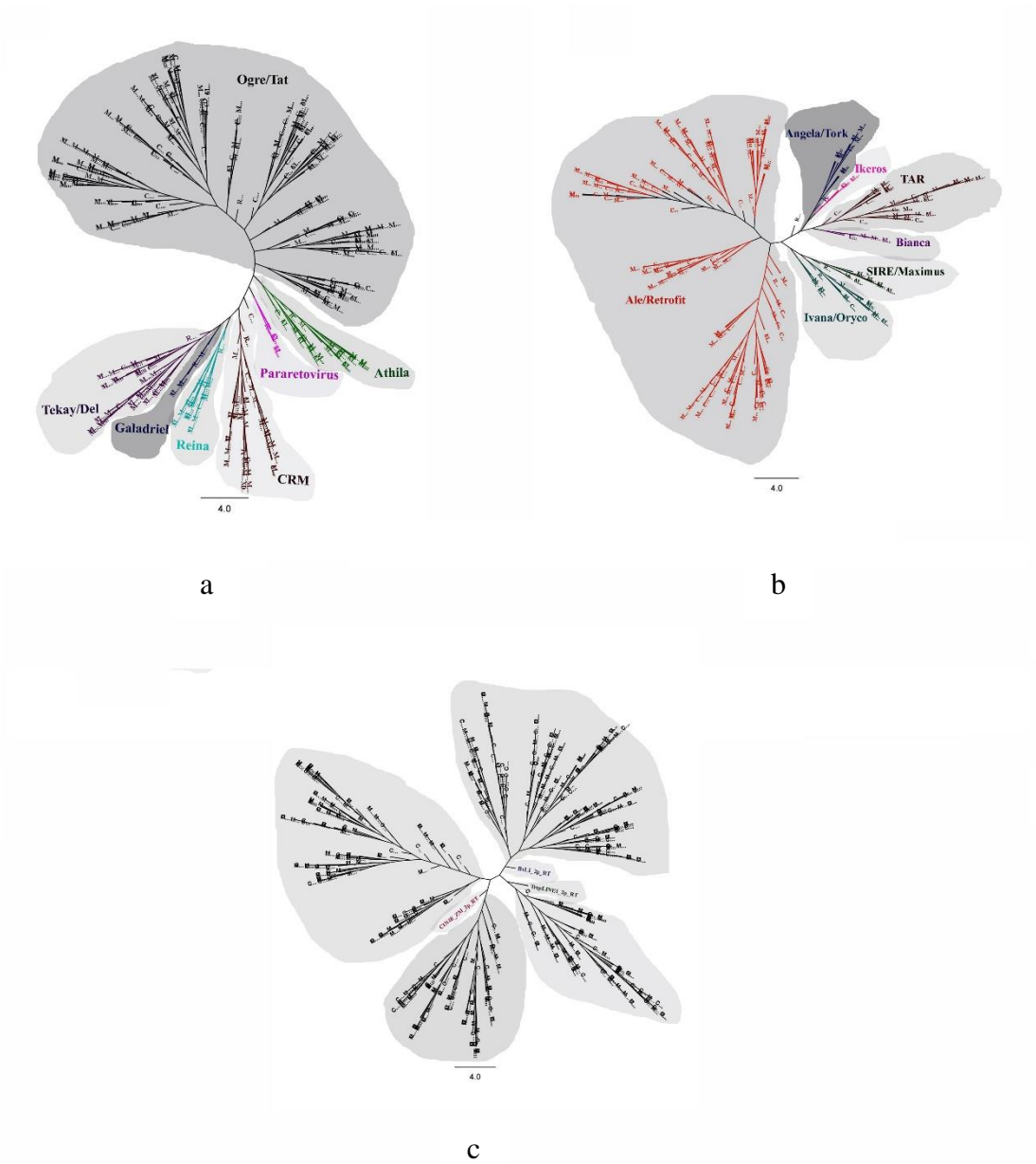


Figure 4.6. Phylogenetic relationship among reverse transcriptase domain sequence of different lineage of Ty3/gypsy (a), Ty1/Copia (b), LINE (c) in *V. corymbosum* and *V. macrocapon* genome

Table 4.5. Calculation of heterogeneity among identified reverse transcriptase protein domain sequences in *V. corymbosum* and *V. macrocarpon*

Superfamily/Family/Lineages	Pairwise Similarity (%) within the identified RT domain in <i>V. corymbosum</i>	Pairwise Similarity (%) within the identified RT domain in <i>V. macrocarpon</i>	Average pairwise (%) similarity for <i>V. corymbosum</i> and <i>V. macrocarpon</i>	Ratio of similarity percentage of <i>V. corymbosum</i> and <i>V. macrocarpon</i>
Class_I LTR Ty1 copia				
Ale	47.60	48.90	47.80	0.97
Alesia	-	-	-	-
Angela	-	-	-	-
Bianca	72.70	73.80	73.30	0.99
Ikeros	58.70	55.20	57.10	1.06
Ivana	56.50	57.90	57.50	0.98
SIRE	67.30	64.10	65.60	1.05
TAR	58.50	59.60	59.20	0.98
Tork	52.40	53.80	53.30	0.97
Class_I LTR Ty3 gypsy				
Chromovirus CRM	65.60	64.90	65.00	1.01
Chromovirus Galadriel	80.00	81.70	81.20	0.98
Chromovirus Reina	73.10	69.30	70.40	1.05
chromovirus Tekay	70.20	68.50	69.20	1.02
non-chromovirus OTA Athila	75.30	75.70	75.60	0.99
non-chromovirus OTA Ogre/Tat	57.40	55.30	55.90	1.04
Class_I LINE	38.30	37.00	37.40	1.04
Class_I pararetrovirus	54.70	60.90	55.80	0.90

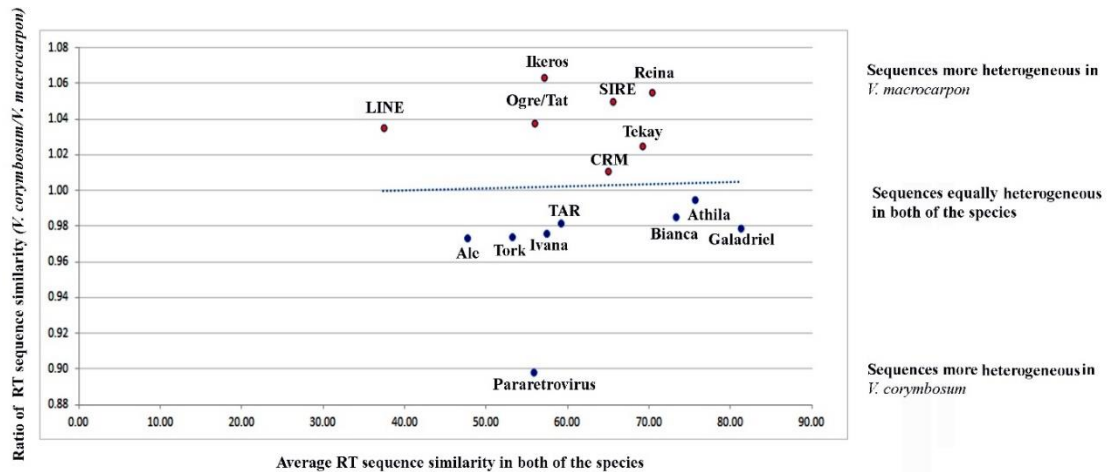


Figure 4.7. Differences in the heterogeneity of different lineage of reverse transcriptase protein domain sequences in *V. corymbosum* and *V. macrocarpon*.

4.3.6 Comparative clustering

The pipeline generated about 453 clusters with the lowest genome proportion of 0.01%. Clusters produced from this Repeat Explorer run were analyzed through scatter plot diagram considering the cluster size and comparative genome proportion of each cluster as an input value.

Comparative genome proportion for a single cluster was calculated with the formula below;

Comparative genome proportion for a single cluster = (genome proportion of one particular cluster in either blueberry or cranberry×100)/ Total genome proportion in blueberry and cranberry

Figure 4.8. shows that the arrangement of clusters in the scatter plot can be divided into three major groups. The first group are the clusters that accumulate at the middle of the plot area. They are common repeat elements for both of the species even if their genomic proportion could be slightly different for each species (Figure 4.8.). The second and third groups are the clusters that take place their position to the two-extreme side of the scatter plot which can be considered as the species-specific cluster. For instance, CL8 and CL9 (color red) with same genome proportion 0.46% were localized only in the area of

blueberry. After in-depth analysis of the contig sequence of these two clusters, it is found that they are species-specific satellite repeat clusters with the monomer size 147 bp named as VaccSat1. However, no representing cluster of VaccSat1 was found in the cranberry individual clustering. When the investigation was further expanded to assembled genome database using blastn search, the result was the same with the comparative clustering and individual clustering that means the VaccSat1 satellite repeats were absent in cranberry genome. Therefore, it was finalized that VaccSat1 is one of the species-specific satellite repeat cluster of *Vaccinium* genome (Figure 4.8.). On the other hand, CL102 localized in the area of cranberry, with genome proportion 0.016% were specific to cranberry genome, found to be a low confidence satellite repeats with monomer size 565. Moreover, there were some other genome specific clusters with low genome proportion for both blueberry and cranberry. Total number of species-specific clusters was 7 (1.5% of total characterized cluster) and 20 (4.41% of the total characterized cluster) for blueberry and cranberry, respectively (Figure 4.8.). In addition, CL399 (color blue) and CL 482 (color yellow) represent the identified satellite repeat VaccSat 5 (monomer size 36-38bp) and 6 (monomer size (49bp) for individual clustering of the blueberry genome. The positions of these two clusters in the scatter plot predict that they could be enriched in the blueberry genome but not in cranberry genome (Figure 4.8.).

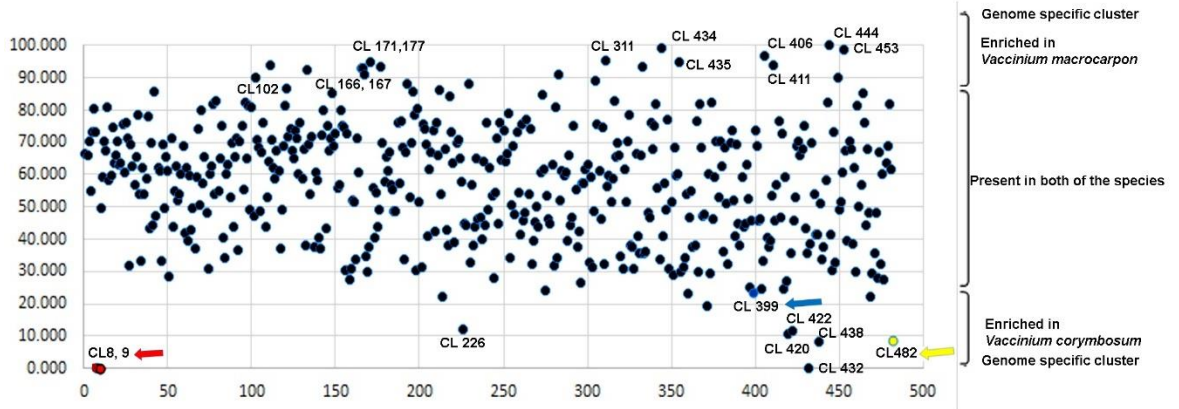


Figure 4.8. Comparative clustering of repeat sequence of *V. corymbosum* and *V. macrocarpon* showing the genome enriched repeat sequence in *Vaccinium* genome. (Color red, yellow and blue represent the satellite repeat VaccSat 1, VaccSat5 and VaccSat6, respectively)

CHAPTER V

CLONING AND MOLECULAR CHARACTERIZATION OF MAJOR SATELLITE REPEATS FROM CULTIVATED AND TURKISH WILD *VACCINIUM* SPECIES

5.1 Background Information

Vaccinium is a taxonomically intrigued genus. Due to extensive intra and interspecific hybridization, identification and characterization of different species of this genus that distributed all over the world is still a quite difficult task (Retamales and Hancock, 2012).

Tandemly organized satellite repeats of the genome are very often diversified and homogenized species specific manner and evolve rapidly in the evolutionary time scale. Therefore, identification and characterization of satellite repeats is an efficient technique to characterize phylogenetically closely related species. (Macas et al., 2006; Begum et al., 2013; Heitkam et al., 2015; Garrido-Ramos, 2017).

In this chapter an attempt has been taken to study satellite repeats from different *Vaccinium* species considering five different sections. Representative sections and species are; *Cyanococcus* (*Vaccinium corymbosum* L. cultivar ‘Jubilee’ and ‘Misty’), *Oxycoccus* (*V. macrocarpon* Ait., cultivar ‘Ben Lear’), *Hemimyrtillus* (*V. arctostaphylos* L.), *Myrtillus* (*V. myrtillus* L.) and *Vaccinium* (*V. uliginosum* L.). Cloning and bioinformatics analysis of repeat sequences in a comparative manner has been performed to examine the diversity, homogeneity and evolution of satellite repeats among different sections. The research focus of this chapter was to unveil the fascinating site of satellite repeats evolution of *Vaccinium* genome.

5.2 Materials and Methods

5.2.1 Collection of plant material and genomic DNA extraction

Blueberry cultivar ‘Jubilee’ and ‘Misty’ (*V. corymbosum* L.) has been grown in greenhouse conditions in Niğde Ömer Halisdemir University for last three years. For cultivated *Vaccinium* species, young leaf tissues were collected in early spring as leaves are breaking from dormancy from the full well-grown blueberry cuttings. For wild *Vaccinium* species, the leaf samples were collected from the field (Kaçkar Mountains, Rize, Turkey) in several occasions during 2016-2017 (July to August) and frozen in liquid nitrogen and lyophilized properly. Genomic DNA from the collected leaf samples for each representative species was extracted from both fresh and lyophilized leaves using DNeasy Plant Maxi kit (Qiagen, Germany) according to the manufacturer’s instructions with some modifications (increased material weight, buffer, RNase volume and increased incubation time) when necessary (Figure 5.1.). For instance, the recommended leaf sample was ≤ 1 g for fresh weight or ≥ 0.2 g for lyophilized tissue, but in case of *Vaccinium* about 3g of fresh tissue or about 1g of lyophilized tissue were used. Buffer volume and RNase were also increased with equal proportion of increased sample weight. Recommended incubation time of crushed sample in buffer was 10 min at 65°C, which was increased to 30 min at 65°C.

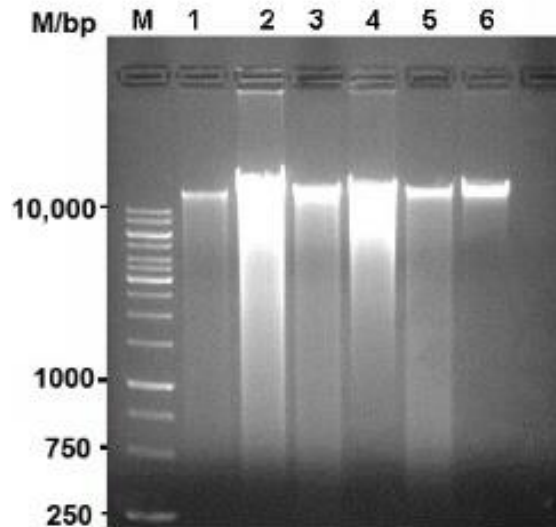


Figure 5.1. Extracted genomic DNA from the *Vaccinium* species. Lane M: 1 kb DNA Marker (ThermoScientific); Lane 1: *V. corymbosum* cultivar ‘Jubilee’; Lane 2: *V. corymbosum*, cultivar ‘Misty’; Lane 3: *V. arctostaphylos*- elongated fruit; Lane 4: *V. arctostaphylos*- round shaped fruit; Lane 5: *V. myrtillus*; Lane 6: *V. uliginosum*

5.2.2 Repeat specific primer designing

Repeat specific PCR primer was designed from the consensus sequence for each specific satellite repeats using Geneious bioinformatics platforms (<http://www.geneious.com>; Kearse et al., 2012). The main steps followed for primer designing was:

1. Each extracted monomer sequence from the RepeatExplorer pipeline was mapped against all contig sequences (RepeatExplorer output).
2. Each mapped assembly was inspected closely to see the region of the mapped area showing more homogeneous or conserved region.
3. The conserved region of the mapped assembly was taken into consideration for primer designing.
4. GC content and length of the primer was adjusted so that the stringency of the primer remains as much as possible (primer length more than 20 bp but less than 30 bp).

5. Properties of seven primer representatives of seven different satellite repeats are presented in (Table 5.1. and Figure 5.2.)

Table 5.1. Features of primer pairs for repeat specific satellite amplification.

Primer name	Primer Sequence (5'-3')	Product Size (bp)	G/C (%)	Tm Geneious
VaccSat 1 for	5'-ATTTAAAATGATTTTGTTCGC-3'	119	5	48.9
VaccSat 1 rev	5'-GCAAATAATAATGGTATTTAGC-3''		6	49.1
VaccSat 2 for	5'-GTACGGGCTACTGACCAC-3'	198	11	56.8
VaccSat 2 rev	5'-TATCGCTCAAACAACAAGTGG-3		9	56.8
VaccSat 3 for	5'-ATTTGACATTGTTGGCTTGC-3'	111	8	55.7
VaccSat 3 rev	5'-GATCTCAATTAGTAGTTTAATTTGGTG-3'		8	55.2
VaccSat 4 for	5'-GAATAATATTTTGTATAATATTTTTTC-3'	92	3	47.3
VaccSat 4 rev	5'-CAAAATTAATTTAGTTAATTTTCG-3'		4	47.2
VaccSat 5 for	5'-ATTAAATCCATTTAAATCATTTTCTG-3'	55	5	51.8
VaccSat 5 rev	5'-GATTTAAATGGATTTAATTAATAATCC-3'		5	51.3
VaccSat 6 for	5'-CTGACGGATTTTAAAAACGATG-3'	54	8	54.4
VaccSat 6 rev	5'-TCCGTCAGGTATTATTATGATTTTC-3'		8	55.1
VaccSat 7 for	5'-CAAGTTAGTTTTTTTGCAAAAC-3'	49-70	6	51.9
VaccSat 7 rev	5'-GTTTTGCAAAAAAACTAACTTG-3'		6	51.9

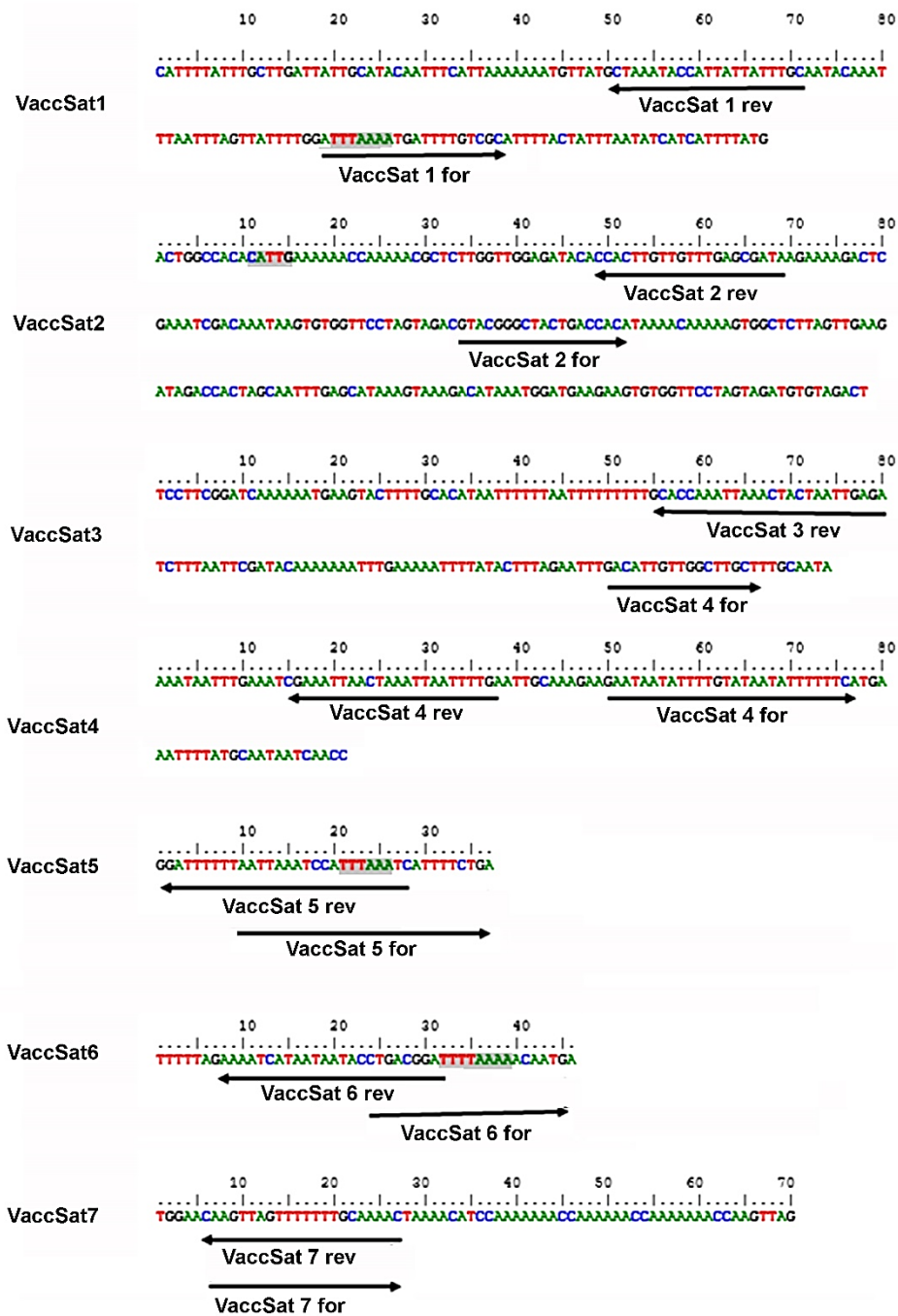


Figure 5.2. Satellite repeat sequence and primer designing site (black arrow).

5.2.3 PCR amplification and gel electrophoresis

Appropriate T_m temperature for each primer pair was determined after several rounds of gradient PCR. Polymerase chain reaction (PCR) was performed using template DNA 20-50 ng, forward primer 20 pM, reverse primer 20 pM, dNTP 10 mM, Dream taq polymerase (2.5 units) and green taq buffer (10X) (Thermoscientific). Standard PCR condition was 94°C for 3 min followed by 35 cycles of 94°C for 30 second, specific annealing temperature for 50 second, extension at 72°C for 50 second and final extension at 72°C for 5 min (Table 5.1.) Primer specific annealing temperature was determined through gradient PCR. Gradient PCR was done separately for each specific primer and each specific plant species.

PCR product was checked by 2% agarose gel electrophoresis (in 3–9 V/cm in 1 x TAE buffer (40 mM Tris-Acetate, 1 mM EDTA, pH 8.0) for 35 min, for appropriate PCR amplification with correct monomer size. After final confirmation of the correct band amplification, ligation was proceeded for cloning of DNA fragment.

5.2.4 Ligation of PCR product with pGEMT-easy vector system

Purification of PCR fragment (multimer of each satellite repeats) from the agarose gel was done with Invisorb Fragment CleanUp kit according to the manufacturer's instruction.

PCR product was cloned with pGEMT-Easy Vector System, according to the manufacturer's instruction (pGEMT-Easy Vector System, Promega Corporation).

5.2.5 Preparation of XL1-blue *E. coli* electrocompetent cell

Fresh cultures of XL1-Blue cells on LB plate with 20 µl Tetracycline (5mg/ml) were grown overnight and used as a pre-culture. Single colony from the pre-cultures was grown in 5 ml LB with 5 µl Tetracycline overnight at 37 °C. Main-culture was prepared from freshly grown cell in 1 L LB with 1 ml Tetracycline. Main culture was incubated in 3-4 h at 37°C in shaker up to OD 600 = 0.6-0.8 (= Correct growth phase). Cells were collected from the main culture through centrifugation for 10 min at 5,800 rpm (4°C) in 50 ml falcon tube. After collection, cells were resuspended in cold 10% glycerin by carefully

pipetting. The protocol was followed by washing of cells with 10% glycerin and collected through centrifugation for 10 min at 5,800 rpm (4°C). Washing step was continued two times. Depending on the final pellet size, cells were re-suspended in 1-2 ml cold, 10% glycerine and 50 µl of cells were aliquated on ice in 1.5 ml Eppendorf tube.

5.2.6 Transformation of XL1-blue electrocompetent bacterial cells with ligation product

For transformation, 1.5 µl of the ligation mixture was mixed with 25-50 µl of the XL1-blue electrocompetent bacterial cells in 0.2 cm cuvettes (Bio Rad) and electroporated with the Bio-Rad's Gene Pulser Xcell™ Electroporation machine with the pulse of voltage 1.8 kV. Transformed cells were immediately recovered in 1 ml of warm (37°C) liquid LB medium for 35 min at 37°C and finally let grown on ampicillin plates with IPTG and X-Gal for overnight at 37°C. The white colonies were assumed to be a positive transformed cell and blue colony assumed to contain empty vector. The white colonies were picked with the toothpick in LB medium with ampicillin and grown overnight at 220 rpm, 37°C for plasmid extraction.

5.2.7 Plasmid extraction and positive clone selection for sequencing

White bacterial colonies were grown in 3 ml LB medium. Plasmid from the grown bacteria was extracted using Thermo scientific GeneJET plasmid Miniprep kit with manufacturer's instruction (Figure 5.3.)

Extracted plasmid size was checked on 1.5% Agarose gel electrophoresis (in 3–9 V/cm in 1 x TAE buffer (40 mM Tris-Ac, 1 mM EDTA, pH 8.0) comparing with 1 kb ladder DNA. The pGEM-T Easy Vector size is 3015 bp. Any plasmid containing longer size than 3kb has been presumed to be positive and prepared for sequencing directly for highly abundant and perfect satellite repeats like VaccSat1, VaccSat2 and VaccSat3 (Figure 5.3.).

However, clones of satellite repeats showing less efficiency of the transformation were checked for positive insert by PCR using satellite specific primer pairs before sending to sequencing. For example, last three satellite repeats (VaccSat5, VaccSat6 and VaccSat7)

have smaller monomer size, species specificity, high genetic diversity or less abundance in the genome and hence show less efficiency during transformation (Figure 5.4).

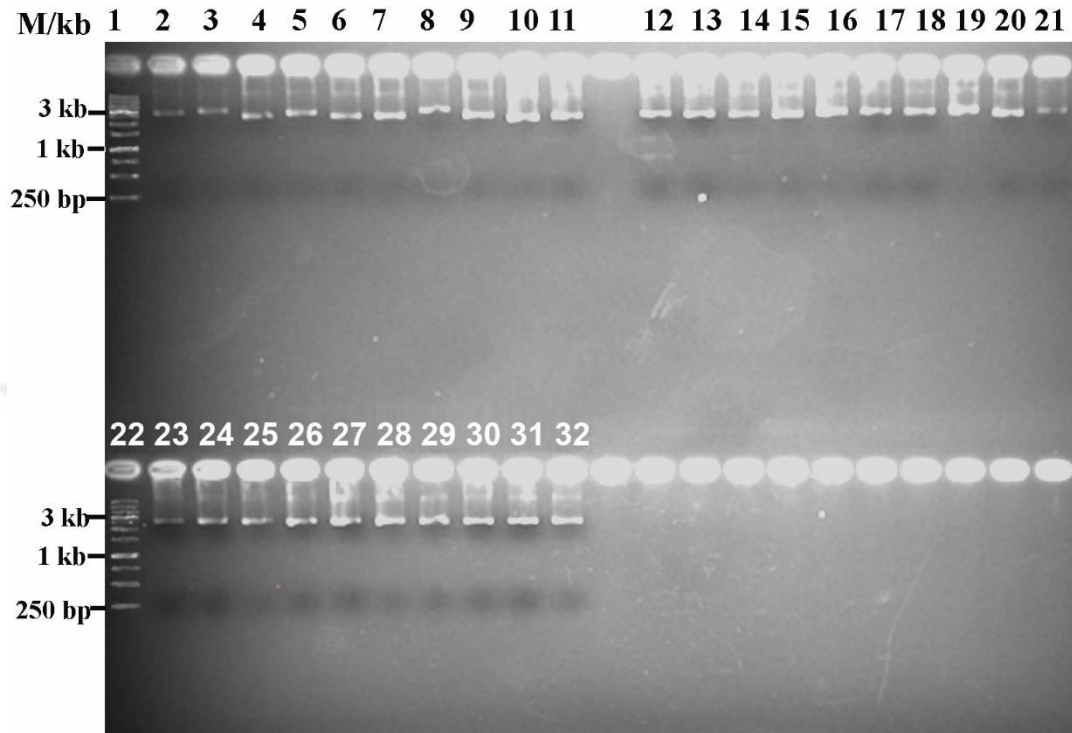


Figure 5.3. Representative photo for plasmid extraction and positive clone selection for VaccSat1, VaccSat2 and VaccSat3 from *V. corymbosum* cultivar 'Jubilee'. Lane 1 and 22: 1 kb DNA Marker (ThermoScientific), Lane 2-11 VaccSat1, Lane 12-21: VaccSat 2 and Lane 23-32: VaccSat3.

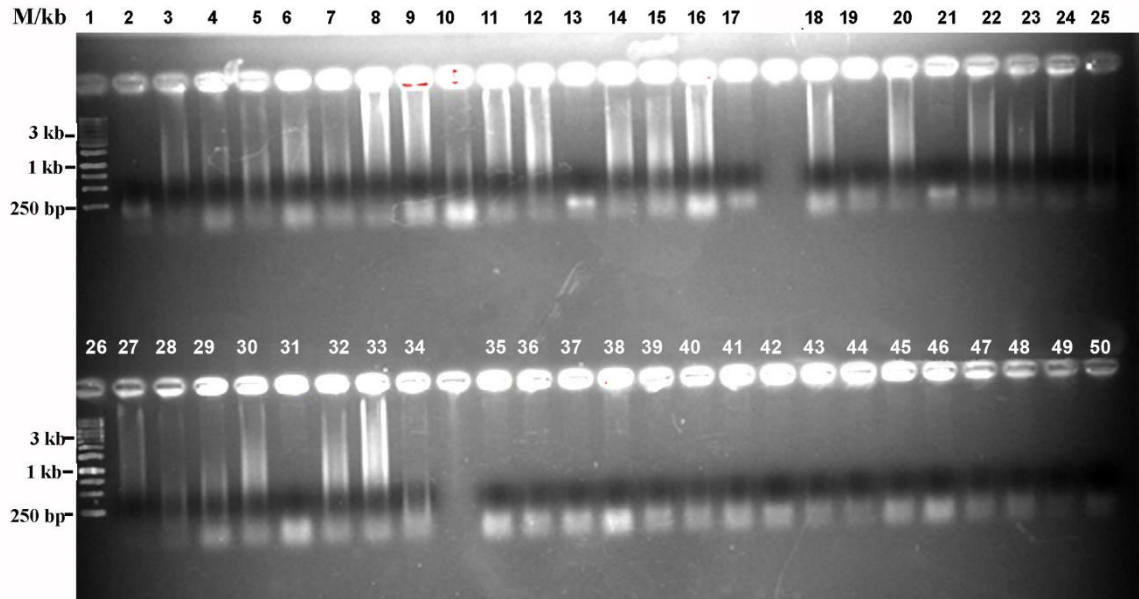


Figure 5.4. Representative photo for PCR amplification from extracted plasmid using satellite specific primer of VaccSat5 in three different *Vaccinium* species. Expected band size was 91 bp for trimer and 199 bp for hexamer. Lane 1 and 26: 1 kb DNA Marker (ThermoScientific), Lane 2-17: *V. arctostaphylos*, Lane 18-34 (except 26): *V. myrtillus* and Lane 35-50: *V. uliginosum*

5.2.8 Sequencing of plasmid

For each specific individual satellite repeat and each individual species, at least five positive plasmid clones were sequenced using automated sequencing machine, a service provided from two different companies (SENTEGEN (<https://www.sentegen.com/>) and BM laboratuvar Sistemleri (<https://www.bmlabosis.com/>)) located in Ankara, Turkey. For DNA sequencing SENTEGEN lab use ABI Prism 3130 Genetic Analyzer and BM laboratuvar use 23 ABI 3730XLs sequencer. Both machines use Sanger Sequencing technique for DNA sequencing.

Before preparing the order form for sequencing each sample was codified based on their satellite information and the plant species.

5.2.9 Computerized sequence analysis

Bioinformatics softwares used in this analysis are:

1. Geneious bioinformatics platform (<http://www.geneious.com>, Kearse et al., 2012)
2. NCBI blast
3. Tandem repeat finder (Benson, 1999)
4. PhyML V3.1 (Guindon et al., 2010)
5. HeatMapper (Babicki et al., 2016)

Identification of positive clone

Sequence misreading was corrected, analyzed and aligned using Geneious bioinformatics based platform (www.geneious.com, Kearse et al., 2012). Protocol followed for positive clone identification was; importing each clone sequence and assembled contig (*V. corymbosum* strain W8520) from RepeatExplorer run to the Geneious platforms. After that, all contigs were sorted based on their size and all contigs were mapped against the cloned sequence as a reference with Geneious mapper tool with fine tuning interaction parameter 10. If the successful mapped assembly found for a certain cloned sequence, it is thought to be positive clone but if no assembly found, the respective clone sequence was concluded to be not positive. After identification of positive satellite clone, multimeric clones were separated into individual monomers and used for further analysis.

Identification of multimeric sequences from assembled genome

To identify the multimeric sequences from the assembled database of *V. corymbosum* strain 'W8520' and *V. macrocarpon* cultivar 'Ben Lear', a artificial heptamer of each specific satellite monomer repeats was used to do blastn search on the custom assembled database sequences in Geneious bioinformatics platforms (Kearse et al., 2012). Multimeric satellite repeats from the scaffolds were extracted and monomer units was used for further analysis. The parameter for blastn search in Geneious windows was; Database = (either assembled database of *V. corymbosum* strain W8520 or *V.*

macrocarpon cultivar ‘Ben Lear’); Program = blastn, Results = Hit table, Retrieve = Full sequence with annotations, Maximum hits = 300, Low complexity filter (Scoring 2-3, Gap cost 5-2, Number of CPU 1, Max E-value 1e-1, word size 11, Max Target seqs 300)

Phylogenetic dendrogram and genetic distance calculation

The steps followed for phylogenetic dendrogram preparation and genetic distance calculation are summarized as follows;

1. Each monomer sequence was named properly respective to their satellite repeats and plant species (Table 5.2.).

Table 5.2. Short name for each plant species

Full scientific name and designation	Short name
<i>V. corymbosum</i> L. cultivar ‘Jubilee’	Vcor_cjubilee
<i>V. corymbosum</i> L. Strain W8520	Vcor_W8520
<i>V. macrocarpon</i> Ait. cultivar ‘Ben Lear’	Vmac
<i>V. arctostaphylos</i> L. (wild-Turkey)	Vacr
<i>V. myrtillus</i> L. (wild-Turkey)	Vmyr
<i>V. uliginosum</i> L.(wild-Turkey)	Vuli

2. For pairwise sequence alignment, MAFFT multiple sequence alignment tools was used as plugin from Geneious software platforms (Kato and Standley, 2013). The parameters for MAFFT multiple sequence alignment was: Algorithm = Auto (selects an appropriate strategy from L-INS-I, FFT-NS-i and FFT-NS-2 according to data size), scoring matrix = 200PAM/K = 2, Gap open penalty = 1.53, offset value = 0.123. Each satellite monomers from all species was taken at a time and grouped into 6 individual alignments (representing six individual satellite family). In these alignments, the shorter sequences were skipped and sequences with a proper monomer length were kept. These 6 alignments (VaccSat1 from all species, VaccSat2 from all species and up to VaccSat7 from all species) again were grouped and aligned. Pairwise distance matrix value and consensus sequence from each individual alignment was exported.

3. Tree was built from this alignment according to maximum likelihood algorithm using PhyML V3.1 for individual satellite family and a combined satellite family (Guindon et al., 2010). The parameter for this tree construction in Geneious window was: Substitution model = HKY85, Branch support = BootStrap, Number of bootstraps = 1000, Transition/transversion rate estimated, Proportion of invariable site 0, Number of substitution rates = 4, Gamma estimation parameter estimated, Optimize = Topology/length/rate, topology search = NNI. Trees generated from these analysis were viewed with Geneious tree viewer tool. Parameter for dendrogram preparation was; tree layout; unrooted, colored node considering species specific monomer.

Analysis of satellite sequence homogeneity

The pairwise similarity distance values generated from multiple sequence alignment for each group of satellite repeats were used to generate heatmap using HeatMapper (Babicki et al., 2016). For generating heatmap, distance matrix was calculated using Euclidean methods from the pairwise similarity distance value.

Each individual alignment representing each group of satellite repeats from different *Vaccinium* species was subjected to individual tree construction using FastTree plugin tools of Geneious software platforms with default parameters (Price et al., 2010). For tree construction with default parameters FastTree use minimum-evolution subtree-pruning-regrafting (SPRs) and maximum-likelihood (NNIs) methods for tree construction with the “CAT” approximation. In addition, rate categories site 20 was used for tree construction. Each tree constructed was viewed in Geneious tree viewer tools. The dendrogram was viewed in rooted tree layout and node was colored according to species specific monomer. Scale bar (below the tree) represent the substitutions per site.

Analysis of satellite repeats structure

For analyzing the subunit structure of satellite repeats Tandem repeat finder tool (Benson, 1999) was used together with Dotplot analysis with EMBOSS 6.5.7 dotmatcher tool from Geneious (Rice et al., 2000). Different parameters was used for Tandem repeat finder tools to detect the perfect subunit structure, for instance, alignment parameters (match, mismatch, indel=2,7,7; 2,5,7; 2, 5, 5; 2,3,5), minimum alignment score to report repeats

(20-60), minimum period size (10-300). The parameters for Dotplot analysis in Geneious window was; window: high sensitivity slow sliding, score matrix: exact, window size: 14 and Threshold: 27.

5.3 Results

5.3.1 DNA fingerprinting of cultivated and wild *Vaccinium* species using seven satellite specific primers

Figure 5.5., shows that all seven satellite primers produced amplified products in all the studied species including cultivated and wild, which also verifies that they are *Vaccinium* specific satellite repeats. Nevertheless, the intensity and banding pattern with correct monomer size of the ladder was different for different species which might be directly correlated with the abundance or availability of the satellite array in different *Vaccinium* species. VaccSat1, for example, produced clear satellite typical ladder-like pattern in all of the studied species with much more intensity and expected size of the first band at 119 bp (monomer to higher monomer motif). According to bioinformatics analysis VaccSat1 belongs to the first cluster with higher genome proportion and the most abundant satellite array in *V. corymbosum*. Although, *V. corymbosum*, *V. arctostaphylos* and *V. myrtillus* produced clear “hexamer”, *V. uliginosum* only produced “tetramer”. This may indicate much shorter array motif of VaccSat1 present in *V. uliginosum* or the satellite might be more diversified throughout the genome of this species (Figure 5.5). VaccSat4 belongs to the same satellite group of VaccSat1, and also produced similar banding pattern like VaccSat1 with expected monomer size of 94 bp, however, some additional shorter band is also visible in the gel parallel to each monomer. The additional short band could predict the presence of subunit-like structure of this first satellite repeats (Figure 5.5).

VaccSat2 also produced clear satellite typical head to tail banding pattern with correct monomer size of 198 bp. However, the intensity of the banding pattern was less compared to that of VaccSat1 or VaccSat4. This is expected with the bioinformatics analysis where cluster number of VaccSat2 was much higher with less genome proportion compared to VaccSat1. The maximum multimer motif amplified was only a dimer may indicate their much shorter or more diversified array motif compared to VaccSat1. The presence of the additional parallel band corresponding to subunit-like structure of the monomer band is

also visible for this satellite repeats. Nonetheless, subunit typical parallel band amplified differently for different species. For instance, the subunit band is much more prominent in wild *Vaccinium* species (*V. arctostaphylos*, *V. myrtillus* and *V. uliginosum*) compared to cultivate *V. corymbosum*. This may also indicate the different homogenization of monomer sequence for wild and cultivated *Vaccinium* species (Figure 5.5).

VaccSat3 was one of the most brightly amplified satellite repeats compared to other satellite repeats with correct monomer size of 111 bp. This may indicate the abundance of this satellite repeats among all *Vaccinium* species and a common satellite repeat. Nevertheless, the maximum multimer motif amplified is only dimer, indicating short array or diversified array motif for this satellite repeats. No additional parallel band with the main monomer band has been detected. This may also indicate that there is no visible subunit-like structure of this satellite repeat (Figure 5.5).

VaccSat5 was one of the most differentially amplified satellite arrays among *Vaccinium* species. The intensity of the banding pattern was also less compared to VaccSat1, 2, 3 or 4. This may indicate that VaccSat5 is a species-specific, less abundant satellite repeat of genus *Vaccinium*. Although the expected first band of 55 bp was clearly visible for all of the species, VaccSat5 could not produce satellite typical ladder-like pattern in *V. corymbosum* and *V. uliginosum*. However, ladder like pattern is observed in *V. arctostaphylos* and *V. myrtillus* species (Figure 5.5)

Similar to VaccSat5, VaccSat6 had also less intense satellite array motif, showed species specificity even within the same species. For instance, the satellite typical ladder-like pattern was observed in *V. corymbosum* cultivar ‘Jubilee’ but not present in *V. corymbosum* cultivar ‘Misty’. The expected first band was of 54 bp and higher motif till hexamer was present in the species *V. arctostaphylos*, *V. myrtillus* and *V. uliginosum* (Figure 5.5).

VaccSat7 was also a commonly amplified satellite repeat with correct monomer size of 79 bp. However, satellite typical banding pattern was not the same for all of the studied species. For instance, the higher monomer band amplified for this satellite was dimer for *V. corymbosum* and *V. uliginosum* but it was trimer for all other *Vaccinium* species like

V. arctostaphylos and *V. myrtillus*, which may indicate the differential array size in different species (Figure 5.5).

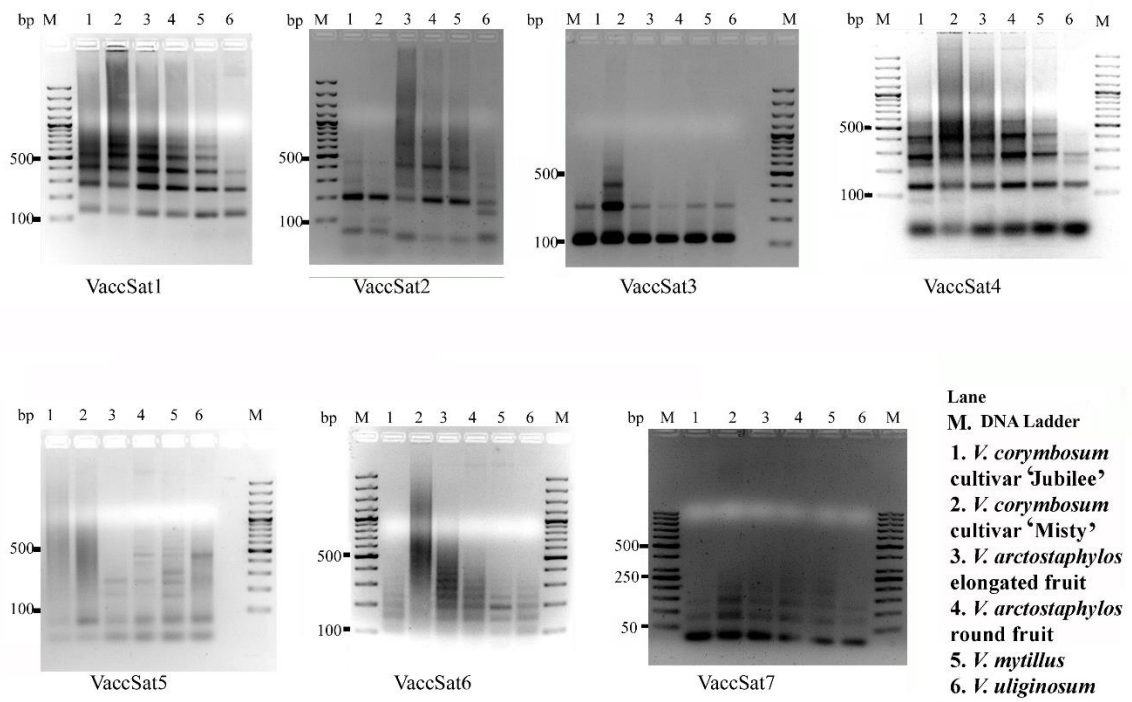


Figure 5.5. DNA fingerprinting of *Vaccinium* species with identified satellite repeats.

5.3.2 Diversity and evolution of satellite repeats among *Vaccinium* species

Clone characteristics

In order to investigate the satellite repeats diversity among different *Vaccinium* species, PCR-amplified satellite repeats were cloned from four different *Vaccinium* species (*V. corymbosum*, *V. arctostaphylos*, *V. myrtillus* and *V. uliginosum*), and cloned sequence were analyzed within a phylogenetic framework. Clone characteristics are shown in Table 5.3.

A total of 21 clones were sequenced from VaccSat1, of which 8 from *V. corymbosum*, 4 from *V. arctostaphylos*, 2 from *V. myrtillus*, and 7 from *V. uliginosum*. Sequence identity with the consensus varied from 67-97%. The most homogeneous group with higher sequence identity with the consensus belongs to *V. uliginosum* which is about 83-97%. Rest of the species show sequence identities varying from 67-92%. Therefore, VaccSat1 is one of the diversified satellite repeats among the species of *Vaccinium* except for *V. uliginosum* (Table 5.3.)

For VaccSat2, a total of 14 clones were sequenced, among these, 2 clones belong to *V. corymbosum* and 4 clones from each of the species of *V. arctostaphylos*, *V. myrtillus* and *V. uliginosum*. Sequence identity with consensus ranged from 71-90% considering all of the species. Nonetheless, species *V. arctostaphylos* and *V. myrtillus* show almost similar sequence identities ranging from 80-90%. Sequence identity for *V. corymbosum* was 86-88%. The lowest sequence identity for this group of satellite repeats belonged to *V. uliginosum* (71-79%) (Table 5.3.)

A total of 15 clones were sequenced for VaccSat3, among which 3 belong to *V. corymbosum* and the rest of the species have 4 clones each. Sequence identity ranged from 65-90% among all of the species. However, species *V. corymbosum* and *V. arctostaphylos* have similar sequence identities ranging from 84-90%. Moreover, *V. myrtillus* and *V. uliginosum* have similar sequence identity ranging from 65-90%. Therefore, VaccSat3 is also a highly diversified group of satellite repeats for *V. myrtillus* and *V. uliginosum* (Table 5.3.)

Due to the fact that satellite repeats VaccSat5, 6 and 7 have some tendency toward species specificity; cloned sequence was much less compared to VaccSat1, 2 and 3. The number of sequenced clones were 7, 10 and 9 for VaccSat5, VaccSat6 and VaccSat7, respectively. Sequence identity ranged from 88-97% for VaccSat5, 81-90% for VaccSat6 and 66-78% for VaccSat7 (Table 5.3.).



Table 5.3. Plasmid clone characteristics of *Vaccinium* satellite repeats

VaccSat1						
Name of the species	Clone name and insert size (bp)	Number of clone	Clone types	Number of monomer	Identity with consensus (%)	Range of consensus identity (%)
<i>V. corymbosum</i> Cultivar 'Jubilee'						
	VC-1-17 and 132	8	Monomer	16	91	78 - 92
	VC-4-3 and 138		Monomer		83	
	VC-3024 and 414		Trimer		92	
	VC-3025 and 268		Dimer		92	
	VC-3029 and 280		Dimer		81.30	
	VC-3030 and 446		Trimer		80	
	VC-3034 and 307		Dimer		78	
	VC-3035 and 287		Dimer		78	
<i>V. arctostaphylos</i>						
	VA-1 and 268	4	Dimer	11	92.60	68 - 92
	VA-2 and 415		Trimer		90.40	
	VA-4 and 328		Trimer		68	
	VA-7 and 415		Trimer		91	
<i>V. myrtilus</i>						
	VM-2 and 406	2	Trimer	4	86	67 - 86
	VM-8 and 61		Partial monomer		67	
<i>V. uliginosum</i>						
	VU-1 and 41	7	Partial Monomer	10	83	83 - 97
	VU-3 and 41		Partial Monomer		85	
	VU-6 and 329		Trimer		92.40	
	VU-9 and 41		Partial Monomer		85	
	VU-35 and 123		Monomer		90	
	VU-38 and 269		Dimer		95.60	
	VU-39 and 120		Monomer		97	

Table 5.3. (Continued) Plasmid clone characteristics of *Vaccinium* satellite repeats

VaccSat2						
Name of the species	Clone name and insert size (bp)	Number of clone	Clone types	Number of monomer	Identity with consensus (%)	Range of consensus identity (%)
<i>V. corymbosum</i> Cultivar 'Jubilee'						
	VC-2-1 and 445	2	Dimer	4	86	86 - 88
	VC-2-7 and 442		Dimer		88	
<i>V. arctostaphylos</i>						
	VA-11 and 194	4	Monomer	7	89.30	82 - 90
	VA-15 and 427		Dimer		90.60	
	VA-16 and 394		Dimer		82.40	
	VA-17 and 434		Dimer		83.20	
<i>V. myrtillus</i>						
	VM-11 and 496	4	Dimer	11	84	83 - 90
	VM-15 and 1087		Pentamer		80	
	VM-19 and 343		Dimer		90	
	VM-20 and 428		Dimer		83	
<i>V. uliginosum</i>						
	VU-15 and 115	4	Partial monomer	6	79	71 - 79
	VU-16 and 394		Dimer		75.80	
	VU-17 and 361		Dimer		76.70	
	VU-18 and 56		Partial monomer		71.40	

Table 5.3. (Continued) Plasmid clone characteristics of *Vaccinium* satellite repeats

VaccSat3						
Name of the species	Clone name and insert size (bp)	Number of clone	Clone types	Number of monomer	Identity with consensus (%)	Range of consensus identity (%)
<i>V. corymbosum</i> Cultivar 'Jubilee'						
	VC-3026 and 103	3	Monomer	3	90.50	88 - 90
	VC-3027 and 103		Monomer		88.60	
	VC-3028 and 103		Monomer		88.60	
<i>V. arctostaphylos</i>						
	VA-21 and 102	4	Monomer	4	84.80	84 - 90
	VA-22 and 110		Monomer		90	
	VA-23 and 111		Monomer		90	
	VA-27 and 120		Monomer		85.10	
<i>V. myrtilus</i>						
	VM-21 and 416	4	Trimer	11	74	65 - 91
	VM-24 and 612		Tetramer		65	
	VM-25 and 264		Dimer		91	
	VM-28 and 304		Dimer		90	
<i>V. uliginosum</i>						
	VU-21 and 262	4	Dimer	10	69	65 - 89
	VU-24 and 264		Dimer		65.60	
	VU-25 and 264		Dimer		89	
	VU-28 and 568		Tetramer		89.90	
VaccSat5						
<i>V. corymbosum</i> Cultivar 'Jubilee'						
	VC-5-1 and 41	1	Monomer	1	88.40	
<i>V. arctostaphylos</i>						
	VA-45 and 116	4	Trimer	20	88	88 - 96
	VA-46 and 301		Nonamer		96	
	VA-47 and 55		Dimer		96	
	VA-48 and 199		Hexamer		95	
<i>V. myrtilus</i>						
	VM-57 and 55	2	Dimer	5	96.40	96 - 97
	VM-58 and 91		Trimer		97	

Table 5.3. (Continued) Plasmid clone characteristics of *Vaccinium* satellite repeats

VaccSat6						
Name of the species	Clone Name and Insert size (bp)	Number of clone	Clone types	Number of monomer	Identity with consensus (%)	Range of identity with consensus (%)
<i>V. corymbosum</i> Cultivar 'Jubilee'						
	VC-6-1 and 66	1	Monomer	1	84.30	
<i>V. arctostaphylos</i>						
	VA-31 and 286	4	Hexamer	14	90	89 - 90
	VA-32 and 93		Dimer		90	
	VA-33 and 199		Tetramer		90	
	VA-34 and 97		Dimer		89	
<i>V. myrtillos</i>						
	VM-31 and 193	5	Tetramer	19	81.50	81 - 88
	VM-32 and 193		Tetramer		81.50	
	VM-33 and 193		Tetramer		81.50	
	VM-34 and 193		Tetramer		88	
	VM-35 and 144		Trimer		88	
VaccSat7						
<i>V. arctostaphylos</i>						
	VA-35 and 111	5	Trimer	13	70	66 - 75
	VA-39 and 137		Trimer		75	
	VA-41 and 64		Dimer		66	
	VA-42 and 126		Trimer		73	
	VA-44 and 96		Dimer		75	
<i>V. myrtillos</i>						
	VM-46 and 118	4	Trimer	10	69	69 - 78
	VM-47 and 115		Trimer		69	
	VM-52 and 77		Dimer		78.50	
	VM-55 and 72		Dimer		74.20	

Satellite repeats diversity in different *Vaccinium* species

To find out the relationship and diversity among *Vaccinium* satellite family, a total of 149 full length monomer repeats representing seven satellite families from six different species (*V. corymbosum*-cultivar 'Jubilee', *V. corymbosum*-strain W8520, *V. macrocarpon*, *V. arctostaphylos*, *V. myrtillus* and *V. uliginosum*) were analyzed through phylogenetic framework (Figure 5.6.).

Result showed that all satellites clearly grouped in individual branches representing each individual satellite repeats. Nevertheless, exceptions were also found particularly in the very diverse Sat5 clade, where two Sat5 sequences go to the Sat6 clade. Sat5 is highly diverse. Therefore the most divergent members of other families might go to this clade. For example, one monomer from VaccSat3 was aligned with VaccSat5. In addition to that VaccSat2 and VaccSat7 are closely related and fell in the same cluster even some of the VaccSat7 monomer sequences were with the VaccSat2 clade. This may indicate the origin of VaccSat2 and 7 from same seeds of sequence and diversification. Moreover, VaccSat6 and VaccSat3 also had high level of sequence similarity and clustered together (Figure 5.6.).

Generally, all sequences seem to be highly homogeneous i.e. most of the diversified monomer sequence is shared by all of the species.

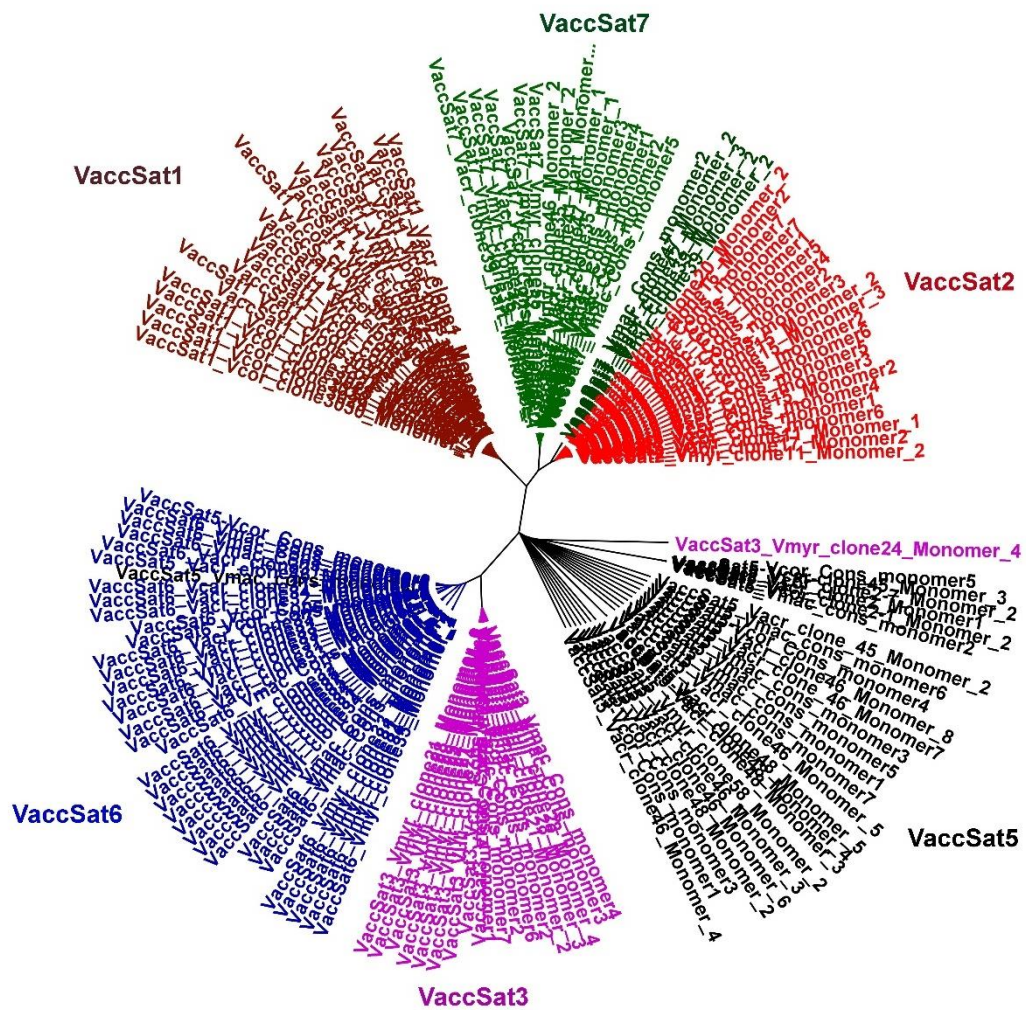


Figure 5.6. Phylogenetic dendrogram from 149 monomer sequences from seven different satellite families of six studied *Vaccinium* species with maximum likelihood algorithm using PhyML V3.1 (Guindon et al., 2010). (For monomer name Table 5.3. was referred).

Structure, homogeneity and evolution of individual satellite families

Monomers from each individual satellite families of different species were analyzed using dot-plot analysis to find out the internal subrepeats that could explain their possible origin. Multiple sequence alignment of the consensus monomer sequences were performed to find out the conserved and polymorphic site in the monomer sequences. Phylogenetic clade analysis of each satellite families revealed the possible divergent clade of satellite groups (Table 5.4.-5.6. and Figure 5.7.-5.11.)

VaccSat1

Dot plot analysis from randomly drawn multimer clone sequences from different *Vaccinium* species revealed that VaccSat1 had a varied monomer length ranging from 146-148 bp. Moreover, this satellite repeats had a periodicity of smaller subunit arranged in higher order repeating structure. This repeating structure was present in all studied plant species (*V. corymbosum*, *V. arctostaphylos*, *V. myrtillus* and *V. uliginosum*) but more prominent in *V. uliginosum* compared to other three species (Figure 5.7a). Figure 5.8a, shows the higher order repeats structure of VaccSat1, where two smaller directly repeated subunits (S1) of 28 bp were followed by a higher subunit (S2) of 94 bp. Therefore, a single monomer of VaccSat1 has a subunit repeat structure organized in S1+S1+S2 (Table 5.4.). Presence of subunit structure or subunit motif predict the origin, proliferation and diversification of larger satellite repeats motif from the smaller repeat motif. Nevertheless, smaller repeat motif could be blurred in some species due to mutation and homogenization but larger motif become fixed and more visible.

To investigate the nature of diversity of VaccSat1 in different *Vaccinium* species, a phylogenetic dendrogram from a total of 46 monomer sequences from four different species has been prepared using maximum likelihood algorithm and FastTree tools from Geneious plugin. From the dendrogram it was found that, VaccSat1 have three main monomer clusters. The first biggest cluster is mostly composed of monomer unit from *V. corymbosum* cultivar-‘Jubilee’ along with some diverse monomer from Turkish wild species (*V. arctostaphylos*). Moreover, monomer repeats from the wild Turkish *Vaccinium* species produce two individual clusters with mixed species distribution. Even though cultivated species and wild *Vaccinium* species show some preferential

accumulation of certain types of monomer unit, this claim further needs to be clarified with much more random clone analysis from taxonomically diverse *Vaccinium* species (Figure 5.9a)

Multiple sequence analysis from the constructed consensus monomer sequence revealed the conserve and polymorphic site of monomer unit. It was found that the polymorphic site is shared among different species (Figure 5.10a). Average genetic distance calculated from the consensus monomer sequences showed that overall genetic distance from different species ranged from 87-97% (Table 5.5.)

To investigate the nature of homogeneity of VaccSat1, the value of percentage of identical bases/residues from 44 monomer units were used to generate heat map. It was very clear that VaccSat1 is a heterogeneous satellite family where variable monomer unit is present between inter and intra-species level. In addition, one monomer unit from *V. uliginosum* shows dramatic heterogeneity. No species specific variants were detected for VaccSat1 (Figure 5.11a)

VaccSat2

To investigate if there is any periodic unit or subunits present in case of VaccSat2, dot plot analysis was performed from the multimer sequenced clone. It was found that the overall monomer length varied from 236-240 bp, constituted with the two direct subunit-like structure (S1+S1) and present in all studied species. However subunit organization into higher order structure was more homogeneously present in *V. corymbosum* cultivar 'Jubilee' and *V. arctostaphylos* but less prominent or fragmented in *V. myrtillus* and *V. uliginosum* (Figure 5.7b and Table 5.4.). The probable subunit size was directly repeated 122 bp sequences (Figure 5.8b). Similar to VaccSat1, presence of subunit like structure in VaccSat2 also predict the origin and evolution of this satellite motif from smaller motif.

Phylogenetic dendrogram from 37 monomer units of VaccSat2 satellite repeats shows that this satellite repeats constituted with four different monomer clusters. The first cluster was mainly composed of cultivated *Vaccinium* species (*V. corymbosum* cultivar 'Jubilee' and *V. macrocarpon*) with some occasional sequences from *V. myrtillus* and *V. uliginosum*. Rest of the two clusters is composed of both wild and cultivated *Vaccinium*

species. The analysis for this satellite repeats is mostly similar with VaccSat1 where the first cluster was also accumulated with cultivated *Vaccinium* species (Figure 5.9b). Therefore, VaccSat2 is also a heterogeneous satellite repeat family with different monomer unit. Heterogeneity in monomer unit is a common phenomenon for all of the species (Figure 5.9b)

Multiple sequence alignment from consensus monomer sequences of VaccSat2 from different *Vaccinium* species revealed that conserved and polymorphic sites are shared among the species (Figure 5.10b). Genetic distance table from the consensus monomer sequence showed that average sequence similarity among different species ranged from 85-97% (Table 5.5.).

To investigate the nature of homogeneity, a heat map analysis was performed from 37 monomer unit of VaccSat2 from five different species. It was very clear that VaccSat2 is a heterogeneous satellite family while variable monomer unit is present between inter and intra-species level. Although, Turkish wild *Vaccinium* species (*V. arctostaphylos*, *V. myrtillus* and *V. uliginosum*) had the most diverse monomer unit no species specific pattern was visible (Figure 5.11b)

VaccSat3

Dot plot analysis for VaccSat3 satellite clones did not reveal any consistent pattern of the periodic subunit in VaccSat3. This phenomenon could indicate that the origin and proliferation of VaccSat3 satellite repeat is different than the satellite shows subunit like structure. However, further investigation is necessary to identify the possible source of this satellite repeats.

Phylogenetic dendrogram from the extracted monomer sequences for VaccSat3 satellite repeats revealed that there were a total of seven individual clusters for this satellite repeats, which make this satellite repeats one of the most diversified satellite repeats of *Vaccinium*. As a common pattern, the first cluster represents the cultivated *Vaccinium* species (*V. corymbosum* and *V. macrocarpon*) together with the monomer from the wild species (*V. arctostaphylos*). On the other hand, *V. myrtillus* and *V. uliginosum* have an

individual cluster. Therefore, some of the species could have a preferential accumulation of certain types of monomer units (Figure 5.9c).

Multiple sequence alignment from consensus monomer sequences of VaccSat3 satellite revealed the conserved and polymorphic sites of repeat sequences in different *Vaccinium* species. It was found that mutated sites was both species specific and shared among all the species (Figure 5.10c). Average genetic distance among the species ranged between 89-98% much higher than VaccSat1 and VaccSat2 (Table 5.5.).

Heatmap analysis of pairwise percentage similarity values of bases from all extracted monomer sequences explained that pattern of heterogeneity was similar like VaccSat1 and VaccSat2 with occasional dramatic heterogeneity in *V. myrtillus* and *V. macrocarpon* (Figure 5.11c). Similar to VaccSat1 and VaccSat2, VaccSat3 is a heterogeneous satellite with variable monomer unit within intra and inter-species level. In addition, no species specific amplification was present for VaccSat3 satellite repeat family (Figure 5.11c)

VaccSat5

Multimer satellite clones were only extracted from *V. arctostaphylos* and *V. myrtillus*. Nonetheless, no multimeric unit had been extracted for *V. corymbosum* (cultivar 'Jubilee') and *V. uliginosum*. Both of the multimers from *V. arctostaphylos* and *V. myrtillus* showed higher order repeat structure. Dot plot analysis for VaccSat5 revealed that higher order satellite repeat structure was constituted with two different subunits (S1+S2) (Figure 5.7c and Table 5.4.). Figure 5.8c showed VaccSat5 monomer constituted with 10 bp smaller subunit (S1) and 26 bp longer subunit (S2) which were directly repeated. Both *V. arctostaphylos* and *V. myrtillus* had similar types of subunit pattern. Therefore, origin and evolution of this satellite repeats is similar to VaccSat1 and VaccSat2.

Phylogenetic dendrogram created from the monomer sequences of five different *Vaccinium* species revealed that VaccSat5 had four major clusters or four major variants of monomer unit. The first cluster was made up with *V. macrocarpon*, *V. corymbosum* and *V. arctostaphylos*, the second cluster with *V. corymbosum* and *V. arctostaphylos*, the third cluster with *V. myrtillus*, *V. macrocarpon* and *V. arctostaphylos* and the fourth

cluster with *V. arctostaphylos* and *V. myrtillus*. One of the biggest clusters was only composed of two cultivated and one wild *Vaccinium* species (*V. arctostaphylos*). Rest of the clusters was composed of different species (Figure 5.9d).

Multiple sequence alignment from the consensus monomer sequences of five different *Vaccinium* species revealed the conserved and polymorphic regions of the sequences among different *Vaccinium* species. It was found that most of the polymorphic sites were species specific (Figure 5.10d). Average genetic distance among different species ranged from 86-100% (Table 5.5.).

Heatmap analysis of pairwise similarity value from 38 monomer unit from different *Vaccinium* species (the cloned sequences of *V. arctostaphylos* and *V. uliginosum* as well as the monomer unit derived from the assembled database of *V. corymbosum* and *V. macrocarpon*) revealed that sequences were more homogeneous in *V. arctostaphylos* and *V. myrtillus* than the species of *V. corymbosum* and *V. macrocarpon*. This analysis revealed that VaccSat5 had the tendency of species specific homogenization (Figure 5.11d)

VaccSat6

Similar to VaccSat5, multimeric satellite clones for VaccSat6 were only extracted from *V. arctostaphylos* and *V. myrtillus* but not for *V. corymbosum* (cultivar 'Jubilee') and *V. uliginosum*. Higher order repeat structure was detected in both of the species (*V. arctostaphylos* and *V. myrtillus*) after dot plot analysis on the multimeric cloned sequences. However, dot plot graph showed slightly different pattern of two directly repeated subunit (S1+S2) structures for *V. arctostaphylos* and *V. myrtillus* (Figure 5.7d). Construction of higher order repeat structure of VaccSat6 with 17 bp smaller subunit (S1) and 32 bp bigger subunit (S2) is presented in Figure 5.8d. The presence of smaller subunit is the seeding smaller motif sequence from which the larger unit originate. This is a common phenomenon of other *Vaccinium* satellite like VaccSat1, 2 and 5.

Phylogenetic dendrogram from 52 monomer unit from four different *Vaccinium* species (extracted monomer from cloned sequences of *V. arctostaphylos* and *V. myrtillus* as well as extracted monomer sequences from assembled database of *V. corymbosum* strain

'W8520' and *V. macrocarpon* cultivar 'Ben Lear') revealed that there was two strict clusters for this satellite repeat. Cluster one was a mixed cluster containing monomer from four different species (*V. arctostaphylos*, *V. macrocarpon*, *V. myrtillus* and *V. corymbosum*). Nevertheless, cluster two contains species mostly from *V. myrtillus* and *V. macrocarpon*. Some of the sequences were highly diversified and produced individual clusters (Figure 5.9e).

Multiple sequence alignment of consensus monomer sequences of VaccSat6 from five different *Vaccinium* species reveals the polymorphic and conserved sites of consensus sequences from five different species. It was found that polymorphic sites were mostly species specific (Figure 5.10e). Average genetic distance among the species ranged between 82-96% which was similar with other studied species (Table 5.5.).

Heatmap analysis from the pairwise sequence similarity value from the 52 monomer units revealed that sequences were highly heterogeneous inter species level but comparatively more homogeneous interspecies level. These results could suggest that VaccSat6 had species specific homogenization and amplification (Figure 5.11e).

VaccSat7

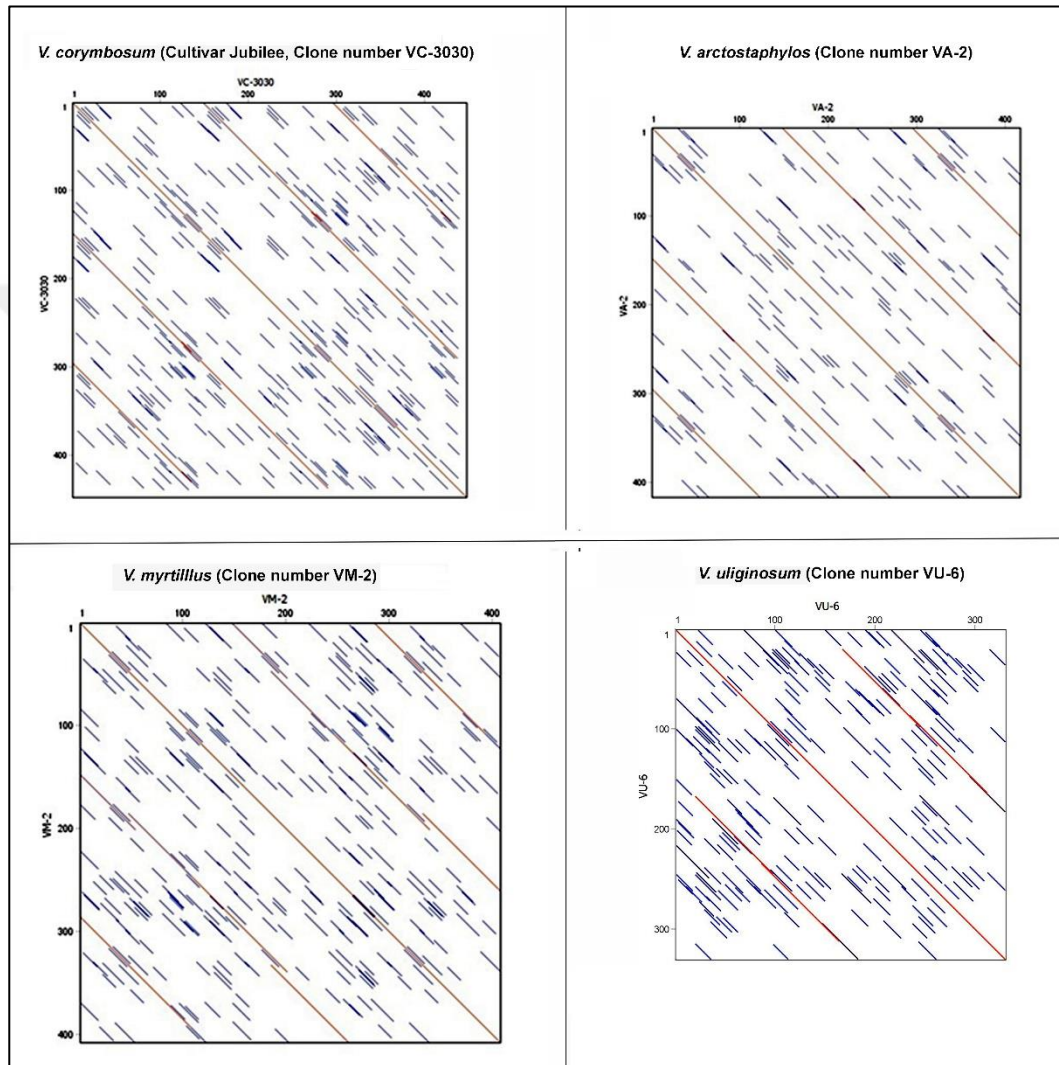
Dot plot analysis from multimer clone sequences from different species does not produce any subunit-like structure. Therefore, similar to VaccSat3, VaccSat7 also do not amplify from the smaller subunit but from different seed sequence motif.

Phylogenetic dendrogram analysis from 27 monomer units showed that this satellite repeats have a total of three clusters. All the clusters had monomer from different species. Therefore, none of the clusters was species-specific (Figure 5.9f).

Multiple sequence alignment of the consensus monomer sequences for this satellite repeats revealed the conserved and highly heterogeneous region of the monomer unit. However, the heterogeneous region was mostly composed of "CAAAAAA" motif which was appeared as a gap in *V. arctostaphylos* and *V. myrtillus* but present in *V. macrocarpon* cultivar 'Ben Lear' and *V. corymbosum* strain 'W8520' (Figure 5.10f). Average genetic distance for the consensus monomer sequence for this satellite repeats

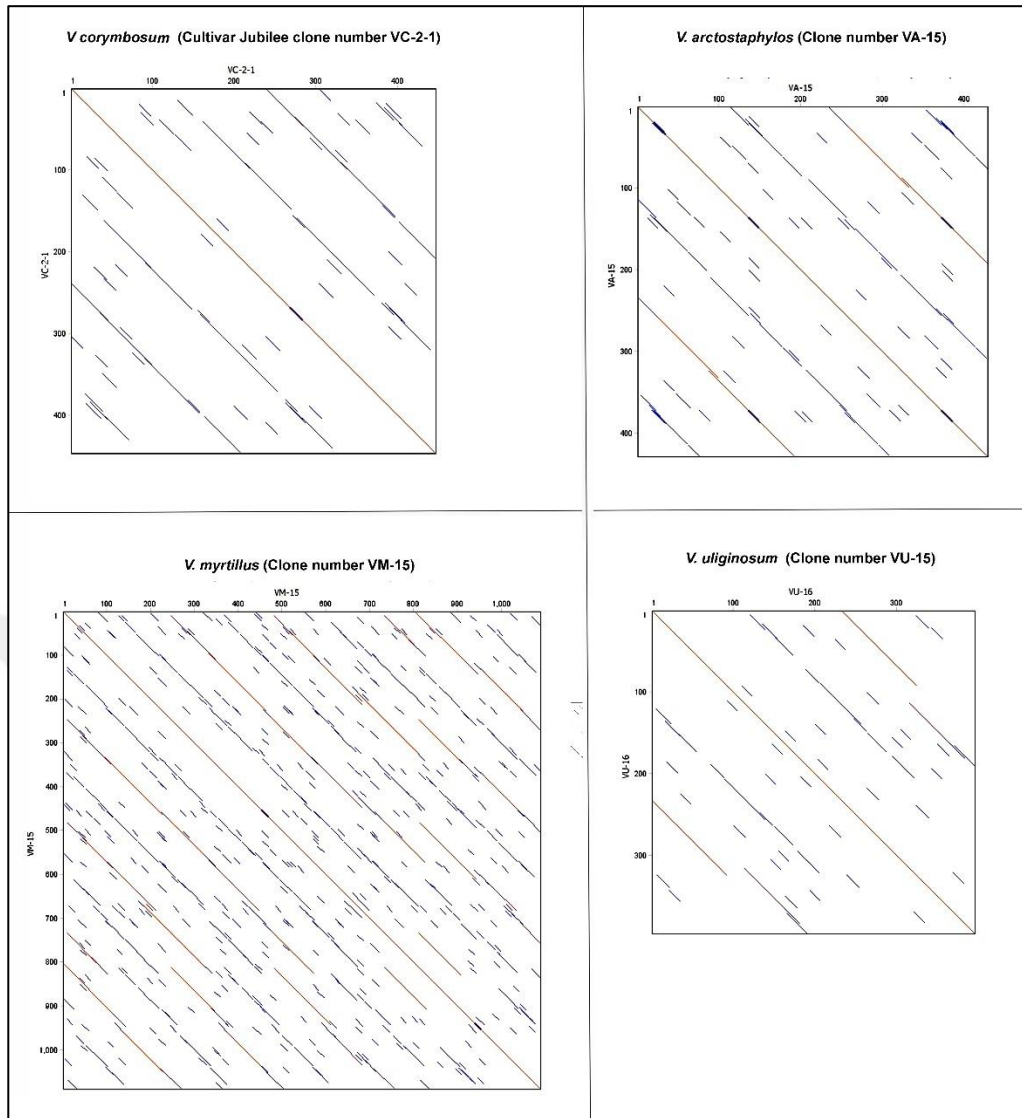
was ranged from 65-97% and hence VaccSat7 was one of most diversified satellite repeats in *Vaccinium* species (Table 5.5.).

Heatmap analysis from the pairwise similarity distance matrix revealed that this satellite repeats are highly heterogeneous intra and inter-species and no species specific variants were present (Figure 5.11f).



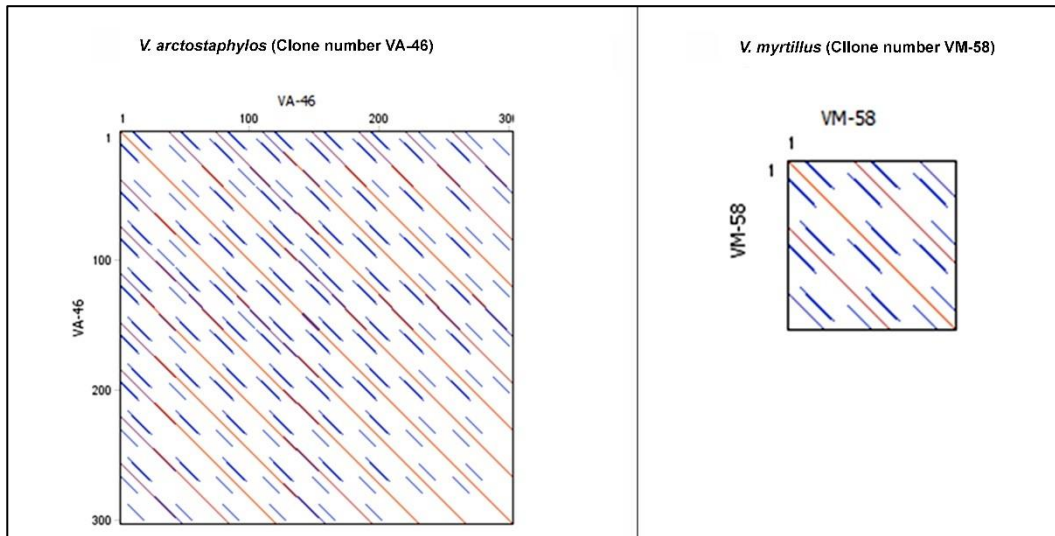
a

Figure 5.7. Dotplot analysis of multimer of *Vaccinium* satellite repeat clone. VaccSat1 (a), VaccSat2 (b), VaccSat5 (c) and VaccSat6 (d). Monomer unit and subunit are showed as parallel line on the dotplot diagram (For clone number Table 5.3. referered)

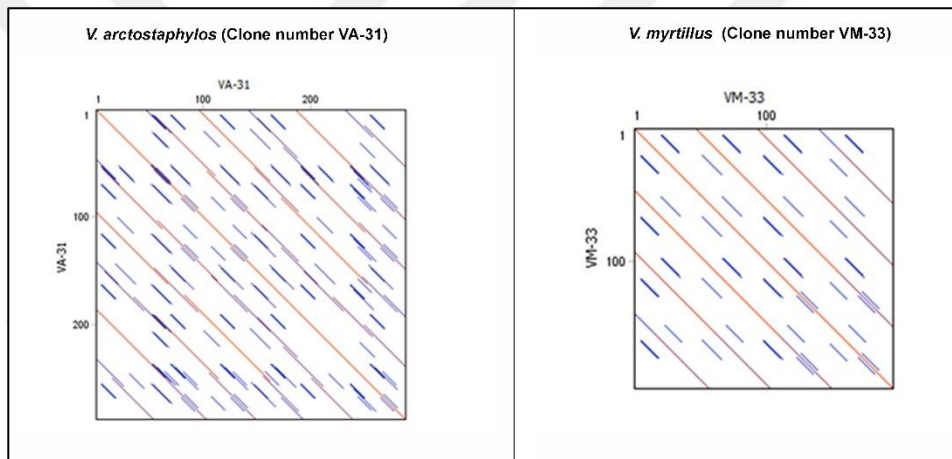


b

Figure 5.7. (Continued) Dotplot analysis of multimer of *Vaccinium* satellite repeat clone. VaccSat1 (a), VaccSat2 (b), VaccSat5 (c) and VaccSat6 (d). Monomer unit and subunit are showed as parallel line on the dotplot diagram (For clone number Table 5.3. referered)

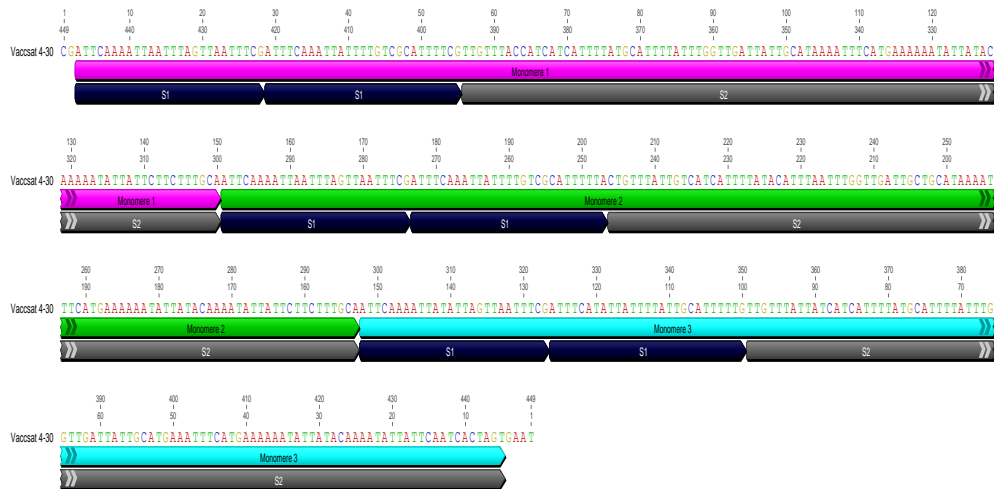


c

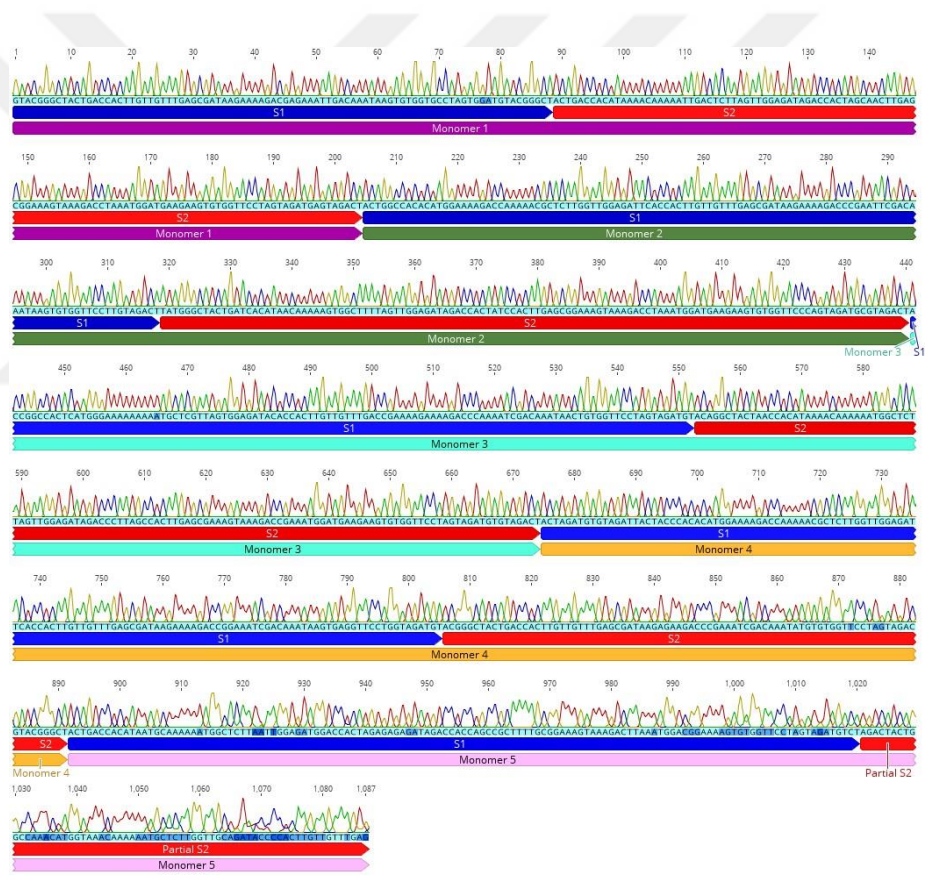


d

Figure 5.7. (Continued) Dotplot analysis of multimer of *Vaccinium* satellite repeat clone. VaccSat1 (a), VaccSat2 (b), VaccSat5 (c) and VaccSat6 (d). Monomer unit and subunit are showed as parallel line on the dotplot diagram (For clone number Table 5.3. referred).

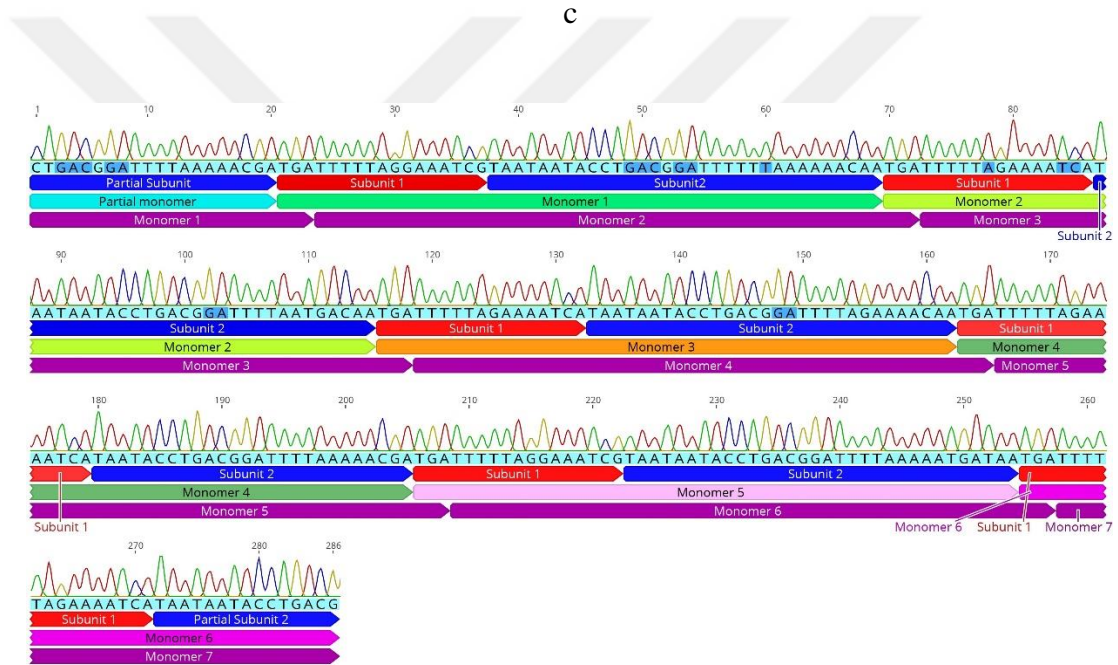
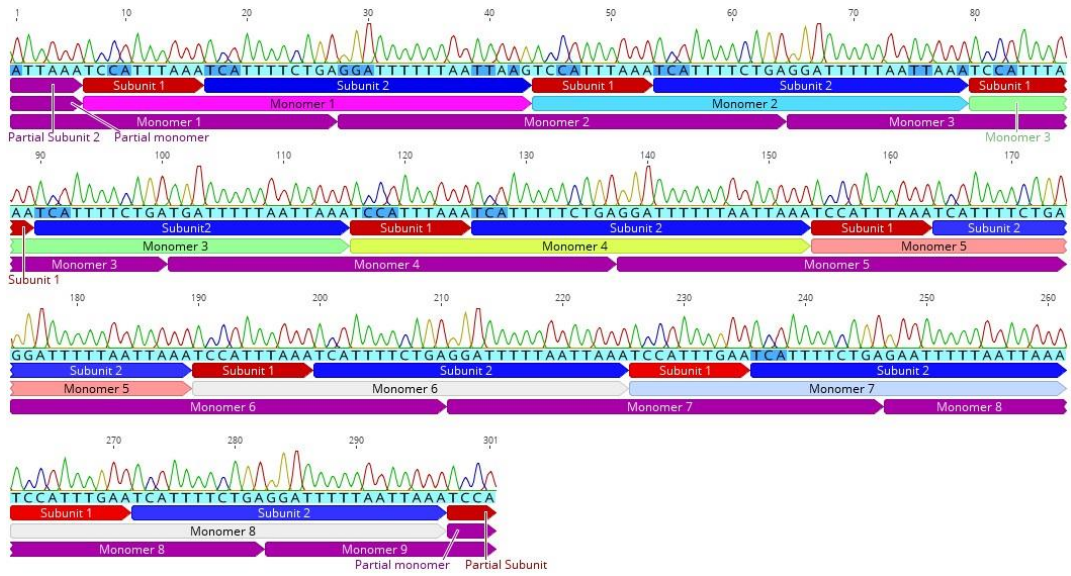


a



b

Figure 5.8. Subunit and higher order repeat unit structure of multimer clone *Vaccinium* satellite. VaccSat1 clone VC-3030 (a), VaccSat2 clone VM-15 (b), VaccSat5 clone VA-46 (c) VaccSat6 clone VA-31 (d). (For clone number Table 5.3. referred)

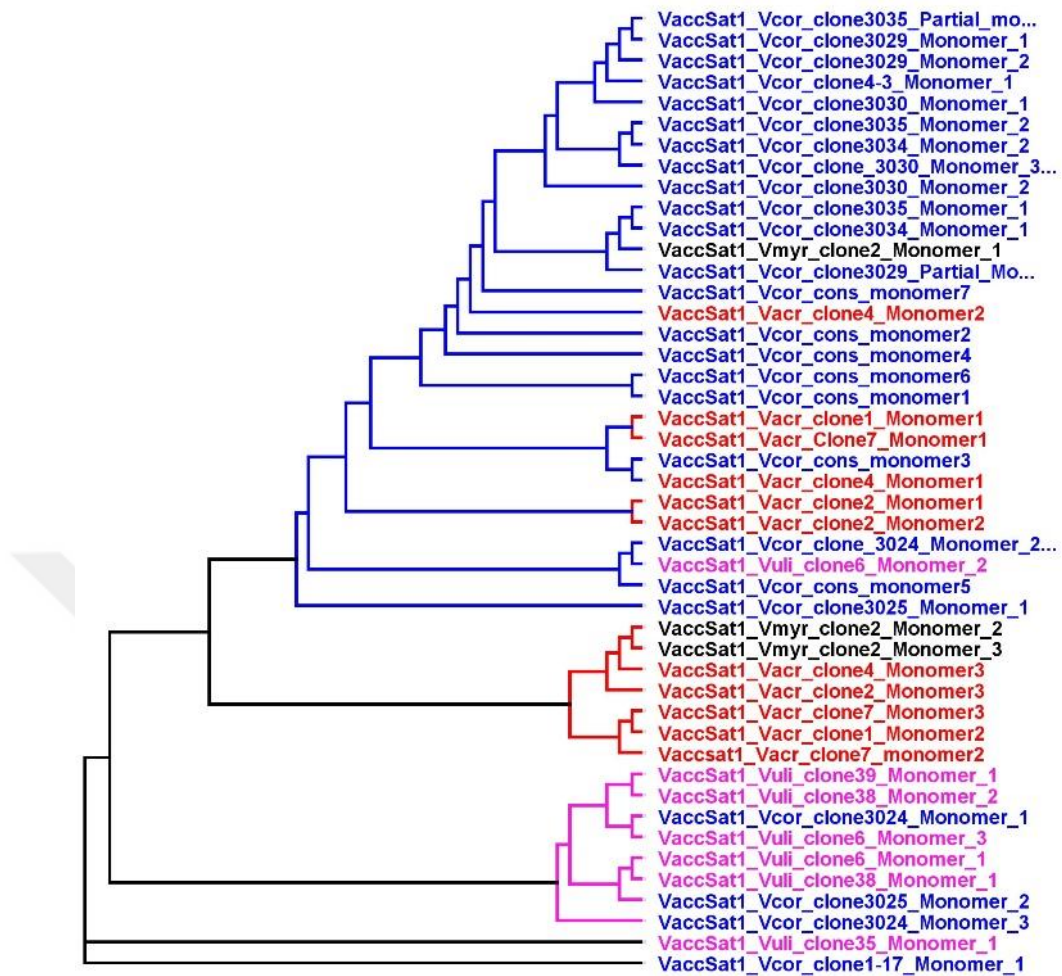


d

Figure 5.8. (Continued) Subunit and higher order repeat unit structure of multimer clone *Vaccinium* satellite. VaccSat1 clone VC-3030 (a), VaccSat2 clone VM-15 (b), VaccSat5 clone VA-46 (c) VaccSat6 clone VA-31 (d). (For clone number Table 5.3. referred)

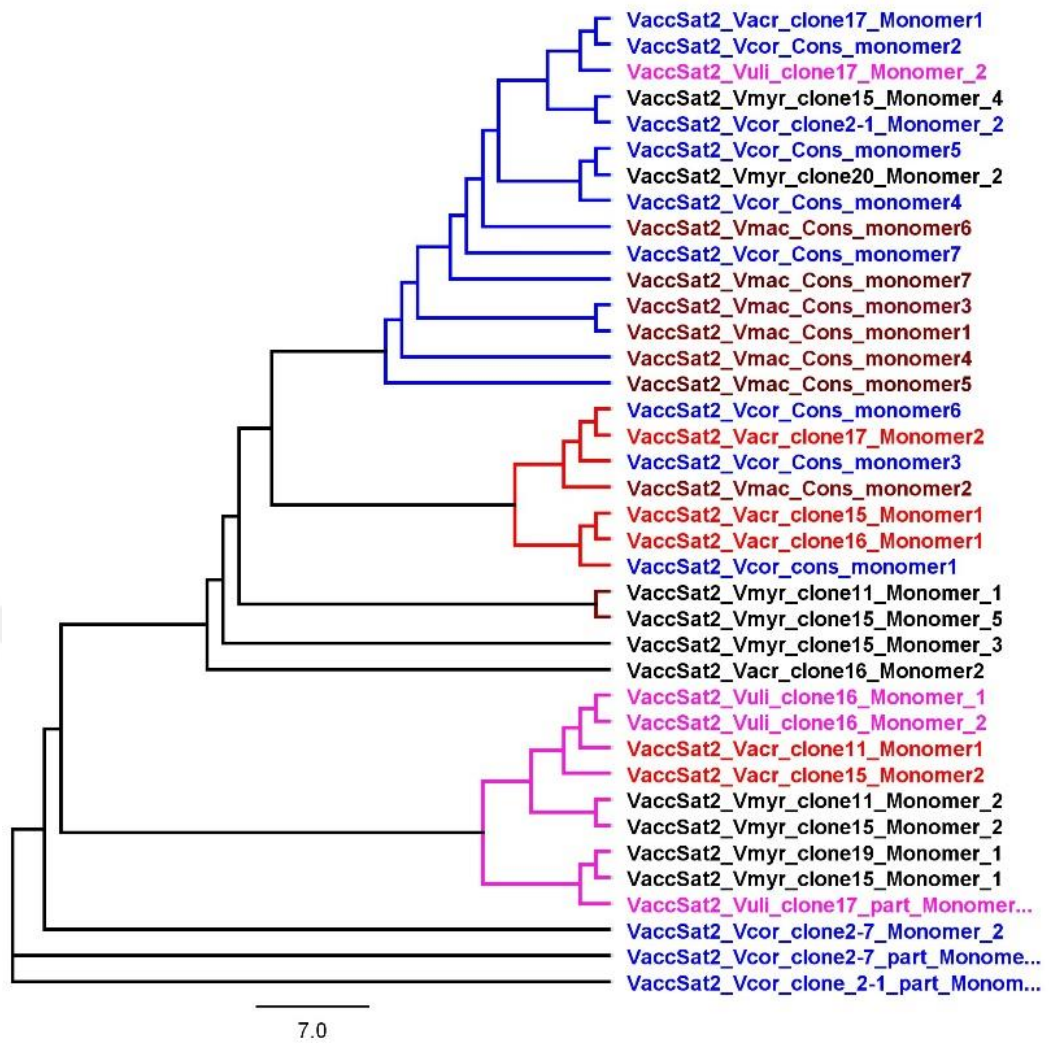
Table 5.4. Summary of satellite repeat structure in the genome of *Vaccinium*.

Name of the Satellite	Range of monomer size	Satellite subunit organization	
VaccSat1	146-148		
		Presence of subunit	Yes
		Total subunit number	Two
		Length of the subunit	S1 = 28 bp S2 = 94 bp
		Subunit organization in full length monomer	S1+S1+S2
		HOR (higher order subunit structure of satellite monomer)	Yes (Figure 5.8a)
VaccSat2	238-240		
		Presence of subunit	Yes
		Total subunit number	One
		Length of the subunit	S1 = 122 bp
		Subunit organization in full length monomer	S1+S1
		HOR (higher order subunit structure of satellite monomer)	No (Figure 5.8b)
VaccSat3	153-154		
		Presence of subunit	No
VaccSat5	36-38		
		Presence of subunit	Yes
		Total subunit number	Two
		Length of the subunit	S1 = 10 bp S2 = 26 bp
		Subunit organization in full length monomer	S1+S2
		HOR (higher order subunit structure of satellite monomer)	Yes (Figure 5.8c)
VaccSat6	49-52		
		Presence of subunit	Yes
		Total subunit number	Two
		Length of the subunit	S1 = 17 bp S2 = 32 bp
		Subunit organization in full length monomer	S1+S2
		HOR (higher order subunit structure of satellite monomer)	Yes (Figure 5.8d)
VaccSat7	49-71	Presence of subunit	No



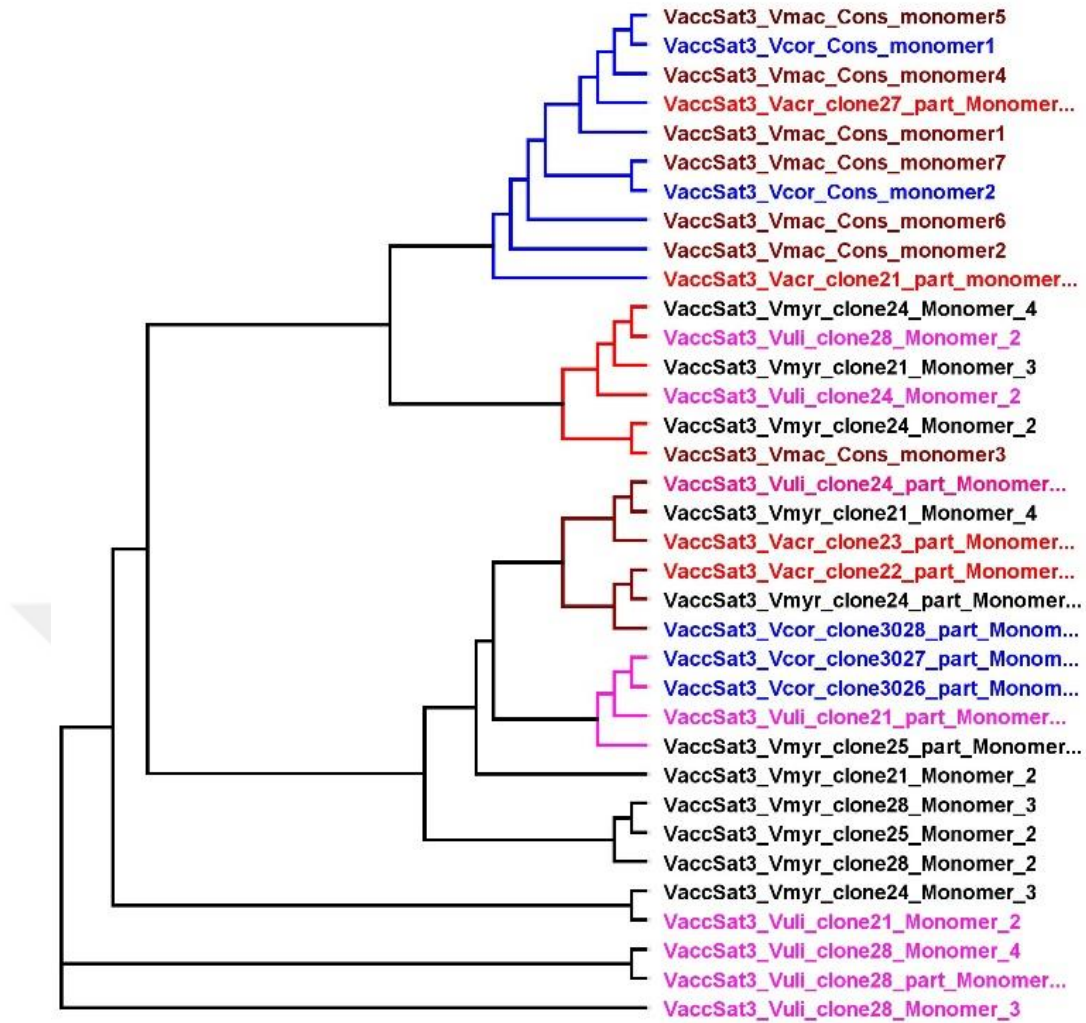
a

Figure 5.9. Phylogenetic dendrogram of extracted monomer sequences from 6 different *Vaccinium* satellite using FastTree 2.1.5 (Price et al., 2010). VaccSat1-VaccSat7 were depicted in (a)-(f). (Referred Table 5.3. for monomer name) Major clusters are colored and minor clusters are given in black.



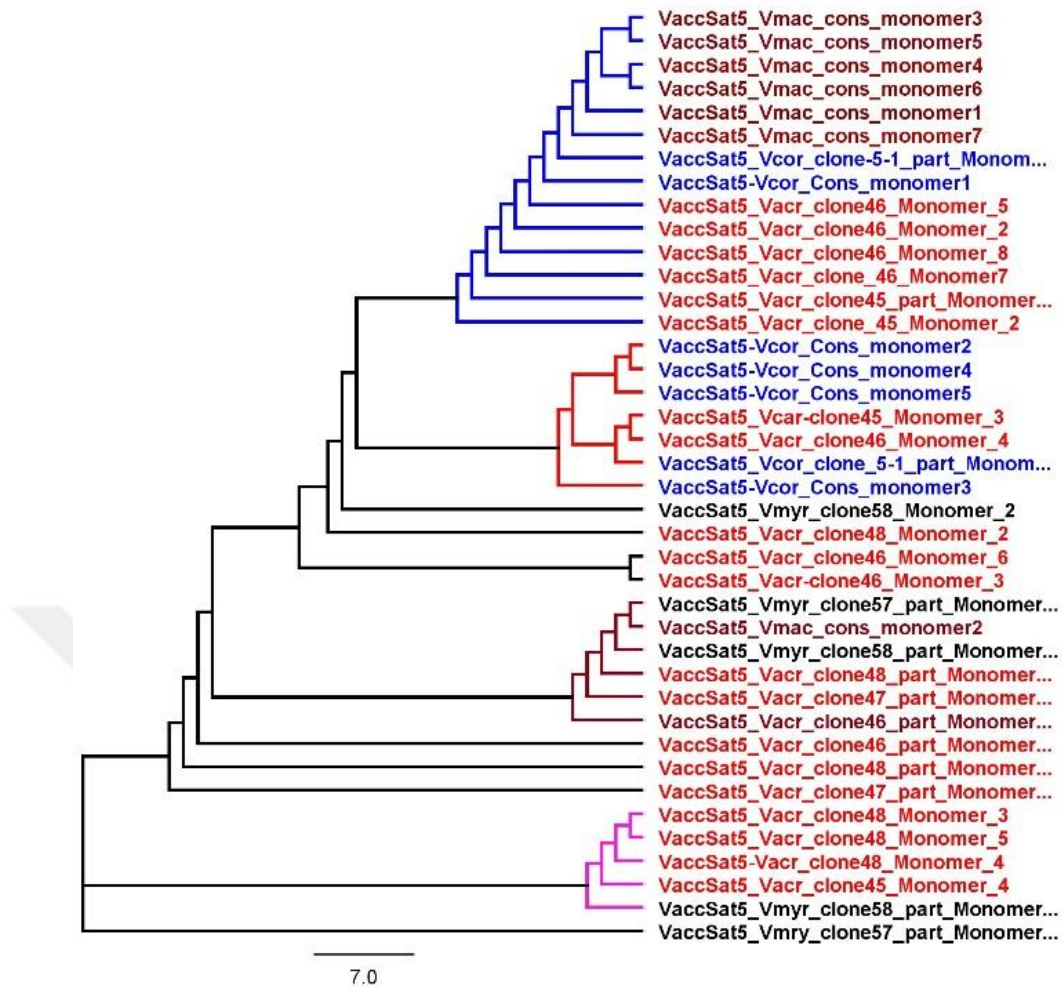
b

Figure 5.9. (Continued) Phylogenetic dendrogram of extracted monomer sequences from 6 different *Vaccinium* satellite using FastTree 2.1.5 (Price et al., 2010). VaccSat1-VaccSat7 were depicted in (a)-(f). (Referred Table 5.3. for monomer name) Major clusters are colored and minor clusters are given in black.



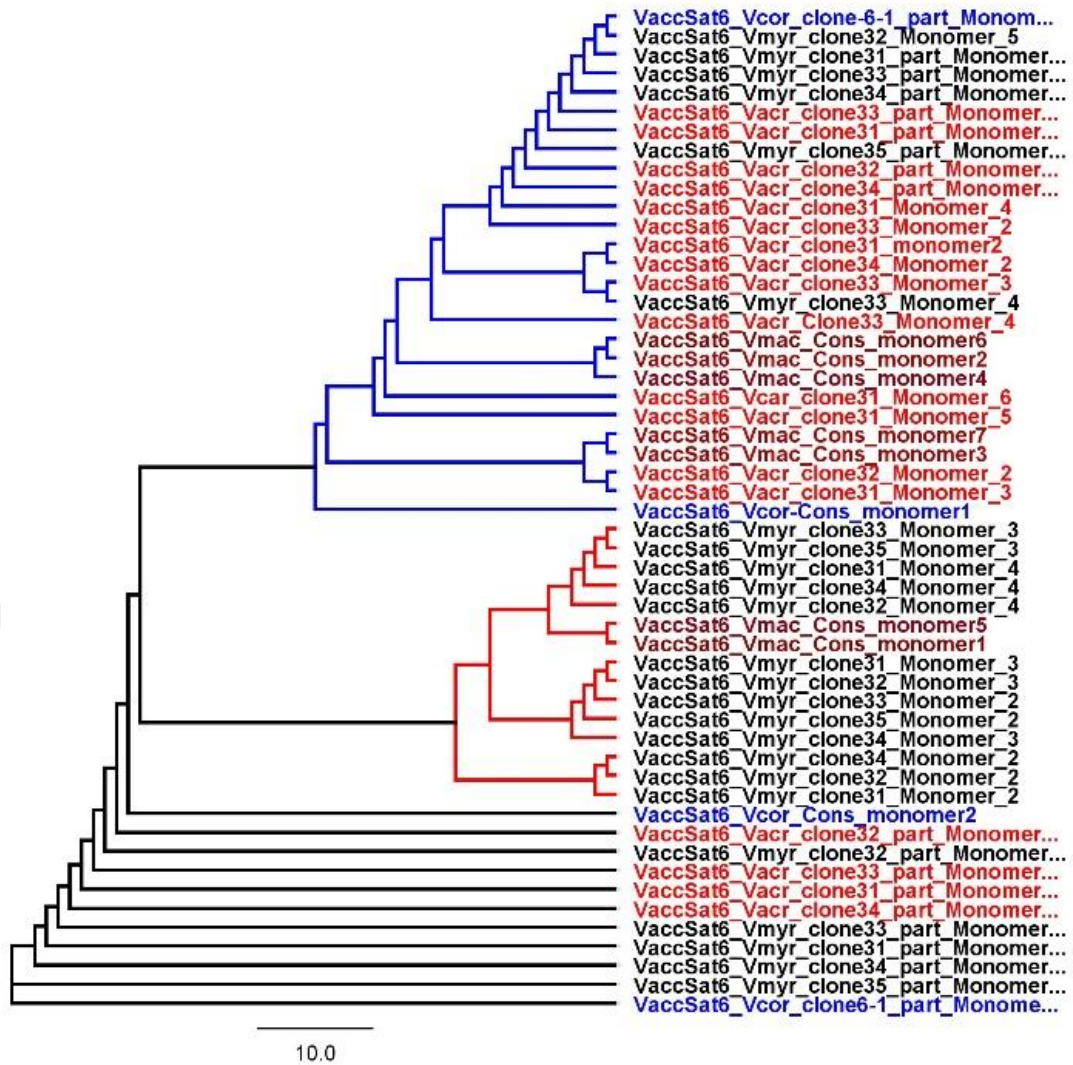
c

Figure 5.9. (Continued) Phylogenetic dendrogram of extracted monomer sequences from 6 different *Vaccinium* satellite using FastTree 2.1.5 (Price et al., 2010). VaccSat1- VaccSat7 were depicted in (a)-(f). (Referred Table 5.3. for monomer name) Major clusters are colored and minor clusters are given in black.



d

Figure 5.9. (Continued) Phylogenetic dendrogram of extracted monomer sequences from 6 different *Vaccinium* satellite using FastTree 2.1.5 (Price et al., 2010). VaccSat1- VaccSat7 were depicted in (a)-(f). (Referred Table 5.3. for monomer name) Major clusters are colored and minor clusters are given in black.



e

Figure 5.9. (Continued) Phylogenetic dendrogram of extracted monomer sequences from 6 different *Vaccinium* satellite using FastTree 2.1.5 (Price et al., 2010). VaccSat1-VaccSat7 were depicted in (a)-(f). (Referred Table 5.3. for monomer name) Major clusters are colored and minor clusters are given in black.

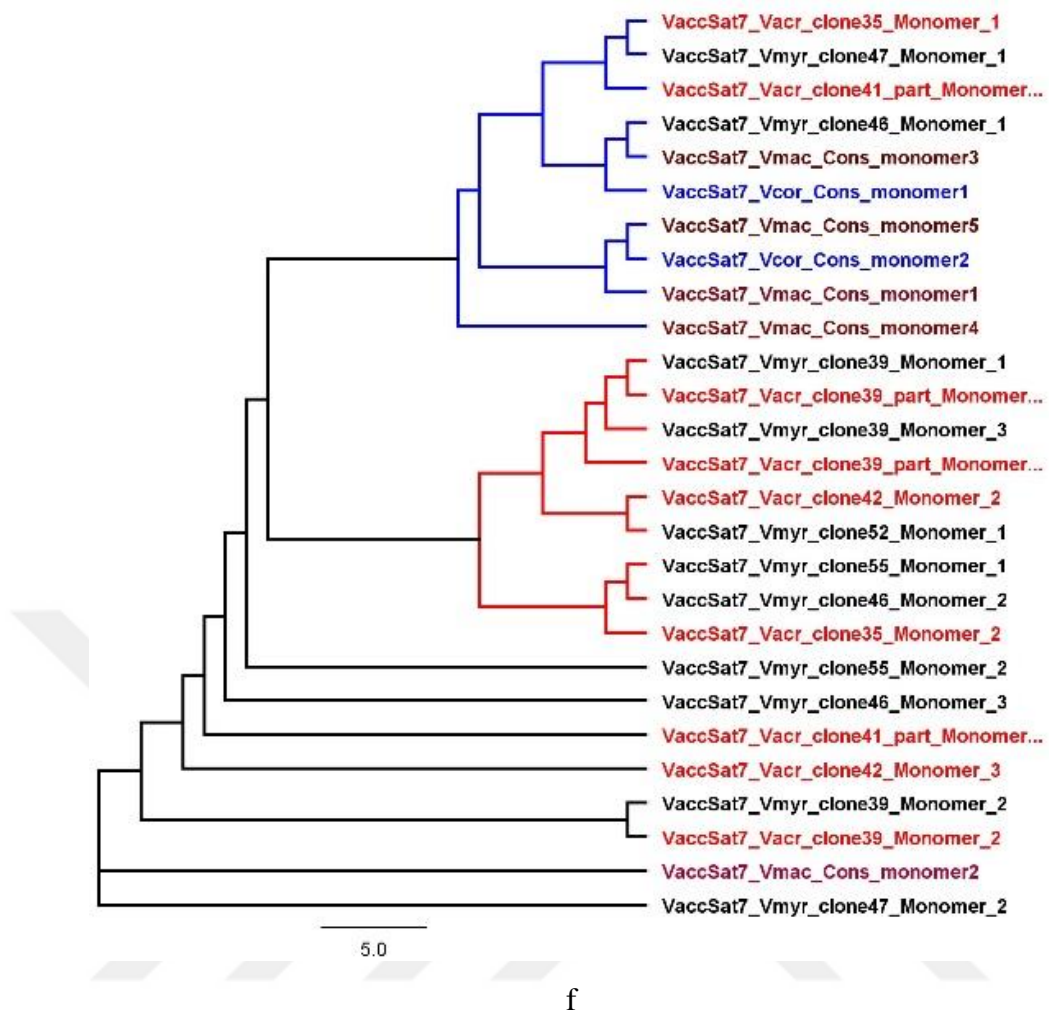
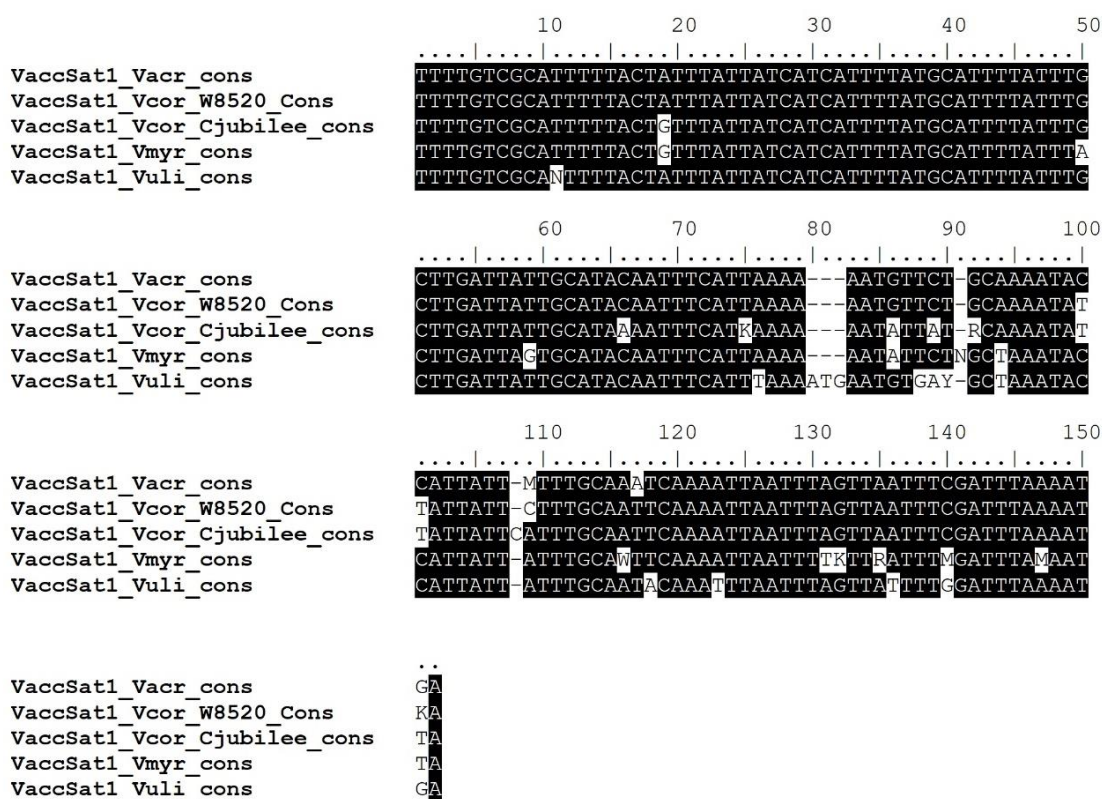
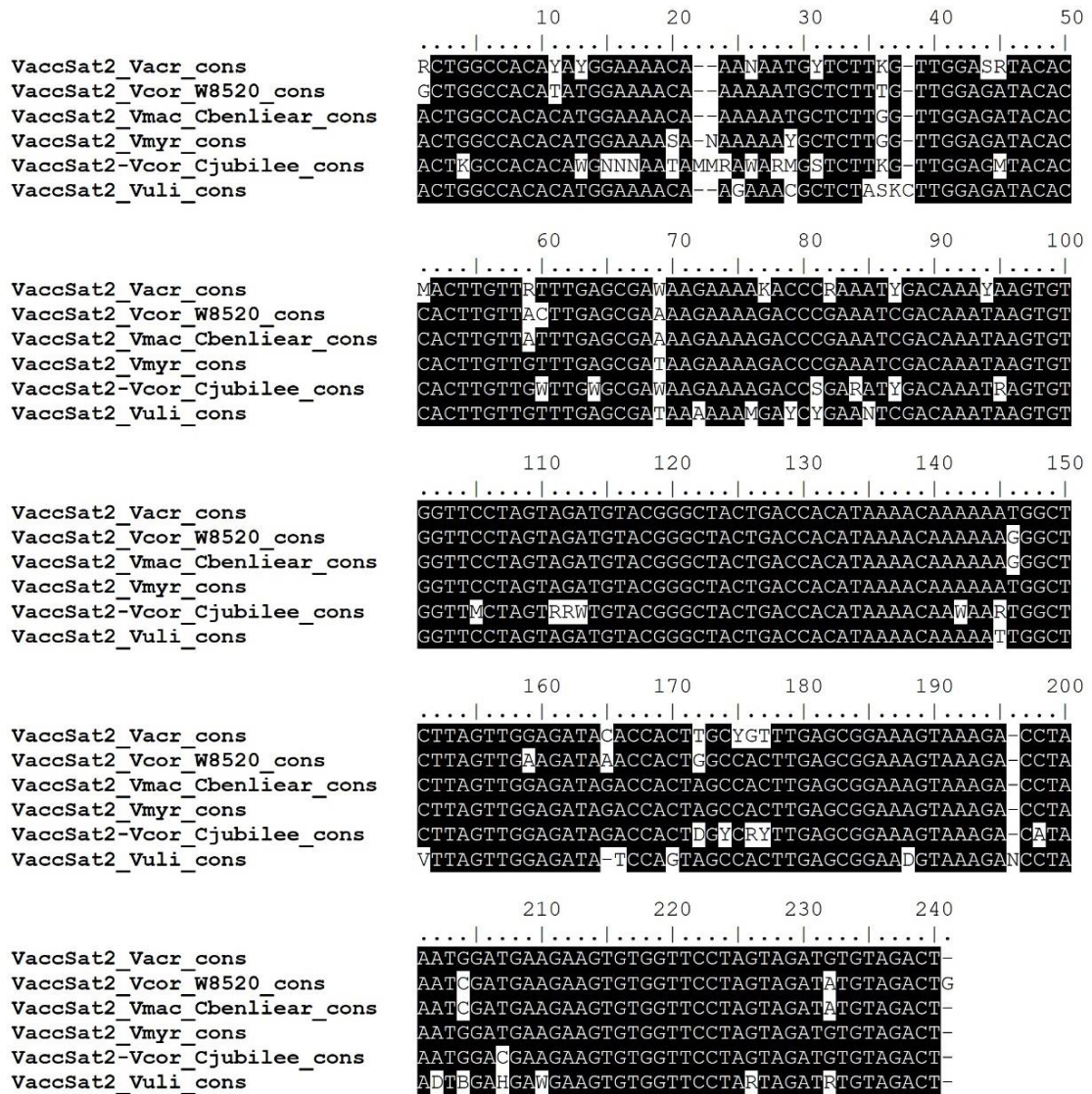


Figure 5.9. (Continued) Phylogenetic dendrogram of extracted monomer sequences from 6 different *Vaccinium* satellite using FastTree 2.1.5 (Price et al., 2010). VaccSat1- VaccSat7 were depicted in (a)-(f). (Referred Table 5.3. for monomer name) Major clusters are colored and minor clusters are given in black.



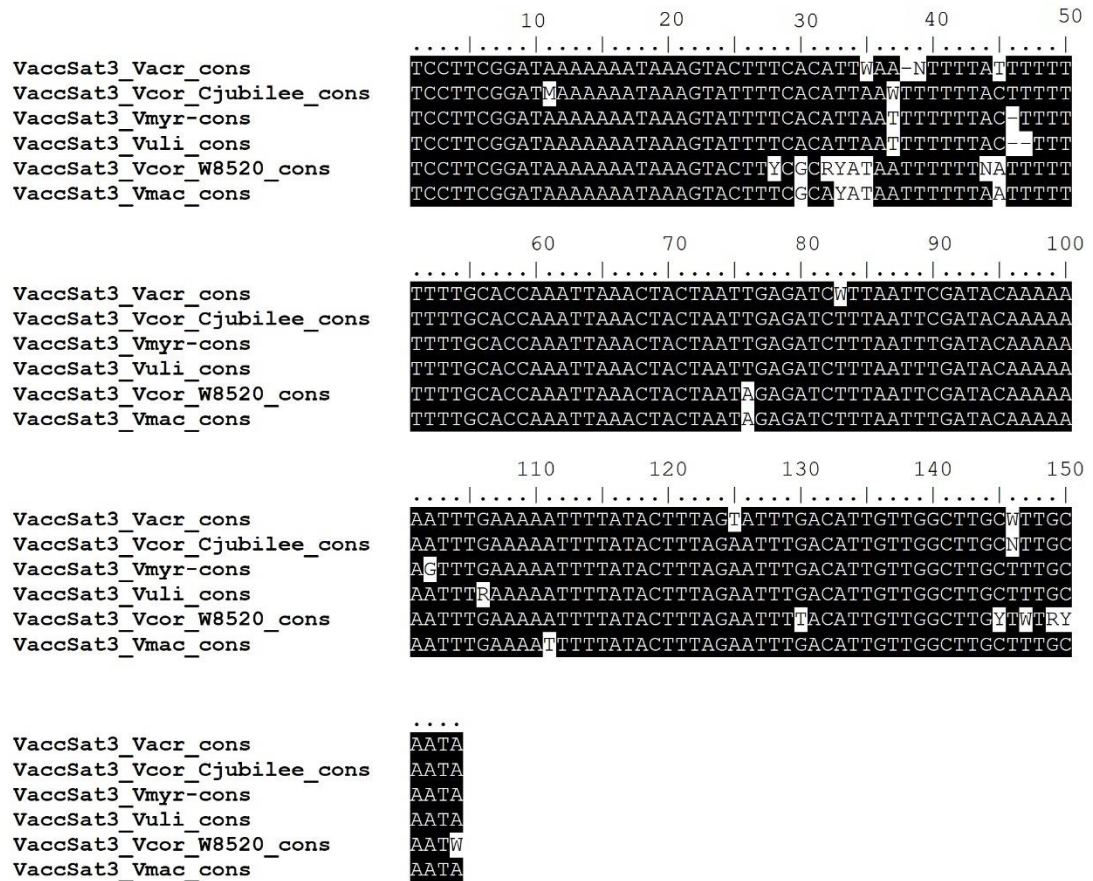
a

Figure 5.10. Multiple sequence alignment of consensus monomer sequence using MAFT multiple sequence alignment tool from six *Vaccinium* satellite. VaccSat1-VaccSat7 are depicted in (a)-(f). (For consensus monomer sequence name, Table 5.3 is referred). Homogeneous region of the multiple alignments is shaded in black and polymorphic site is given with no shading.



b

Figure 5.10. (Continued) Multiple sequence alignment of consensus monomer sequence using MAFT multiple sequence alignment tool from six *Vaccinium* satellite. VaccSat1-VaccSat7 are depicted in (a)-(f). (For consensus monomer sequence name, [Table 5.3](#) is referred). Homogeneous region of the multiple alignments is shaded in black and polymorphic site is given with no shading.



c

Figure 5.10. (Continued) Multiple sequence alignment of consensus monomer sequence using MAFT multiple sequence alignment tool from six *Vaccinium* satellite. VaccSat1-VaccSat7 are depicted in (a)-(f). (For consensus monomer sequence name, Table 5.3. is referred). Homogeneous region of the multiple alignments is shaded in black and polymorphic site is given with no shading.

```

                                10      20      30
VaccSat5_Vacr_cons      -GGATT-TTTAATTAAATCCATTTAAATCATTTTCTGA
VaccSat5_Vmyr_cons      -GGATT-TTTAATTAAATCCATTTAAATCATTTTCTGA
VaccSat5_Vcor_Cjubilee_cons -GGATTWTTTAATTAAATCCATTTAAATCATTTTHTGA
VaccSat5_Vcor_W8520_cons GGGATT-TTTAATTAAATCCATTTAAATCATTTT-TGA
VaccSat5_Vmac_cons      -GGATTATTTHATTTAATCCGTTTAAATCATTTTCTGA

```

d

```

                                10      20      30      40      50
VaccSat6_Vacr_cons      TTTT TAGAAAATCATAATA---ATACCTGACGGATTTTAAAAACGAYAAAT
VaccSat6_Vcor_Cjubilee_cons TTTT TAGAAAATCATAATA---ATACCTGACGGATTTTAAAAAYGAWRAT
VaccSat6_Vcor_W8520_cons  TTTT TAGAAAATCANAATA---GTRCCTGACGGAYTTTAAAAATDATDAT
VaccSat6_Vmyr_cons      TTTT TAGAAAATCGGAATA---ATACCTGACGGATTTTAAAAACAATGA-
VaccSat6_Vmac_cons      TTTT TAGAAAATNAGAATAATGATACCTGACGGATTTTAAAAACAATGA-

```

```

VaccSat6_Vacr_cons      GA
VaccSat6_Vcor_Cjubilee_cons GA
VaccSat6_Vcor_W8520_cons GA
VaccSat6_Vmyr_cons      --
VaccSat6_Vmac_cons      --

```

e

Figure 5.10. (Continued) Multiple sequence alignment of consensus monomer sequence using MAFT multiple sequence alignment tool from six *Vaccinium* satellite. VaccSat1-VaccSat7 are depicted in (a)-(f). (For consensus monomer sequence name, Table 5.3. is referred). Homogeneous region of the multiple alignments is shaded in black and polymorphic site is given with no shading.

```

          10      20      30      40      50
    .....|.....|.....|.....|.....|.....|.....|.....|.....|.....|
VaccSat7_Vacr_cons  TGGACAAGTTAGTTTTTTTGCAAAACAAAAA-----
VaccSat7_Vmyr_cons  TGGACAAGTTAGTTTTTTTGCAAAACAAAAA-----C
VaccSat7_Vmac_cons  TGGACAAGTTAGTTTTTTTGCAAAACAAAAAACC-----A
VaccSat7_Vcor_W8520_cons  TGGACMASTTAGTTTTTTTGCAAAACAAAAANNNA-----C-----A-----C-----CANNN

          60      70
    .....|.....|.....|.....|
VaccSat7_Vacr_cons  CAAAAAAGTTAGTTAG-----
VaccSat7_Vmyr_cons  CAAAAAAGTTAGTTAG-----
VaccSat7_Vmac_cons  AAAAAAAGTTAGTTAG-----
VaccSat7_Vcor_W8520_cons  AAAAAACCAAGTTAGTTAGTTGGAAC

```

f

Figure 5.10. (Continued) Multiple sequence alignment of consensus monomer sequence using MAFT multiple sequence alignment tool from six *Vaccinium* satellite. VaccSat1-VaccSat7 are depicted in (a)-(f). (For consensus monomer sequence name, Table 5.3. is referred). Homogeneous region of the multiple alignments is shaded in black and polymorphic site is given with no shading.

Table 5.5. Genetic distance among different *Vaccinium* species based on the consensus monomer sequence of VaccSat1-7. The lowest values were colored with yellow and the highest values were with cyan.

VaccSat1						
	Vacr_cons	Vcor_W8520_ Cons	Vcor_Cjubilee_ cons	Vmyr_cons	Vuli_cons	
Vacr_cons						
Vcor_W8520_ Cons	97.279					
Vcor_Cjubilee_ cons	92.905	94.932				
Vmyr_cons	91.892	91.216	90.94			
Vuli_cons	90.833	89.5	87.252	86.589		
VaccSat2						
	Vacr_cons	Vcor_W8520_ cons	Vmac_Cbenliear_ _cons	Vmyr_cons	Vcor_Cjubilee_ _cons	Vuli_cons
Vacr_cons						
Vcor_W8520_ cons	92.691					
Vmac_Cbenliear_ _cons	93.538	97.034				
Vmyr_cons	93.987	94.093	97.046			
Vcor_Cjubilee_ cons	88.48	87.64	89.531	91.422		
Vuli_cons	87.675	88.831	90.931	91.806	85.104	
VaccSat3						
	Vacr_cons	Vcor_Cjubilee_ cons	Vmyr_cons	Vuli_cons	Vcor_W8520_ cons	Vmac_cons
Vacr_cons						
Vcor_Cjubilee_ cons	95.13					
Vmyr-cons	93.344	96.916				
Vuli_cons	93.019	96.591	98.366			
Vcor_W8520_ cons	90.909	91.234	89.773	89.448		
Vmac_cons	93.019	93.344	93.182	92.857	94.968	

Table 5.5. (Continued) Genetic distance among different *Vaccinium* species based on the consensus monomer sequence of VaccSat1-7. The lowest values were colored with yellow and the highest values were with cyan

VaccSat5					
	Vacr_cons	Vmyr_cons	Vcor_Cjubilee_cons	Vcor_W8520_cons	Vmac_cons
Vacr_cons					
Vmyr_cons	100				
Vcor_Cjubilee_cons	95.495	95.495			
Vcor_W8520_cons	95.833	95.833	93.243		
Vmac_cons	90.09	90.09	89.64	86.036	
VaccSat6					
	Vacr_cons	Vcor_Cjubilee_cons	Vcor_W8520_cons	Vmyr_cons	Vmac_cons
Vacr_cons					
Vcor_Cjubilee_cons	96.429				
Vcor_W8520_cons	88.605	89.626			
Vmyr_cons	90.217	90.217	86.775		
Vmac_cons	85.204	85.204	81.973	90.306	
VaccSat7					
	Vacr_cons	Vmyr_cons	Vmac_cons	Vcor_W8520_cons	
Vacr_cons					
Vmyr_cons	97.959				
Vmac_cons	75.41	75.41			
Vcor_W8520_cons	65	65.385	80.769		

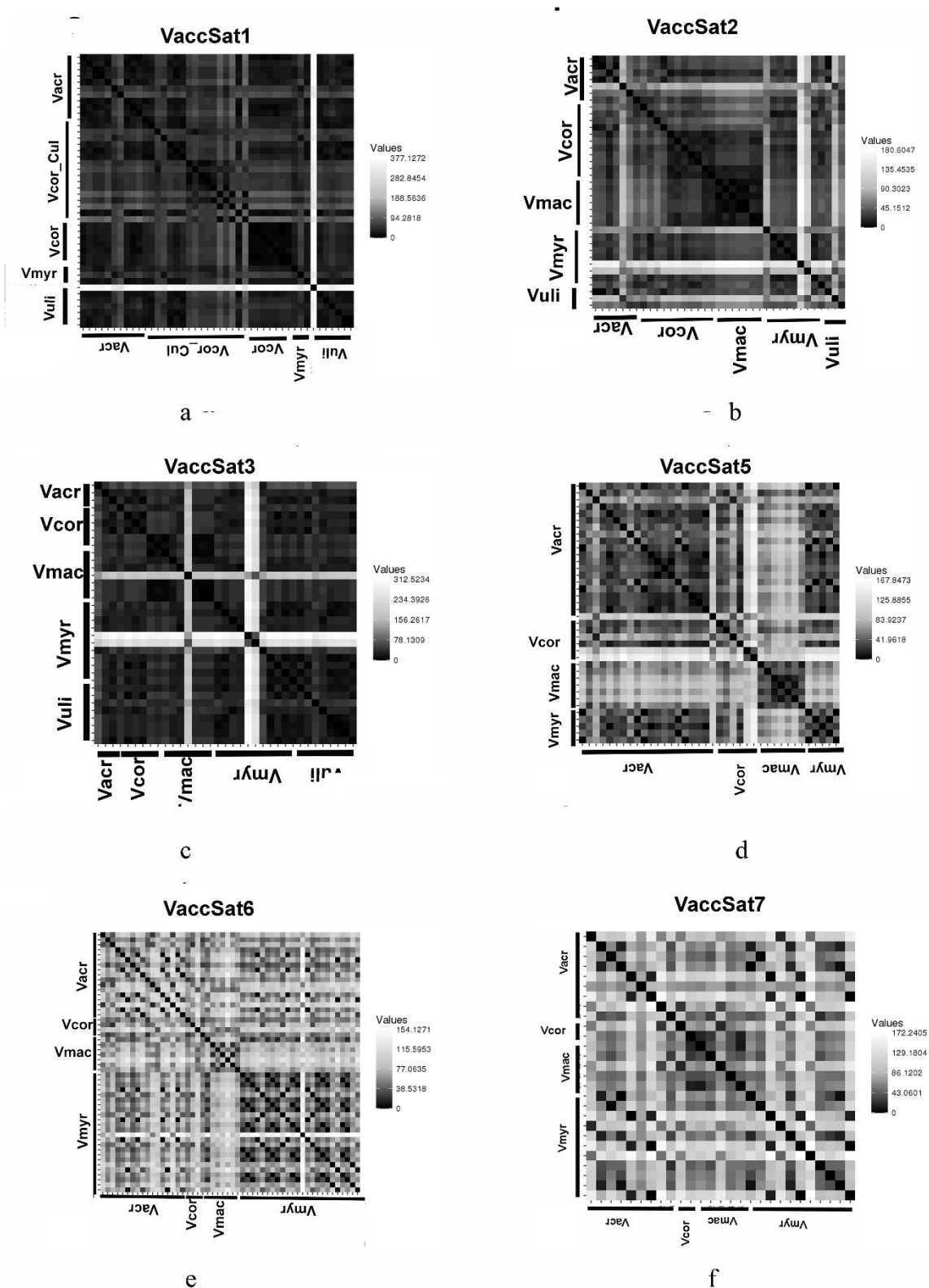


Figure 5.11. Heatmap analysis of monomer unit of six *Vaccinium* satellite. (Pairwise similarity distance heatmap of extracted monomer unit. VaccSat1-VaccSat7 are depicted in (a)-(f) (Scale bar represent the calculated distance matrix value))

Table 5.6. Overview of satellite repeat diversity in *Vaccinium* genome.

Satellite	Species specificity	Average similarity percentage among the species and total clusters from phylogenetic dendrogram	Satellite tandemly organized in the species	Satellite absent/ less abundant or highly diversified in the species
VaccSat1	Yes	87-97% (3)	<i>V. corymbosum</i> -strain W8520; <i>V. arctostaphylos</i> ; <i>V. corymbosum</i> -cultivar 'Jubilee'; <i>V. myrtillus</i> ; <i>V. uliginosum</i>	<i>V. macrocarpon</i> -cultivar 'Ben Lear'
VaccSat2	No	85-97% (4)	<i>V. corymbosum</i> -strain W8520; <i>V. arctostaphylos</i> ; <i>V. corymbosum</i> -cultivar 'Jubilee'; <i>V. myrtillus</i> ; <i>V. uliginosum</i> ; <i>V. macrocarpon</i> -cultivar 'Ben Lear'	-
VaccSat3	No	89-98% (7)	<i>V. corymbosum</i> -strain W8520; <i>V. arctostaphylos</i> ; <i>V. corymbosum</i> -cultivar 'Jubilee'; <i>V. myrtillus</i> ; <i>V. uliginosum</i> ; <i>V. macrocarpon</i> -cultivar 'Ben Lear'	-
VaccSat5	Yes	86-100% (5)	<i>V. corymbosum</i> -strain W8520; <i>V. arctostaphylos</i> ; <i>V. myrtillus</i> ; <i>V. macrocarpon</i> -cultivar 'Ben Lear'	<i>V. corymbosum</i> -Cultivar 'Jubilee'; <i>V. uliginosum</i>
VaccSat6	Yes	82-96% (2)	<i>V. corymbosum</i> -strain W8520; <i>V. arctostaphylos</i> ; <i>V. myrtillus</i> ; <i>V. macrocarpon</i> -cultivar 'Ben Lear'	<i>V. corymbosum</i> -Cultivar 'Jubilee'; <i>V. uliginosum</i>
VaccSat7	Yes	65-98% (2)	<i>V. corymbosum</i> -strain W8520; <i>V. arctostaphylos</i> ; <i>V. myrtillus</i> ; <i>V. macrocarpon</i> -cultivar 'Ben Lear'	<i>V. corymbosum</i> -Cultivar 'Jubilee'; <i>V. uliginosum</i>

CHAPTER VI

OVERVIEW OF MAJOR LTR-RETROTRANSPOSON LANDSCAPE IN *V. MACROCARPON* GENOME

6.1 Background Information

Recent technological explosion makes it possible to understand the genome dynamics of organisms in a fine scale level. It is now well understood that the plant genome evolution is the result of repeated cyclical whole genome duplication and dynamic interaction of different types of transposable elements (Wendel et al., 2016). It is also well-established that these transposable elements are the ancient elements possibly exist before the emergence of the eukaryotic organism and have the huge impact on the eukaryotic genome structure and evolution. It is also hypothesized that all transposable elements originated from the same ancestral element and diversified through differential accumulation and organization conserved module or protein domain sequence (Capy and Maisonhaute, 2002; Oliver et al., 2013).

Two major types of transposable elements are DNA transposons (class I) and retrotransposons (class II) (Llorens et al., 2010). Angiospermic plant genome is mainly enriched with LTR retrotransposons, a group of retrotransposons having two directly repeated long terminal repeats at two opposite ends of main protein coding domain *Gag/coat* and *Pol* (Oliver et al., 2013). Although *Gag* is considered as a single gene, *Pol* domain is composed of several gene like protease, integrase, RNase H and reverse transcriptase where each gene has specific function. For instance, integrase recognizes the binding site; RNase H hydrolyses the RNA and DNA duplex; protease recognizes the protein and performs degradation; reverse transcriptase produces the protein responsible for the conversion from RNA to DNA. In addition to protein domain sequences two extra regions are also found in full length elements called primer binding site (PBS) and polypurine tract (PPT). PBS is complementary to tRNA sequences and used as a primer for reverse transcriptase protein to synthesize complementary DNA minus strand. On the other hand, PPT is responsible for the plus strand synthesis (Llorens et al., 2010). *Ty3/gypsy* and *Ty1/copia* are two well-annotated superfamilies of the order LTR retrotransposons. The superfamily *Ty3/gypsy* and *Ty1/copia* have been divided based on

their order of protein domains and sequence similarity. In angiosperms, Ty3/*gypsy* is subdivided into Errantiviruses/Athila, Chromoviruses (with sublineages Del/Tekay, Reina, Galadriel and CRM), Ogre/Tat (with sublineages Ogre and Tat) and Ty1/*copia* is subdivided into Sireviruses/Maximus, Tork (with sublineages TORK, TAR and Angela), Retrofit (with sublineages Ale and Alesia), Oryco/Ivana, Bianca (Marín and Lloréns, 2000; Benabdelmouna and Darmency, 2003; Eickbush and Jamburuthugoda, 2008).

Here, the identification and characterization of Ty3/*gypsy* and Ty1/*copia* retrotransposons is discussed in details in the genome of *V. macrocarpon* Ait.

6.2 Materials and Methods

Construction of full length LTR retrotransposons

To reconstruct the full-length LTR retrotransposon sequences several steps were followed. First, supercluster and clusters belonging to particular types of retrotransposon were identified based on their shape and protein domain hits. Second, contigs from the identified clusters were imported in Geneious (<http://www.geneious.com>, Kearse et al., 2012). Third, mapping all contig sequences against the longest contig sequence as a reference with Geneious mapper tool with fine tuning interaction parameter 10, and finally extracting the consensus sequence from the mapped reads. If the identified LTR retrotransposon sequences belonged to only a single cluster than the consensus sequences were directly used as a reconstructed full length element. However, if the LTR retrotransposons do not belong to single cluster than some additional steps were taken under consideration for reconstruction of full-length elements.

The additional steps were constituted with progressive multiple sequence alignment of the consensus sequences of all clusters, using Geneious alignment tools. The parameter for Geneious alignment tools were: Automatically determine direction, Alignment type = Global alignment, cost matrix = 65% similarity (5.0/-4.0), Gap open penalty = 12, Gap extension penalty = 3, Refinement iterations = 2). Finally, exporting the final consensus reconstructed full length elements.

The reconstructed full-length sequences were searched for respective structural features of LTR retrotransposon like protein domain sequences, LTR region, primer binding region sites (PBS) and Polypurine Tract sequences (PPT). Different software and databases were used for these purposes. Software and web-based computational tools used in these analyses are listed below.

1. LTR_Finder (Xu et al., 2007)
2. Dot plot (<http://www.geneious.com>, Kearse et al., 2012)
3. Protein domain tools implemented in RepeatExplorer (Novák et al., 2013)
4. Geneious (<http://www.geneious.com>, Kearse et al., 2012)

Databases used to characterize each full-length sequences are

1. RepBase (Bao et al., 2015)
2. The Gypsy Database (GyDB) (Llorens et al., 2010)

Sequence divergence among the sublineages of both Ty1/*copia* and Ty3/*gypsy* elements were analyzed through pairwise sequence similarity analysis of the identified LTR and RT elements using MAFFT plugin tools in Geneious (Kato and Standley, 2013). The parameters for MAFFT multiple sequence alignment tool was: Algorithm = Auto (selects an appropriate strategy from L-INS-I, FFT-NS-i and FFT-NS-2 according to data size), scoring matrix = 200PAM/K = 2, Gap open penalty = 1.53, offsetvalue = 0.123.

6.3 Results

Structural diversity of LTR retroelements

A total of 26 and 28 full-length elements were identified and characterized from Ty1/*copia* and Ty3/*gypsy* retrotransposon elements, respectively. Important features of identified full-length elements are shown in the (Table 6.1.-6.7. and Figure 6.1.-6.2.).

Ty1/*copia*

Table 6.1. indicate that Ale/Retrofit family had a total of 7 full-length elements from sublineages Ale and Alesia, with genome proportion for each individual sequences ranging from 0.01-0.5%. The functional *Gag* and *Pol* protein domain regions were identified and depicted. Even though intact primer binding site (PBS) sequence was found for only two full-length elements, intact polypurine tract (PPT) region was found for six full length elements. Average sequence length and LTR length varied between 4.9-5.5 kb and 127 bp-562 bp, respectively (Table 6.1., Figure 6.1 and Figure 6.1). Pairwise sequence similarity of LTR and RT ranged between 17-54% and 58-72%, respectively. Therefore, Ale/Retrofit was one of the diversified families of Ty1/*copia* superfamily (Table 6.2. and Table 6.3.).

Family Ivana/Oryco has total three full-length elements with genome proportion ranged between 0.02-0.19%, length 4.9-5.5%, LTR length of 417-438 bp (Table 6.1.). Internal protein domain sequences (*Gag* and *Pol* gene) and intact PPT sequence were identified for all the three elements but intact PBS was only found for one element (Table 6.1. and Figure 6.1). Pairwise sequence similarity for LTRs region ranged between 41.45-47.37% and RT element ranged between 62-69% (Table 6.2. and Table 6.3.).

Family Sireviruses/Maximus had only one full-length element, with genome proportion 0.19%, full length 7 kb and LTR length 1,545 bp (Table 6.1 and Figure 6.1). Therefore, Sireviruses was one of the longest Ty1/*copia* elements whose LTR was also the longest among all other Ty1/*copia* elements. Even though internal protein domain sequences were identified and depicted, no intact PBS or PPT regions were identified.

Family Bianca had only one element with genome proportion of 0.12%, total size of 7 kb and LTR length ranged between 313-326 bp (Table 6.1). Characteristics *Gag* and *Pol* protein domain sequences were identified and characterized. Although intact PBS region was not identified, 16 bp PPT region was found and depicted (Table 6.1. and Figure 6.1). Therefore, Bianca was one of the less diversified families of Ty1/*copia* type LTR retrotransposons of *V. macrocarpon*.

The highest number of full-length elements belonged to the family Tork (total 14 sequences) in the sublineages TAR, Angela and Tork. Genome proportion of the full elements varied between 0.06-0.33%. Size of the full length elements and length of LTR ranged between 4.8-6.8kb and 329-994bp, respectively (Table 6.1.). Internal protein coding domain sequences (*Gag* and *Pol* gene) were found for all full-length elements. Most of the Tork elements had intact PBS (tRNA¹⁶²-MetCAT-type) and PPT regions (Table 6.1., Figure 6.1 and Figure 6.1). Pairwise sequence similarity for LTRs and RT region were ranged between 17-43% and 57-74%, respectively (Table 6.2. and Table 6.3.). Therefore, family Tork was one of the diversified Ty1/*copia* type LTR-retrotransposons group in *V. macrocarpon*.

The common primer binding site belonged to MetCAT group with some variation in most of the Ty1/*copia* element.

Table 6.1. Cluster, supercluster, genome proportion, structural protein domain, PBS and PPT features identified in Class_I|LTR|Ty1/*copia* retrotransposon in *V. macrocarpon* (GAG = *Gag* domain, PROT = Protease, RT = Reverse transcriptase domain, RH = RNase H domain, INT = Integrase domain)

SL	Supercluster	Cluster	Genome proportion (%)	Short name	length (bp)	Length of LTR (bp)	Internal protein domain
1	5	6	0.49	SCL5_Ale	4251	127	PROT, INT, RT, RH
2	22	28, 30	0.57	SCL22_Ale	4880	183-189	GAG, PROT, INT, RT RH
3	110	173	0.09	SCL110_Ale	4893	351	GAG, PROT, INT, RT RH
4	112	175	0.09	SCL112_Ale	5054	356-401	GAG, PROT, INT, RT, RH
5	288	384	0.02	SCL288_Ale	5227	233	GAG, PROT, INT, RT, RH
6	129	198	0.08	SCL129_Alesia	5,385	562	GAG, PROT, INT, RT, RH
7	344	441	0.01	SCL344_Alesia	4944	289-296	GAG, PROT, INT, RT, RH
8	35	50	0.21	SCL35_Angela	6806	796	GAG, PROT, INT, RT RH
9	79	120	0.12	SCL79_Bianca	7016	313-326	GAG, PROT, INT, RT RH
10	42	63	0.19	SCL42_Ivana	5528	307-316	GAG,PROT, INT, RT, RH
11	178	260	0.06	SCL178_Ivana	4914	417-447	GAG, PROT, INT, RT, RH
12	310	406	0.02	SCL310_Ivana	5052	438	GAG, PROT, INT, RT, RH
13	13	16	0.33	SCL13_TAR	6550	637-695	GAG, PROT, INT, RT RH
14	26	38, 293	0.30	SCL26_TAR	6471	994	GAG, PROT, INT, RT RH
15	44	66	0.18	SCL44_TAR	6285	833-982	GAG,PROT, INT, RT, RH
16	98	156	0.10	SCL98_TAR	6209	855-899	GAG, PROT, INT, RT RH
17	182	265	0.06	SCL182_TAR	5536	585	GAG, PROT, INT, RT, RH
18	183	266	0.06	SCL183_TAR	6383	801-803	GAG, PROT, INT, RT, RH
19	101	159	0.10	SCL101_Tork	5546	329-337	GAG, PROT, INT, RT, RH
20	37	52, 165	0.29	SCL37_Tork	5082	459-465	GAG, PROT, INT, RT, RH
21	173	254	0.06	SCL173_Tork	5615	600-746	GAG, PROT, INT, RT, RH
22	243	335	0.03	SCL243_Tork	5586	750-773	GAG, PROT, INT, RT, RH
23	12	15	0.33	SCL12_Tork	5127	502	GAG, PROT, INT, RT, RH
24	89	140	0.11	SCL89_Tork	5049	509	GAG, PROT, INT, RT, RH
25	52	81	0.16	SCL52_Tork	4852	358-393	GAG, PROT, INT, RT, RH
26	40	58	0.19	SCL40_SIRE	7045	1,545	GAG, PROT, INT, RT, RH

Table 6.1. (Continued) Cluster, supercluster, genome proportion, structural protein domain, PBS and PPT features identified in Class_I|LTR|Ty1/ *copia* retrotransposon in *V. macrocarpon* (GAG = Gag domain, PROT = Protease, RT = Reverse transcriptase domain, RH = RNase H domain, INT = Integrase domain)

Short name	types of PBS	Sequence of PBS	Length of PPT	Sequence of PPT
SCL5_Ale	tRNA type: - AsnGTT		9	GAGGGGGAG
SCL22_Ale			10	TGAGGGGGAG
SCL110_Ale	At-chr1.tRNA162- MetCAT		14	AGCTTGAGGGGGAG
SCL112_Ale			9	GAGGGGGAG
SCL288_Ale				
SCL129_Alesia		TGGTATCAGAGCCGAGTTGGT CGGG	9	GAGGGRGAG
SCL344_Alesia		TGGTATCAAAGCTATAKYAA GG	15	AGCACAAAGGGGGAG
SCL35_Angela	At-chr1.tRNA162- MetCAT	TGGTATCAGAGCGGGCTGTA AT	10	ATAAAGTGAGA
SCL79_Bianca			16	AAGGAAAGAAGAAGAA
SCL42_Ivana	At-chr1.tRNA162- MetCAT	TGGTATCAGAGCTCAGGCTCA T	9	AAGGGGGGA
SCL178_Ivana			14	AGAAAAVATAAAAA
SCL310_Ivana	At-chr1.tRNA162- MetCAT		14	AGTTTAAGGGGGGA
SCL13_TAR	tRNA type GluTTT	AAGCTTAACAAAGATT		
SCL26_TAR	At-chr3.tRNA65- SerGCT	TAAAGCAGATTCTTTAAA	10	ATGTGGGGGA
SCL44_TAR	At-chr1.tRNA8- MetCAT		15	GAAAATGGTGGGGGA
SCL98_TAR	At-chr2.tRNA13- TyrGTA	ATTTTAAGACGTTTTTATCCG CAA	19	AAAAAATACAAGTGGGGGA
SCL182_TAR				
SCL183_TAR	At-chr2.tRNA6- TrpCCA		10	AAAAATTATA
SCL101_Tork	At-chr1.tRNA162- MetCAT	TGGTATCAGAGCAGTGTAGT	10	AAGGTGGAGA
SCL37_Tork	At-chr1.tRNA162- MetCAT	TGTTATCAGAGCCGCGGTTGC GGT	10	AAGTGGGAGA
SCL173_Tork	At-chr1.tRNA162- MetCAT	TGGTATCATAGCTCTARGTTT C	11	GCATTGGGGAA
SCL243_Tork			13	GGGAGGGGGAGAA
SCL12_Tork	At-chr1.tRNA162- MetCAT	TGGTATCAGAGCGGGGTTTA	10	AAGGTGGAGA
SCL89_Tork	At-chr1.tRNA162- MetCAT	TGGTATCAGAGCTCCGGG	11	AAGTGGGAGAA
SCL52_Tork	At-chr1.tRNA162- MetCAT	TGGTATCAGAGCTCTTGGTTG	10	AGGTGGAGAA
SCL40_SIRE	At-chr1.tRNA162- MetCAT			

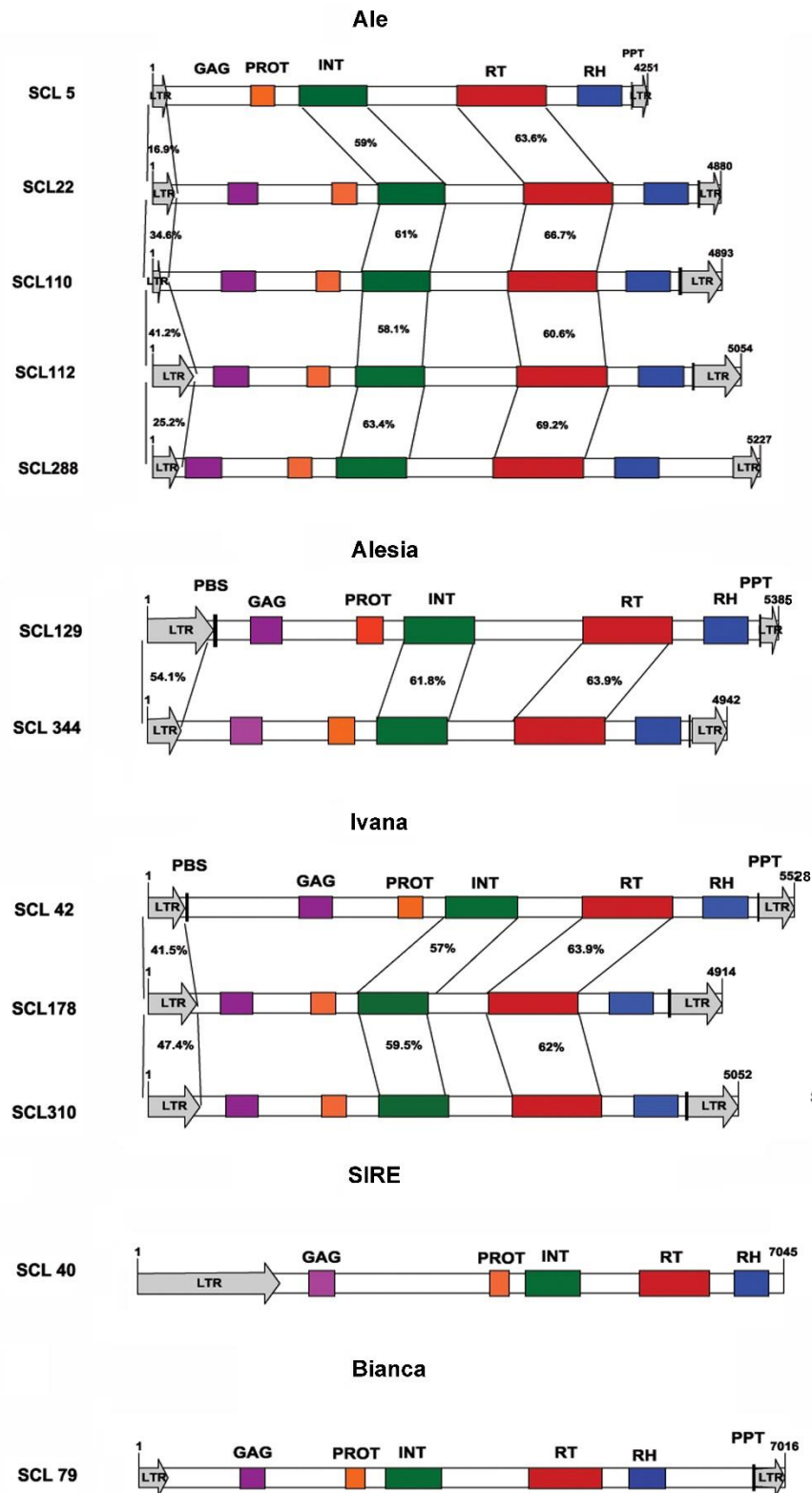


Figure 6.1. In depth structural diversity analysis of Ty1/*copia* retrotransposon in *V. macrocarpon*

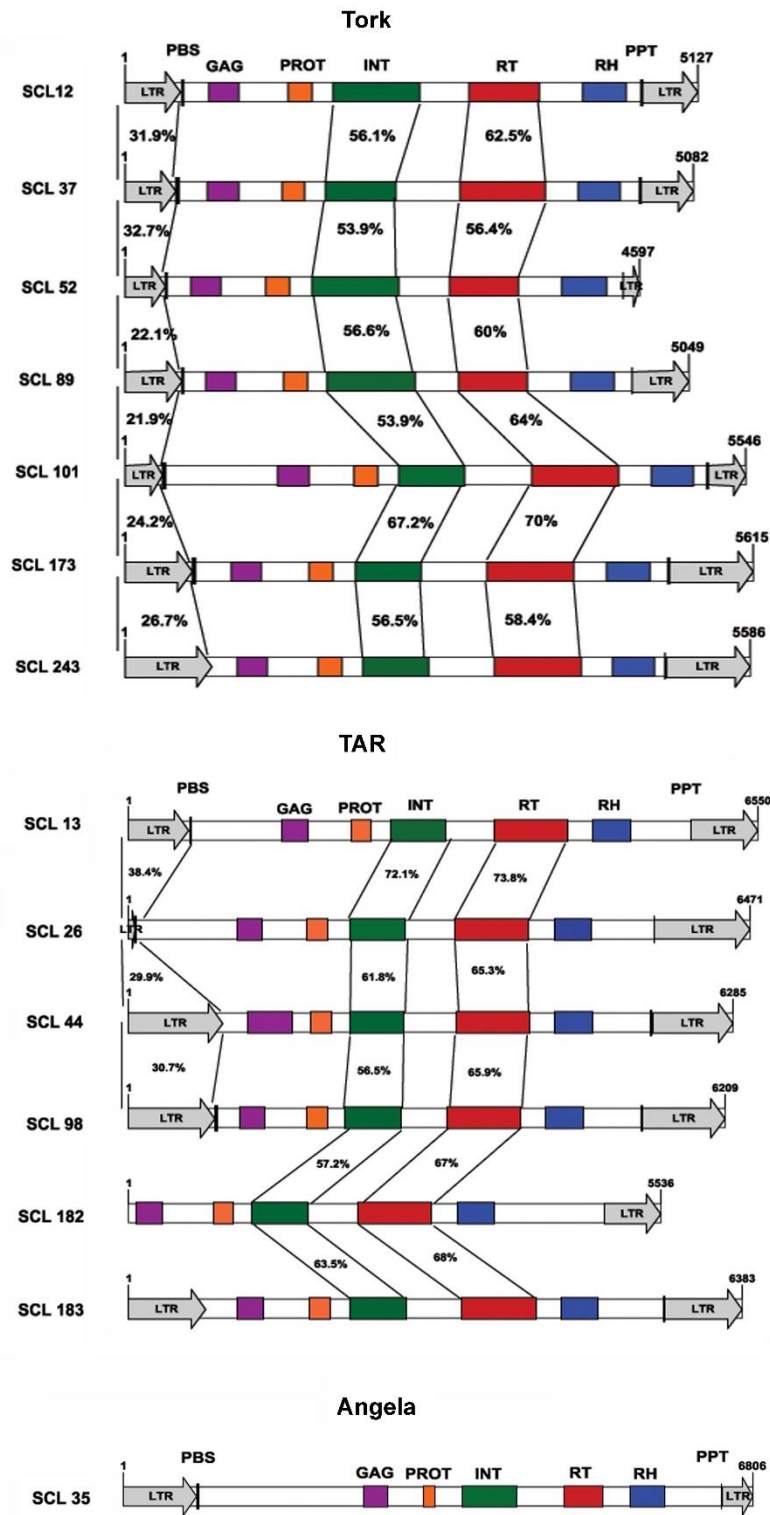


Figure 6.1. (Continued) In depth structural diversity analysis of Ty1/copia retrotransposon in *V. macrocarpon*

Table 6.2. Pairwise similarity distance matrix of LTR regions of identified full length Ty1/*copia* element in *V. macrocarpon* genome. The lowest values were colored with cyan and the highest values were with yellow.

	SCL5_ Ale	SCL22_ Ale	SCL110_ Ale	SCL112_ Ale	SCL288_ Ale		
SCL5_ Ale							
SCL22_ Ale	16.905						
SCL110_ Ale	13.559	34.568					
SCL112_ Ale	22.165	22.989	41.176				
SCL288_ Ale	22.532	21.687	33.333	25.223			
	SCL42_ Ivana	SCL178_ Ivana	SCL310_ Ivana				
SCL42_ Ivana							
SCL178_ Ivana	41.492						
SCL310_ Ivana	41.758	47.367					
	SCL12_ Tork	SCL37_ Tork	SCL52_ Tork	SCL89_ Tork	SCL101_ Tork	SCL173_ Tork	SCL243_ Tork
SCL12_ Tork							
SCL37_ Tork	31.882						
SCL52_ Tork	27.447	32.716					
SCL89_ Tork	27.199	26.207	22.122				
SCL101_ Tork	25.912	29.63	30.465	21.891			
SCL173_ Tork	30.17	29.435	23.417	42.743	24.225		
SCL243_ Tork	24.721	22.533	23.166	26.015	17.772	26.688	
	SCL13_ TAR	SCL26_ TAR	SCL44_ TAR	SCL98_ TAR	SCL182_ TAR	SCL183_ TAR	
SCL13_ TAR							
SCL26_ TAR	38.356						
SCL44_ TAR	22.818	29.851					
SCL98_ TAR	21.351	21.591	30.745				
SCL182_ TAR	28.694	21.901	23.817	24.842			
SCL183_ TAR	36.86	40.909	29.471	32.93	37.601		

Table 6.3. Pairwise similarity distance matrix of RT regions of identified full length Ty1/*copia* element in *V. macrocarpon* genome. The lowest values were colored with cyan and the highest values were with yellow.

	SCL5_ Ale	SCL22_ Ale	SCL110_ Ale	SCL112_ Ale	SCL288_ Ale		
SCL5_ Ale							
SCL22_ Ale	63.554						
SCL110_ Ale	61.77	66.732					
SCL112_ Ale	68.234	59.576	60.56				
SCL288_ Ale	71.79	62.855	58.85	69.218			
	SCL42_ Ivana	SCL178_ Ivana	SCL310_ Ivana				
SCL42_ Ivana							
SCL178_ Ivana	69.326						
SCL310_ Ivana	62.84	61.997					
	SCL12_ Tork	SCL37_ Tork	SCL52_ Tork	SCL89_ Tork	SCL101_ Tork	SCL173_ Tork	SCL243_ Tork
SCL12_ Tork							
SCL37_ Tork	62.701						
SCL52_ Tork	64.63	56.726					
SCL89_ Tork	63.826	70.718	59.968				
SCL101_ Tork	66.117	59.69	64.108	64.338			
SCL173_ Tork	64.503	58.154	64.744	61.058	70.463		
SCL243_ Tork	59.646	60.314	60.772	62.825	59.199	58.407	
	SCL13_ TAR	SCL26_ TAR	SCL44_ TAR	SCL98_ TAR	SCL182_ TAR	SCL183_ TAR	
SCL13_ TAR							
SCL26_ TAR	73.785						
SCL44_ TAR	65.734	65.343					
SCL98_ TAR	64.898	63.487	65.864				
SCL182_ TAR	62.905	63.986	67.164	67.045			
SCL183_ TAR	67.969	67.73	69.206	68.218	67.964		

Ty3/gypsy

Full-length elements were found in three main families (Ogre/Tat, Errantiviruses/Athila, and Chromoviruses) for Ty3/gypsy retrotransposons superfamily (Table 6.4.-6.6.).

The longest and most diversified LTR/gypsy element belonged to the Ogre/Tat group (Table 6.4.). Some of the elements of Ogre/Tat group are fragmented into many smaller cluster and constituted a single supercluster (Table 6.4.). This phenomenon was less frequent in other types of elements belonged to two other families of Ty3/gypsy (Athila and chromovirus). The main possible reason could be the size of the elements which are the longest among all other Ty3/gypsy elements. The size of Ogre/Tat elements ranged between 10.5-19.7 kp. A total of 13 full-length elements have been characterized from this family in the sublineages Ogre/Tat with the genome proportion ranged between 0.12-1.34%. The lengths of full-length LTR elements ranged from 646-4560 bp (Table 6.4.). Intact protein domain sequences of all full length have been identified (*Gag* and *Pol* gene). *Pol* gene of Ogre/Tat lineages was found to have dual ribonuclease H domain (RH and aRH) (Table 6.4. and Figure 6.2.). Pairwise sequence similarity for LTR and RT ranged between 10.9-35.6% and 55.4-74.67%, respectively (Table 6.5.-6.6.).

Family Athila had a total of 5 full-length elements with genome proportion ranged between 0.01-1.92%. Size of the elements ranged between 5.7-9.8 kb, with LTR length ranged between 854-1403 bp (Table 6.4.). All the important protein domain sequences (*Gag* and *Pol* gene) were found and depicted (Table 6.4. and Figure 6.2). Although intact primer binding sites (PBS) were found for all the full length Athila elements except one, intact polypurine tract (PPT) were not found for any of them (Table 6.4. and Figure 6.2). Pairwise sequence similarity of LTR and RT regions ranged between 8.7-49.6% and 69-81.8%, respectively (Table 6.5.-6.6.). Therefore, Athila elements were one of the most diversified LTR-retrotransposon elements in *V. macrocarpon*.

Family chromoviruses had a total of 10 full length elements. Elements belonged to three different sublineages (Tekay, CRM and Galadriel). Genome proportion ranged between 0.02-0.23%, full length 5.3-8.6 kb and LTR length 361-1925 bp. Among three sublineages Tekay had the highest length (Table 6.4.). All the internal protein domain (*Gag* and *Pol* gene) sequences were found and depicted. However, chromodomain

regions specific for chromovirus family were found for Tekay and Galadriel but not for CRM clade. PBS were found for most of the elements of CRM and Galadriel but not for all elements of Tekay clade. Moreover, some Tekay and Galadriel elements had two PPT sites (Table 6.4. and Figure 6.2.). Pairwise sequence similarities of LTR and RT elements were 47.6-54.5% and 67.05-87.5%, respectively (Table 6.5.-6.6.). Nevertheless, primer binding site is more diverse in case of Ty3/*gypsy* element than Ty1/*copia* group. Overall the length differences among the different lineages within the same family are mainly because of the differences in the LTR region or intergenic region compared to protein domain sequences of *gag/pol* gene (Table 6.5.-6.7. and Figure 6.2.).



Table 6.4. Cluster, supercluster, genome proportion, structural protein domain feature, sequence information of PBS and PPT of Ty3/*gypsy* retrotransposon (GAG = Gag domain, PROT = Protease, RT = Reverse transcriptase domain, RH = RNase H domain, a RH = addition RNase H domain, INT = Integrase domain, CHDII = Chromodomain

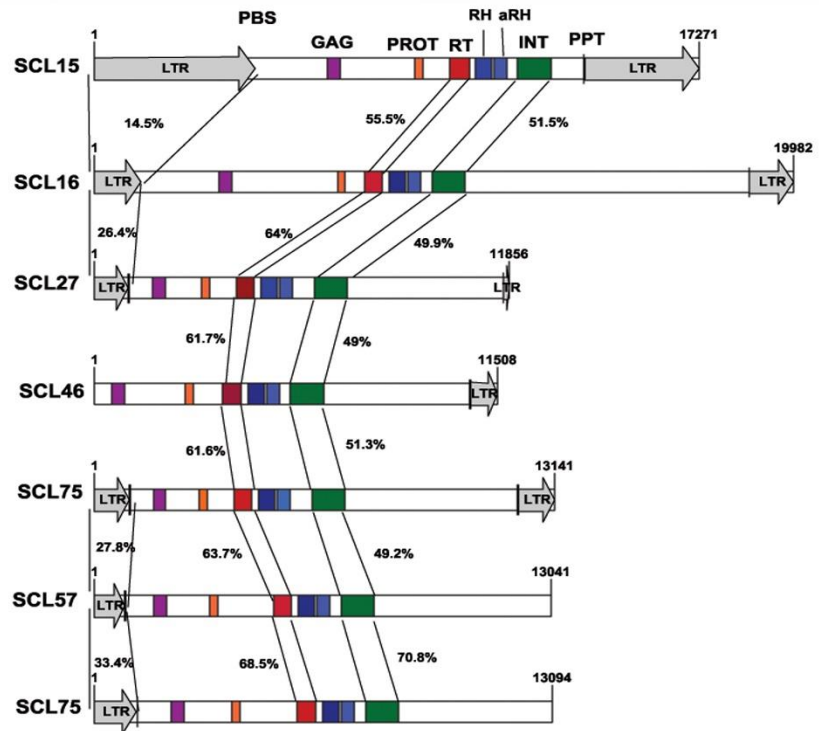
II)

SL	Super cluster	Cluster	Genome proportion (%)	Short name	length (bp)	Length of LTR (bp)	Internal protein domain
1	15	18, 36, 61, 73, 76, 101, 153	1.34	SCL15_Ogre	17271	3246-4590	GAG, PROT, RT, RH, aRH, INT
2	16	19, 13, 32, 34, 70, 114, 141	0.99	SCL16_Ogre	19982	1251-1326	GAG, PROT, RT, RH, aRH, INT
3	27	41, 90	0.23	SCL27_Ogre	11,856	955	GAG, PROT, RT, RH, aRH, INT
4	32	46	0.21	SCL32_Ogre	11,508	761	GAG, PROT, RT, RH, aRH, INT
5	49	75	0.17	SCL49_Ogre	13,141	1010-1014	GAG, PROT, RT, RH, aRH, INT
6	57	89, 373	0.15	SCL57_Ogre	13,041	864	GAG, PROT, RT, RH, aRH, INT
7	75	116, 359	0.12	SCL75_Ogre	13,094	1,211	GAG, PROT, RT, RH, aRH, INT
8	18	23, 24, 37, 80	1.02	SCL18_TatV	11,780	1643-1646	GAG, PROT, RT, RH, aRH, INT
9	46	69	0.17	SCL46_TatV	10,530	747	GAG, PROT, RT, RH, aRH, INT
10	55	87, 26, 112	0.15	SCL55_TatV	12,927	901-906	GAG, PROT, RT, RH, aRH, INT
11	59	94, 129	0.14	SCL59_TatV	11,611	726-752	GAG, PROT, RT, RH, aRH, INT
12	64	103	0.14	SCL64_TatV	11,279	858	GAG, PROT, RT, RH, aRH, INT
13	111	174, 195	0.17	SCL111_TatV	10,513	646	GAG, PROT, RT, RH, aRH, INT
14	6	7, 11, 29, 86, 122, 330	1.92	SCL6_Athila	9886	849-854	GAG, PROT, RT RH, INT
15	34	49	0.21	SCL34_Athila	6774	1,121	GAG, PROT, RT RH, INT
16	36	51	0.20	SCL36_Athila	8,673		GAG, PROT, RT RH, INT
17	100	158	0.10	SCL100_Athila	7,392	1403	GAG, PROT, RT RH, INT
18	352	449	0.01	SCL352_Athila	5781		GAG, PROT, RT RH, INT
19	125	193	0.08	SCL125_Tekay	6075	1149	GAG, PROT, RT, RH, INT, CHDII
20	248	341	0.03	SCL248_Tekay	5412		GAG, PROT, RT, RH, INT, CHDII
21	269	364	0.03	SCL269_Tekay	5545		GAG, PROT, RT, RH, INT, CHDII
22	28	9, 42	0.23	SCL28_Tekay	8697	1518-1925	GAG, PROT, RT, RH, INT, CHDII
23	31	45	0.22	SCL31_CRM	5501	396-553	GAG, PROT, RT, RH, INT
24	56	88	0.15	SCL56_CRM	5419	457-484	GAG, PROT, RT, RH, INT
25	80	124	0.12	SCL80_CRM	5284	361-556	GAG, PROT, RT, RH, INT
26	272	367	0.02	SCL272_CRM	5315		GAG, PROT, RT, RH, INT
27	213	300	0.05	SCL213_Galadriel	5319	465	GAG, PROT, RT, RH, INT, CHDII
28	214	301	0.05	SCL214_Galadriel	6413	1054-1061	GAG, PROT, RT, RH, INT, CHDII

Table 6.4. (Continued) Cluster, supercluster, genome proportion, structural protein domain feature, sequence information of PBS and PPT of Ty3/*gypsy* retrotransposon (GAG = Gag domain, PROT = Protease, RT = Reverse transcriptase domain, RH = RNase H domain, a RH = addition RNase H domain, INT = Integrase domain, CHDII = Chromodomain II)

Short name	types of PBS	types of PBS	Length of PPT	Sequence of PPT
SCL15_Ogre	Asn-1x	TTCAAATCCTTCTTGAGGAGcca	10	GGAGAAGGAA
SCL16_Ogre			10	
SCL27_Ogre	Arg-6x	TTCGATCCCCAGCAGAGTCGcca	12	GCGGGGGGGCCA
SCL32_Ogre			9	GGGGGGCAA
SCL49_Ogre	Arg-1x	TTGGATCCCCAGCGGAGTCGcca	9	GGAGGGAAA
SCL57_Ogre	Arg-27x	TTCGAGTCCCCTGGGCGTGcca		
SCL75_Ogre	Thr-2x	TCGAATCCGATAATTTGTTGcca		
SCL18_TatV	tRNA type-LysTTT	tRNA type-LysTTT	10	
SCL46_TatV	Lys-1x	TTTGATTCCCACAGACGGCGcca		
SCL55_TatV	Arg-1x	TTTGATCCCCAGTGGAGTTGcca		
SCL59_TatV		CTGGGTTCGATCCCCAGCAGAGTC	16	GGAGGAATCGAGG GG
SCL64_TatV			16	GGAAGAAGTTGAGG GG
SCL111_TatV	Lys-1x	TTTGATTCCCACAGACGGCGcca		GGGAACAGGGGAG
SCL6_Athila	tRNA type-AspGTC	tRNA type-AspGTC TGGCGCCGTTGCCGGGACGGCA		
SCL34_Athila	Asp-235x	TTCGATCCCCGGCAACGGCGcca		
SCL36_Athila	Asp-235x	TTCGATCCCCGGCAACGGCGcca		
SCL100_Athila	At-chr1.tRNA5-AspGTC,	At-chr1.tRNA5-AspGTC, CCGGGTTCGATCCCCGG CAACGGCG		
SCL352_Athila				
SCL125_Tekay			10 and 12	GAGGACGAAA and AAGGGGGATAGG
SCL248_Tekay				
SCL269_Tekay				
SCL28_Tekay	Gln-1x	TCGATTCCTTCTGGTGCCAAcca	10 and 13	AAGGAGGGTAGAA and GAGGACAAAA
SCL31_CRM	Met-1x	ATTGAAACCTCGCTCTGATAcca	13	AAAGAATAAGGRA
SCL56_CRM	Met-1x	ATTGAAACCTCGCTCTGATAcca	9	AATAAATAAG
SCL80_CRM	Met-1x	ATTGAAACCTCGCTCTGATAcca	12	AGAAGTGAAGA
SCL272_CRM				
SCL213_Galadriel	Met-1x	ATCAAAACCTGGCTCTGATAcca	11 and 13	AAGGAGGGAGA and AGGTGGGGGAGAG
SCL214_Galadriel			11	AGTGGGGGTGG

Ogre/Tat|TatIV/Ogre



Ogre/Tat|TatV

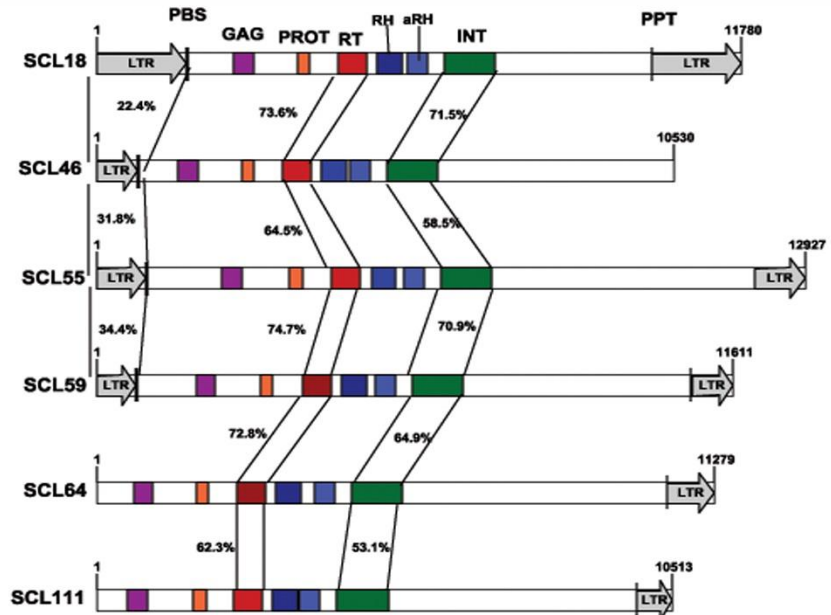
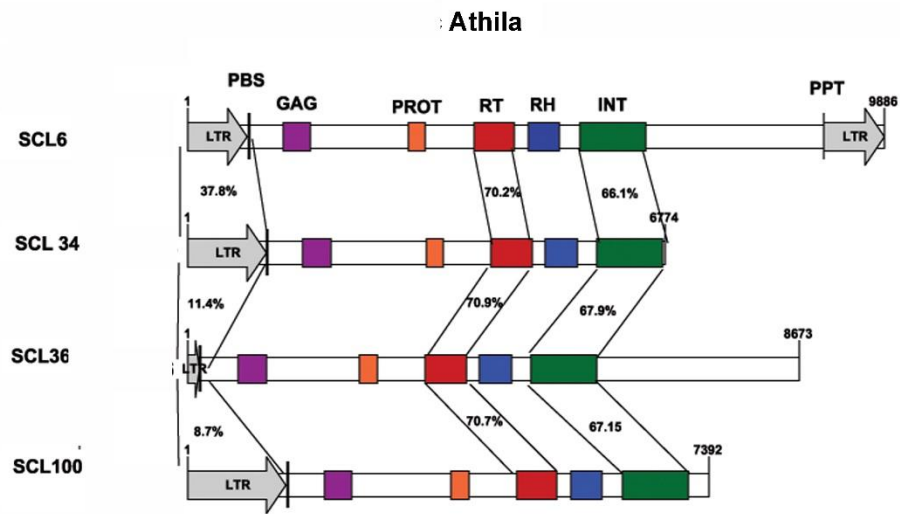


Figure 6.2. In depth structural diversity analysis of Ty3/gypsy retrotransposon in *V. macrocarpon*

Non-chromovirus|OTA



Chromovirus

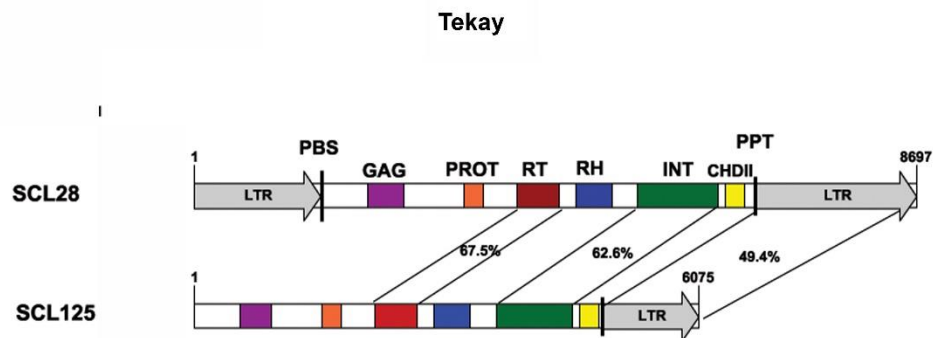
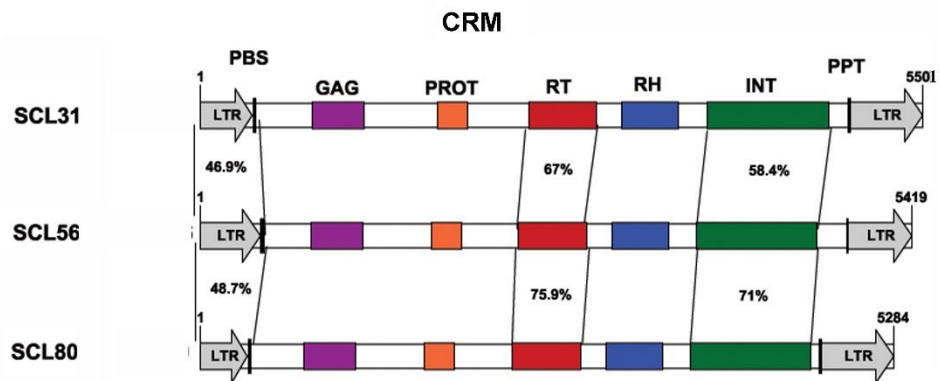


Figure 6.2. (Continued) In depth structural diversity analysis of Ty3/*gypsy* retrotransposon in *V. macrocarpon*

Chromovirus

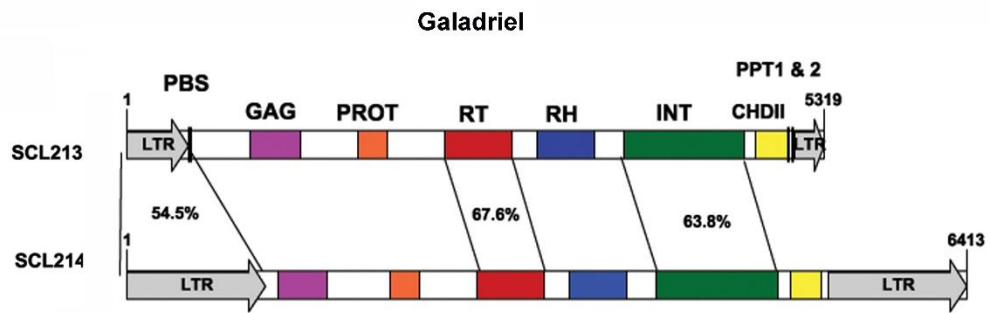


Figure 6.2. (Continued) In depth structural diversity analysis of Ty3/gypsy retrotransposon in *V. macrocarpon*

Table 6.5. Pairwise similarity distance matrix of LTR regions of identified full length Ty3/gypsy element in *V. macrocarpon* genome. The lowest values were colored with cyan and the highest values were with yellow.

	SCL15_ Ogre	SCL16_ Ogre	SCL27_ Ogre	SCL32_ Ogre	SCL49_ Ogre	SCL57_ Ogre	SCL75_ _Ogre
SCL15_Ogre							
SCL16_Ogre	14.497						
SCL27_Ogre	10.29	26.436					
SCL32_Ogre	8.437	23.627	24.868				
SCL49_Ogre	11.062	26.609	36.013	28.777			
SCL57_Ogre	10.152	28.889	28.446	27.694	27.773		
SCL75_Ogre	14.271	30.55	28.501	25.661	28.721	33.391	
	SCL18_ TatV	SCL46_ TatV	SCL55_ TatV	SCL59_ TatV	SCL64_ TatV	SCL111_ TatV	
SCL18_TatV							
SCL46_TatV	22.395						
SCL55_TatV	27.749	31.751					
SCL59_TatV	23.317	34.211	34.404				
SCL64_TatV	26.804	31.68	30.721	34.696			
SCL111_TatV	20.296	35.558	30.031	34.747	28.308		
	SCL6_ Athila	SCL34_ Athila	SCL36_ Athila	CL100_ Athila			
SCL6_Athila							
SCL34_Athila	37.448						
SCL36_Athila	12.888	11.398					
CL100_Athila	32.777	49.632	8.689				
	SCL31_ CRM	SCL80_ CRM	SCL80_ CRM				
SCL31_CRM							
SCL56_CRM	47.622						
SCL80_CRM	54.296	48.687					

Table 6.6. Pairwise similarity distance matrix of RT regions of identified full length Ty3/gypsy element in *V. macrocarpon* genome. The lowest values were colored with cyan and the highest values were with yellow.

	SCL15_ Ogre	SCL16_Ogre	SCL27_Ogre	SCL32_Ogre	SCL49_ Ogre	SCL57_ Ogre	SCL75_ _Ogre
SCL15_Ogre							
SCL16_Ogre	55.463						
SCL27_Ogre	58.241	64					
SCL32_Ogre	66.728	56.006	61.676				
SCL49_Ogre	58.673	62.73	73.73	61.613			
SCL57_Ogre	55.638	58.651	60.825	56.84	63.746		
SCL75_Ogre	57.839	61.032	63.514	59.118	63.921	68.464	
	SCL18_ TatV	SCL46_TatV	SCL55_TatV	SCL59_TatV	SCL64_ TatV	SCL111_ _TatV	
SCL18_TatV							
SCL46_TatV	73.563						
SCL55_TatV	70.172	64.532					
SCL59_TatV	66.826	65.01	74.665				
SCL64_TatV	65.805	64.847	74.187	72.753			
SCL111_TatV	66.667	63.889	65.105	63.48	62.261		
	SCL6_ Athila	SCL34_ Athila	SCL36_ Athila	SCL100_ Athila	SCL352_ Athila		
SCL6_Athila							
SCL34_Athila	70.191						
SCL36_Athila	77.47	70.855					
SCL100_Athila	70.451	81.802	70.711				
SCL352_Athila	70.104	72.155	69.324	72.27			
	SCL31_ CRM	SCL80_CRM	SCL80_CRM				
SCL31_CRM							
SCL56_CRM	67.05						
SCL80_CRM	68.87	75.865					

Comparative structural feature analysis of Ty1/copia and Ty3/gypsy

Five Ty1/*copia* lineages (Retrofit, Bianca, Ivana/Oryco, Tork, and Sireviruses/Maximus) and three Ty3/*gypsy* lineages (Ogre/Tat, Errantiviruses/Athila, and Chromoviruses) were found and depicted in *V. macrocarpon* genome (Table 6.7.). Errantiviruses/Athila (a total of 5 full length elements and genome proportion range (0.01-1.92%)) and Ogre/Tat (a total of 13 full length elements and genome proportion range (0.12-1.34%)) lineages were found to be the most diversified and abundant LTR-retrotransposon in *V. macrocarpon* genome (Table 6.7.). In addition, Ogre/Tat lineage was the biggest LTR-retrotransposon (full length range 11.5-19.9 kb) among all other LTR-retrotransposons. Nonetheless, the longest LTR-length was belonged to Sireviruses/Maximus (1,545 bp) of Ty1/*copia* group. The most common primer binding site (tRNA) in Ty1/*copia* and Ty3/*gypsy* were belonged to Met- tRNA. Pairwise sequence similarity of reverse transcriptase domain sequences (RT) were much higher (57-81%) compared to their LTR region (8-54%). Therefore, sequences were more homogeneous in the reverse transcriptase domain region compared to the LTR region (Table 6.7.).

Table 6.7. General structural features and repeat compositions of full length LTR-retrotransposons

Super famil/y Family/Lineage	Subfamily/ Sublineage	Total full length elements	Genome proportion (%)	Size (kb)	LTR (bp)	PBS (tRNA types)	Range of pairwise similarity (%)	
							LTR	RT
Ty1/ <i>Copia</i>								
Retrofit							LTR	RT
	Ale	5	0.02-0.57	5.2-4.9	401-127	Asn; Met	17-41	58-72
	Alesia	2	0.01-0.08	5-5.4	289-562		54	63
Bianca								
	Bianca	1	0.12	7	313-326			
Ivana/Oryco								
	Ivana	3	0.02-0.19	4.9-5.5	417-438	Met	41.45- 47.37	62-69
Tork								
	TAR	6	0.06-0.33	5.5-6.5	585-994	Glu, Ser, Met, Tyr, Trp	21-41	63-74
	Angela	1	0.21	6.8	796	Met		
	Tork	7	0.06-0.33	4.8-5.6	329-773	Met	17-43	57-70
Sireviruses/ Maximus								
	SIRE	1	0.19	7	1,545	Met		
Ty3/ <i>gypsy</i>								
Errantiviruses/ <i>Athila</i>								
	Athila	5	0.01-1.92	5.7-9.8	854-1403	Asp	8.7- 49.6	69-81.8
Chromoviruses								
	Tekay	4	0.03-0.23	5.4-8.6	1149- 1925	Gln	49.4	87.5
	CRM	4	0.02-0.22	5.3-5.5	361-556	Met	47.6- 54.29	67.05- 75.86
	Galadriel	2	0.05	5.3-6.4	465-1061	Met	54.5	67.6
Ogre/Tat								
	Ogre	7	0.12-1.34	11.5- 19.9	761-4560	Asn, Arg, Thr	10.9- 28.9	55.4- 73.73
	TatV	6	0.14-1.02	10.5- 11.7	646-1646	Lys, Arg	20.3- 35.6	62.26- 74.67

CHAPTER VII

DISCUSSION

Significant morphological differences (leaf shape, leaf color, fruit shape, fruit color, plant shape and size) have been observed among cultivated and wild Turkish *Vaccinium* species. In addition, it is found that the cultivated *V. corymbosum* L. (cultivar ‘Jubilee’, $2n = 4x = 48$) and Turkish wild *V. arctostaphylos* L. ($2n = 4x = 48$) are tetraploid species which is complementary of the study of Çelik, (2012). Therefore, these two high bush type species are morphologically quite similar to each other than Turkish low bush type species (*V. myrtillus* L., $2n = 2x = 24$ and *V. uliginosum* L., $2n = 2x = 24$) which are actually diploid species. In the natural condition of the Kaçkar Mountains of Black Sea region of Turkey, normally the diploid species occupied separate location (distant cluster) than the tetraploid species. Smith et al. (2015) studied the similar phenomenon in case of *Vaccinium* section *Oxycoccus* and they reported the occurrence of morphologically distinguishable diploid and tetraploid species, with distinct distribution in the natural condition. Polyploidization is a natural phenomenon of genome evolution and speciation arise through the interplay within repetitive sequences mainly transposable elements. Therefore, analysis of repetitive sequences in a comparative manner could reveal the impact of polyploidization on the genome evolution of related species or vice versa (Wendel et al., 2016; Vicient and Casacuberta, 2017).

Repetitive sequences are obvious challenges for sequence alignment and assembly program because of unavoidable technical drawbacks and due to their ambiguous nature and abundant distribution throughout the genome. For instance, human genome is still not well assembled especially in the repeat regions (Miga, 2015). Therefore, annotation of these vast diverse repeat sequences with fast and efficient technology is a prime strategy to have a fine-annotated improved genome assembly with closed gap (Treangen and Salzberg, 2012; Yin et al., 2015). As far as our knowledge, *Vaccinium* genome assembly (blueberry and cranberry) is still ongoing and the current genome assembly contains many gaps and repetitive sequences are not yet characterized (Brown et al., 2011; Mudd et al., 2013; Polashock et al., 2014; Bian et al., 2014; Gupta et al., 2015). In this study, RepeatExplorer based estimation and characterization of the repetitive sequences were carried out in *Vaccinium* genome, a recent and powerful technique to detect overall

repeat sequence landscape of the studied genome (Novák et al., 2013; Novák et al., 2017). It was found that the genus *Vaccinium* is rich in repetitive sequences and the percentage ranges from 80-90% supporting the studies of Polashock et al. (2014) and Li et al. (2016). Although genic sequences are more or less constant, percent of repetitive sequences in the genome is highly variable within the plant kingdom (especially closely related species) and could be a characteristic feature for a particular species and its genome size (Macas et al., 2015; Wendel et al., 2016; Bombarely et al., 2016; Pellicer et al., 2018). Current knowledge on the genome size variation of *Vaccinium* reveals that genome sizes are variable across the species regardless the ploidy level differences (Hummer et al., 2015).

In-depth analysis of repeat sequences from the two common *Vaccinium* species (blueberry and cranberry) reveals that the most abundant repeat sequences are LTR-retrotransposons, followed by DNA transposons, satellites, minisatellite repeats and other types of repetitive sequences. Comparative genome analysis of repetitive sequences among blueberry and cranberry shows that satellite repeats are more diversified in the blueberry genome (6 satellite families), but LTR-retrotransposon sequences are more diversified in cranberry genome (3 satellite families) (Sultana et al., 2017). Differential accumulation of different types of repetitive sequences in closely related species is an interesting species-specific phenomenon and directly related with the environmental impact that species has been gone through (Dodsworth et al., 2014, 2015 and 2017). For instance, the main driving factor responsible for the genome size differences in the family Fabaceae is LTR-retrotransposons (57%), mainly Ogre/Tat lineages of the Ty3/gypsy family (Macas et al., 2015). Meanwhile, in *Fritillaria affinis* of the family Liliaceae, satellite repeats are highly diversified and accounts for about 35% of the genome; however, in other species of this genus *Fritillaria*, satellite repeats is not much diversified (Kelly et al., 2015).

The percentage of satellite repeats varies between 0.06-3.00% in the genome of *Vaccinium*. As far as our knowledge, this is the first record of satellite repeats in the genus *Vaccinium*. Moreover, sequence diversity, periodicity and homogeneity among seven individual tandemly organized repeats sequences have been characterized from six different *Vaccinium* species (*V. corymbosum* cultivar 'Jubilee', *V. corymbosum* strain 'W850', *V. macrocarpon* cultivar 'Ben Lear', *V. arctostaphylos*, *V. myrtillus*, *V.*

uliginosum) representing five different economically important section (*Cyanococcus*, *Oxycoccus*, *Hemimyrtillus*, *Myrtillus*, *Vaccinium*). Primers were designed for satellite amplification in such a way so that the maximum diversity within and among the species would be revealed, followed by random cloning of PCR fragments from different species. Such a combined strategy (NGS Seq analysis, cloning and molecular analysis) was believed to increase the chance of getting more representative satellite repeats from different species. Combined bioinformatics, PCR amplification and clone analysis reveal that VaccSat2, 3 and 7 are the common satellites in all studied *Vaccinium* species. In contrast, VaccSat1, 5 and 6 show some degree of species-specific amplification. Species specific amplification of satellite repeats is a well-known phenomenon and studied extensively in the species of the genus *Solanum* (Tek et al., 2005), *Beta* (Zakrzewski et al., 2013), *Vicia* (Macas et al., 2006; Robledillo et al., 2017) and *Camellia* (Heitkam et al., 2015) and many other plant and animal genome (Sharma and Raina, 2005; Melters et al., 2013). Even though exact mechanism behind the rapid satellite repeat amplification or elimination within closely related species is still unknown, evidence indicates that unequal chromosome exchange, rolling circle replication, and gene conversion could be responsible for this phenomenon (Macas et al., 2002; Garrido-Ramos, 2017).

Nonetheless, phylogenetic analysis of seven satellites repeats from different *Vaccinium* species produce seven individual clusters representing the seven individual satellite repeats, significant level of sequence similarity is present among the satellite families. For instance, VaccSat1 and 4 seem to belong to the same subfamily and can be considered as a single satellite family. Moreover, VaccSat2 and 7 as well as VaccSat3 and VaccSat6 have significant sequence similarities. Ruiz-Ruano et al. (2016) propose the satellite “library” hypothesis to explain the phenomenon of inter-satellite similarity. According to this hypothesis, closely related species shared a common conserved ancestral set of satellite repeats (seed of satellite) and each of them differentially amplified in different species. Another well-established phenomenon of origin, evolution, spreading and dissemination of satellite DNA from transposable elements could also explain the situation of inter-satellite similarity (Meštrović et al., 2015).

Monomer length of the satellite repeats seems to be consistent among the species (ranged between 36-234 bp for seven satellite families). Several authors pointed out that monomer size ranged between 160-180 bp and 320-360 bp could facilitate DNA wrapping and

phasing around the nucleosome of the centromeric region, an important feature of centromeric DNA packaging and centromere function in many species (Schmidt and Heslop-Harrison, 1998; Melters et al., 2013; Heslop-Harrison and Schwarzacher, 2013; Yang et al., 2017). All the satellite repeats were found to have AT-rich regions, while VaccSat1, 5 and 6 showed an overrepresentation of AA/TT region. AT-rich regions are thought to be linked with the intrinsic bending of the DNA molecule and differential DNA packaging in euchromatic and heterochromatic regions (Macas et al., 2002; Melters et al., 2013). VaccSat2 have high GC region which is very often targeted for RNA-directed DNA methylation and epigenetic fingerprinting (Macas et al., 2002; Mehrotra and Goyal, 2014; Garrido-Ramos, 2017). VaccSat7 is highly heterogeneous due to the addition or deletion of “CAAAAAA” motif, the similar pentanucleotide repeat motif “CAAAA” in the satellite repeats are assumed to be linked with the breakage-reunion cycle of satellite repeats (Macas et al., 2002).

VaccSat1, 2, 5 and 6 showed higher order repeat (HOR) unit structure in all of the studied species, however, no higher order repeat structure has been found for VaccSat3 and 7. In VaccSat1, the subunit structure was two direct repeats of the same length of 28 bp followed by a 101 bp longer subunit constituting a total monomer length of 147-149 bp. For VaccSat2, the subunit is approximately 122 bp and two direct subunits (122 bp) constitute a 234 bp monomer unit. Nevertheless, these two subunits seem to be highly heterogeneous. For VaccSat5, subunits are 10 bp and 26 bp and constitute the final monomer length of 36-38 bp. In case of VaccSat6, two direct subunits of 17 bp and 32 bp constitute the final monomer length of 47-48 bp. The formation of satellite subunit, organization of monomer in higher order structure and subsequent multimer formation has been described in many different species by several authors in different plant species in the family of Fabaceae (Macas et al., 2006), Chenopodiaceae (Menzel et al., 2008; Zakrzewski et al., 2010), Theaceae (Heitkam et al., 2015). HOR structure of satellite repeats are directly linked with the formation of new tandem repeats of larger monomer unit and evolution of satellite repeats (Garrido-Ramos, 2017). Melters et al. (2013) analyzed satellite repeats from 100 different plant and animal species and found that HOR structure in the tandem repeats of centromeric region are particularly related with centromere evolution. Therefore, polyploidization and HOR structure and concerted evolution of centromeric satellite repeats are the well-studied topics of species genome evolution (Plohl et al., 2012; Garrido-Ramos, 2017; Yang et al., 2018)

Sequence homogeneity analysis among the studied satellite repeats prove that VaccSat1, 2 and 3 are more homogeneous and show high sequence similarity among and within the species with occasional deletion of the certain sequence motif. In different *Vaccinium* species, variants of VaccSat1, 2 and 3 are found to be intermixed within intra and inter-species level and devoid of genome specific sequence homogeneity. In contrast, VaccSat5, 6 and 7 are more heterogeneous as well as VaccSat5 and 6 show species-specific sequence variants. High genome wide sequence homogeneity is thought to be the result of concerted evolution, while sequences are more homogeneous in intra-species level but more heterogeneous in inter-species level (Plohl, 2010; Garrido-Ramos, 2015). Garrido-Ramos, (2017) explained that sequence homogenizations are directly linked with several important factors like biological (reproductive biology), environmental, chromosomal localization, copy number, functional constraint, and evolutionary biology. For instance, the Y chromosomes of human (Skaletsky et al., 2003) and *Rumex acetosa* (Mariotti et al., 2009) do not recombine during meiosis which could explain homogenizing events of Y-linked satDNA. In addition, subtelomeric and pericentromeric satellite repeats often show high sequence diversity because those particular locations of the chromosome are the hot spots for recombination mechanism (Macas et al., 2006; Torres et al., 2011). Moreover, high heterogeneous nature could also be linked with comparatively young satellite repeats where concerted evolution was not efficient enough (Macas et al., 2006).

These results shed the light on the points to study the taxonomic position of widely distributed *Vaccinium* species throughout the world based on the identified satellite repeats covering its diverse ploidy level and section. Satellite repeats diversity has been used to study genome dynamics and phylogenetics analysis of many species of *Beta* (Menzel et al., 2008), *Dendrobium* (Begum et al., 2009), *Camellia* (Heitkam et al., 2015), *Fabeae* (Macas et al., 2015). Such analysis is very efficient to study the taxonomically intrigued but economically important species and genus whose genomes are the result of repeated cyclical polyploidization and interspecific hybridization.

Transposable elements are the most abundant repetitive sequence and RepeatExplorer based analysis only identify 9.53% and 22.59% of the transposable element from the blueberry and cranberry genomes. However, this estimation is noticeably less according to the report of Polashock et al. (2014) who characterized that 39.53% of cranberry

genome belongs to transposable elements. The main reason of this deviation could be related with the abysmal quality of publicly available ILLUMINA next generation sequence reads (Macas et al., 2015, Novak et al., 2013 and 2017). Ty3/*gypsy* and Ty1/*copia* LTR retrotransposon are the most abundant transposable elements followed by L1 type non-LTR retrotransposon. Comparative heterogeneity analysis of the LTR-retrotransposon sequences from blueberry and cranberry genome reveal differential diversity pattern for the individual transposable elements in two common cultivated *Vaccinium* species. Differential heterogeneity of retrotransposon elements is also studied in many plants species, for instance Rice (Kubis et al., 1998), Maize (Meyers et al., 2001), *Setaria* (Benabdelmouna et al., 2003), *Pear* (Yin et al., 2015), *Populus* (Natali et al., 2015), and several other economically important plant species (Wendel et al., 2016). This phenomenon is thought to be directly linked with the phylogenetic relationship, evolution and environmental adaptation (Wendel et al., 2016). Phylogenetic analysis of reverse transcriptase protein domain sequences from *V. corymbosum* and *V. macrocarpon* reveals that Ogre/Tat and Ale/Retrofit is the highly diversified lineages of Ty3/*gypsy* and Ty1/*copia* retrotransposons, respectively. These features could be the characteristics of *Vaccinium* genome because plant-specific lineage diversification is reported by many authors in *Arabidopsis thaliana* (Marín and Lloréns, 2000), Triticeae, rice and *Arabidopsis* (Wicker and Keller, 2007), Sugarcane (Domingues et al., 2012), *Chenopodium quinoa* (Kolano et al., 2013), *Fabeae* (Macas et al., 2015).

A total of 28 full-length Ty3/*gypsy* and 26 Ty1/*copia* elements are identified and reconstructed from the *V. macrocarpon* genome. In this study, Ty3/*gypsy* and Ty1/*copia* elements are subdivided into lineage. Lineages are further subdivided into sublineages because sequences are highly diverse considering their sequence similarities and sizes. The lineages for Ty3/*gypsy* are Errantiviruses/Athila, Chromoviruses with sublineages Del/Tekay, Reina, Galadriel and CRM, and Ogre/Tat with sublineages Ogre and Tat. On the other hand, Ty1/*copia* is subdivided into Sireviruses/Maximus, Tork with sublineages TORK, TAR and Angela, Retrofit with sublineages Ale and Alesia, and Oryco/Ivana. Subclassification of the lineages into sublineages was done according to Novak et al. (2013). However, there are other classification systems of LTR retrotransposons, which does not consider subclassification into lineages (Wicker et al., 2007; Llorens et al., 2010).

For all the full-length elements, the characteristic features like functional protein domain sequence (*Gag* and *pol* gene), primer binding site (PBS), polypurine tract (PPT) and long terminal repeats (LTR) regions are identified and characterized. While *Gag* is a single gene, *pol* are constituted with multiple gene and common domains are like reverse transcriptase, protease, ribonuclease H, and integrase. Ustyantsev et al. (2017) explain that reverse transcriptase is the only core and ancient protein domain of LTR retrotransposons and acquiring of other polyproteins domain are the results of independent evolutionary history of each specific lineages happened independently in the organisms life history. This kind of theory of retrotransposon evolution is also termed “modular view”. Therefore, structural differences within the lineages are quite significant. For instance, in *V. macrocarpon*, all the Ogre/Tat lineages of Ty3/*gypsy* family have dual ribonuclease H domain. Ustyantsev et al. (2015) speculate that dual ribonuclease H domain is necessary for successful strand transfer a property common for retroviruses. Therefore, dual ribonuclease H domain is acquired by Ogre/Tat lineage through convergent evolutionary process which has natural selection pressure on the elements (Ustyantsev et al., 2015). Moreover, chromodomain region is present in Galadriel and Tekay sublineages but not in CRM sublineage. Weber et al. (2013) reported that in the genus *Beta*, CRM clade has more diversified chromodomain sequences than the Galadriel and Tekay clade. Weber et al. (2013) also explained that chromodomain protein has the similarity with the heterochromatin protein (HP1) and have the ability to bind specific histone variants. Therefore, chromoviruses could have impact on chromatin structure and function.

It was found that elements within the lineages are structurally similar with high sequence similarity in the protein domain region, although the noncoding and LTR regions are highly diverse or incomplete for a significant number of the elements of both Ty3/*gypsy* and Ty1/*copia*. The range of pairwise genetic distance of LTR region is 17-54% and 8-54% for Ty1/*copia* and Ty3/*gypsy* retrotransposons, respectively. The highly diverse noncoding and LTR regions among sublineages level may indicate their ancient origin and diversification in the genome of *Vaccinium* (Weber et al., 2013). In addition, presences of incomplete or truncated LTR-regions indicate the inactive elements (Wollrab et al., 2012). On the other hand, elements those have mainly intact LTR regions with high sequence similarity are evolutionarily young elements and mostly transcriptionally active sequences (Weber et al., 2010). Domingues et al. (2012)

summarize that in sugarcane, Ty1/*copia* elements are more transcriptionally active than Ty3 *gypsy* elements. Moreover, most of transcriptionally active lineages of Ty1/*copia*, like Ale/Retrofit and Sireviruses are preferentially accumulate in euchromatic region and hence could have significant effect on the gene regulation of the genome (Bousios et al., 2010; Domingues et al., 2012; Bousios et al., 2012). On the other hand, Wollrab et al. (2012) explain that some of the members of Athila/Errantiviruses of Ty3/*gypsy* family are truncated and localized in the heterochromatic regions.

Although in this study the first insight of the landscape of repeat sequence of the economically important *Vaccinium* species (cultivated and wild) was analyzed and quantified, the abundance, copy number, localization, and transcriptional activity of these diversified repeat elements from the *Vaccinium* genome is still missing. Comparative analysis of repetitive sequence focusing on the abundance and distribution is an advance strategy of the present era to study genome evolution based on taxon-specific manner and ploidy level (Menzel et al., 2008; Macas et al., 2015; Heitkam et al., 2015). The other important research area related with repetitive DNA analysis are the transcriptional activity of repetitive DNA, regulation of repetitive DNA through small RNAs mediated mechanism, importance of repetitive DNA on chromatin regulation and gene regulation, and dynamism of repetitive DNA on polyploidy and speciation (Meyers et al., 2001; Neumann et al., 2011; Wendel et al., 2016; Garrido-Ramos, 2017; Chuong et al., 2017; Pellicer et al., 2018). As *Vaccinium* is an old genus and significant amount of autopolyploidization and allopolyploidization have been recorded, analysis of repeat sequences will facilitate the investigation of the taxonomic complexity as well as genome dynamics of this economically important plant.

CHAPTER VIII

CONCLUSION

In conclusion, in this study *Vaccinium* genome structure was revealed for the first time through ploidy level estimation and repetitive DNA analysis. A comprehensive survey of repetitive DNA analysis was performed in a comparative manner to understand the impact of repetitive DNA on *Vaccinium* genome evolution. Identification, quantification and analysis of repetitive DNA sequences were performed through advance bioinformatics technique using publicly available sequence data and cloning based approach. The most dynamic parts of repetitive DNA showing the species specific genome enrichment within the evolutionary timescale were satellite repeats. A total of six satellite families have been identified and characterized based on their monomer length diversity and sequence homozenization. Although these satellite repeats constitute a non-significant portion of the genome of different *Vaccinium* species, their homozenization and amplification pattern were highly diverse among different species. Therefore, a new door has been opened to use satellite repeats diversity to charactreize the wild and cultivated *Vaccinium* species. Meanwhile, a total of 54 full length LTR-retrotransposon sequences were reconstructed and assigned to three lineages of Ty3/*gypsy* and five lineages of Ty1/*copia* LTR-retrotransposon. In addition, exciting structural features of identified LTR-retrotransposons have been characterized. Overall this study expands the genomic insight of *Vaccinium*. However, more detailed studies considering the diverse species and ploidy level are necessary to better unveil the effect of repetitive DNA on *Vaccinium* genome evolution.

REFERENCES

Babicki, S., Arndt, D., Marcu, A., Liang, Y., Grant, J.R., Maciejewski, A. and Wishart, D.S., “Heatmapper: web-enabled heat mapping for all”, *Nucleic Acids Research* 44(W1),W147-W153, 2016.

Ballington, J.R., “Collection, utilization, and preservation of genetic resources in *Vaccinium*”, *HortScience* 36(2), 213-220, 2001.

Bao, W., Kojima, K.K. and Kohany, O., “Rebase Update, a database of repetitive elements in eukaryotic genomes”, *Mobile DNA* 6(1), 11, 2015.

Begum, R., Zakrzewski, F., Menzel, G., Weber, B., Alam, S.S. and Schmidt, T., “Comparative molecular cytogenetic analyses of a major tandemly repeated DNA family and retrotransposon sequences in cultivated jute *Corchorus* species (Malvaceae)”, *Annals of Botany* 112(1), 123-134, 2013.

Begum, R., Alam, S.S., Menzel, G. and Schmidt, T., “Comparative molecular cytogenetics of major repetitive sequence families of three *Dendrobium* species (Orchidaceae) from Bangladesh”, *Annals of Botany* 104(5), 863-872, 2009.

Benabdelmouna, A. and Darmency, H., “Copia-like retrotransposons in the genus *Setaria*: Sequence heterogeneity, species distribution and chromosomal organization”, *Plant Systematics and Evolution* 237(3-4), 127-136, 2003.

Bennett, M.D. and Leitch, I.J., “Genome size evolution in plants”, In *The evolution of the Genome*, 89-162, 2005.

Bennetzen, J.L., “The contributions of retroelements to plant genome organization, function and evolution”, *Trends in Microbiology* 4(9), 347-353, 1996.

Benson, G., “Tandem repeats finder: a program to analyze DNA sequences”, *Nucleic Acids Research* 27(2), 573, 1999.

Besse, P., Molecular Plant Taxonomy, Methods and protocols, *Humana Press*, 2014.

Bian, Y., Ballington, J., Raja, A., Brouwer, C., Reid, R., Burke, M., Wang, X., Rowland, L.J., Bassil, N. and Brown, A., “Patterns of simple sequence repeats in cultivated blueberries (*Vaccinium* section *Cyanococcus* spp.) and their use in revealing genetic diversity and population structure”, *Molecular Breeding* 34(2), 675-689, 2014.

Biémont, C. and Vieira, C., “Genetics: junk DNA as an evolutionary force”, *Nature* 443(7111), 521, 2006.

Bolger, A.M., Lohse, M. and Usadel, B., “Trimmomatic: a flexible trimmer for Illumina sequence data”, *Bioinformatics* 30(15), 2114-2120, 2014.

Bombarely, A., Moser, M., Amrad, A., Bao, M., Bapaume, L., Barry, C.S., Bliiek, M., Boersma, M.R., Borghi, L., Bruggmann, R. and Bucher, M., “Insight into the evolution of the Solanaceae from the parental genomes of *Petunia* hybrid”, *Nature Plants* 2(6), 16074, 2016.

Bousios, A., Darzentas, N., Tsaftaris, A. and Pearce, S.R., “Highly conserved motifs in non-coding regions of Sirevirus retrotransposons: the key for their pattern of distribution within and across plants?”, *BMC Genomics* 11(1), 89, 2010.

Bousios, A., Kourmpetis, Y.A., Pavlidis, P., Minga, E., Tsaftaris, A. and Darzentas, N., “The turbulent life of Sirevirus retrotransposons and the evolution of the maize genome: more than ten thousand elements tell the story”, *The Plant Journal* 69(3), 475-488, 2012.

Brevis, P.A., Bassil, N.V., Ballington, J.R. and Hancock, J.F., “Impact of wide hybridization on highbush blueberry breeding”, *Journal of the American Society for Horticultural Science* 133(3), 427-437, 2008.

- Britten, R.J. and Kohne, D.E., “Repeated sequences in DNA”, *Science* 161(3841), 529-540, 1968.
- Brown, A., Raja, A., Reid, R., Wright, G., Brouwer, C., Main, D., and Burke, M., “The genomic sequencing of diploid blueberry (*Vaccinium corymbosum*)”, *ASHS Annual Conference September*, Waikoloa, Hawaii, 25–28, (2011).
- Camp, W.H., “On the structure of populations in the genus *Vaccinium*”, *Brittonia* 4(2), 189-204, 1942.
- Capy, P. and Maisonhaute, C., “Acquisition/loss of modules: the construction set of transposable elements”, *Russian Journal of Genetics* 38(6), 594-601, 2002.
- Çelik, H., “Yüksek boylu maviyemiş (highbush blueberry) yetiştiriciliği”, *Gifimey Mesleki Yayınlar Serisi-III*, 152, 2012.
- Česonienė, L., Daubaras, R., Paulauskas, A., Žukauskienė, J. and Zych, M., “Morphological and genetic diversity of European cranberry (*Vaccinium oxycoccos* L., Ericaceae) clones in Lithuanian reserves”, *Acta Societatis Botanicorum Poloniae* 82(3), 211, 2013.
- Chavez, D.J. and Lyrene, P.M., “Hybridization of two diploid *Vaccinium* section *Cyanococcus* species with diploid *Vaccinium arboreum* in section *Batodendron*”, *Euphytica* 171(2), 263, 2010.
- Chuong, E.B., Elde, N.C. and Feschotte, C., “Regulatory activities of transposable elements: from conflicts to benefits”, *Nature Reviews Genetics* 18(2), 71, 2017.
- Churikov, D. and Price, C.M., “Telomeric and subtelomeric repeat sequences”, *eLS*, 2008.
- Costich, D.E., Ortiz, R., Meagher, T.R., Bruederle, L.P. and Vorsa, N., “Determination of ploidy level and nuclear DNA content in blueberry by flow cytometry”, *Theoretical and Applied Genetics* 86(8), 1001-1006, 1993.

Davidson, E.H. and Britten, R.J., "Regulation of gene expression: possible role of repetitive sequences", *Science* 204(4397), 1052-1059, 1979.

Darnell, R.L. and Williamson, J.G., "Feasibility of blueberry production in warm climates", *VI International Symposium on Vaccinium Culture 446*, pp. 251-256, August, 1996.

De Lange, T., "T-loops and the origin of telomeres", *Nature Reviews Molecular Cell Biology* 5(4), 323, 2004.

Dhanaraj, A.L., Alkharouf, N.W., Beard, H.S., Chouikha, I.B., Matthews, B.F., Wei, H., Arora, R. and Rowland, L.J., "Major differences observed in transcript profiles of blueberry during cold acclimation under field and cold room conditions", *Planta* 225(3), 735-751, 2007.

Dhanaraj, A.L., Slovin, J.P. and Rowland, L.J., "Analysis of gene expression associated with cold acclimation in blueberry floral buds using expressed sequence tags", *Plant Science* 166(4), 863-872, 2004.

Die, J.V. and Rowland, L.J., "Advent of genomics in blueberry", *Molecular Breeding* 32(3), 493-504, 2013.

Domingues, D.S., Cruz, G.M., Metcalfe, C.J., Nogueira, F.T., Vicentini, R., de S Alves, C. and Van Sluys, M.A., "Analysis of plant LTR-retrotransposons at the fine-scale family level reveals individual molecular patterns", *BMC Genomics* 13(1), 137, 2012.

Dodsworth, S., "Genome skimming for next-generation biodiversity analysis", *Trends in Plant Science* 20(9), 525-527, 2015.

Dodsworth, S., Chase, M.W., Kelly, L.J., Leitch, I.J., Macas, J., Novák, P., Piednoël, M., Weiss-Schneeweiss, H. and Leitch, A.R., "Genomic repeat abundances contain phylogenetic signal", *Systematic Biology* 64(1), 112-126, 2014.

Dodsworth, S., Jang, T.S., Struebig, M., Chase, M.W., Weiss-Schneeweiss, H. and Leitch, A.R., “Genome-wide repeat dynamics reflect phylogenetic distance in closely related allotetraploid *Nicotiana* (Solanaceae)”, *Plant Systematics and Evolution* 303(8), 1013-1020, 2017.

Ehlenfeldt, M.K. and Ballington, J.R., “*Vaccinium* species of section *Hemimyrtillus*: their value to cultivated blueberry and approaches to utilization”, *Botany* 90(5), 347-353, 2012.

Eickbush, T.H. and Jamburuthugoda, V.K., “The diversity of retrotransposons and the properties of their reverse transcriptases”, *Virus Research* 134(1-2), 221-234, 2008.

Fajardo, D., Senalik, D., Ames, M., Zhu, H., Steffan, S.A., Harbut, R., Polashock, J., Vorsa, N., Gillespie, E., Kron, K. and Zalapa, J.E., “Complete plastid genome sequence of *Vaccinium macrocarpon*: structure, gene content, and rearrangements revealed by next generation sequencing”, *Tree Genetics and Genomes* 9(2), 489-498, 2013.

Fajkus, J. and Zentgraf, U., Structure and maintenance of chromosome ends in plants, *Telomeres and Telomerases: Cancer and Biology*, *Landes Bioscience*, 314-331, 2002.

FAOSTAT, F., Statistical data, *Food and Agriculture Organization of the United Nations, Rome*, 2017.

Feschotte, C., Jiang, N. and Wessler, S.R., “Plant transposable elements: where genetics meets genomics”, *Nature Reviews Genetics* 3(5), 329, 2002.

Folta, K.M. and Kole, C. eds., Genetics, genomics and breeding of berries, *CRC Press*, 2016.

Garcia, S. and Kovařík, A., “Dancing together and separate again: gymnosperms exhibit frequent changes of fundamental 5S and 35S rRNA gene (rDNA) organization”, *Heredity* 111(1), 23, 2013.

Garcia, S., Lim, K.Y., Chester, M., Garnatje, T., Pellicer, J., Vallès, J., Leitch, A.R. and Kovařík, A., “Linkage of 35S and 5S rRNA genes in *Artemisia* (family Asteraceae): first evidence from angiosperms”, *Chromosoma* 118(1), 85, 2009.

Garrido-Ramos, M.A., “Satellite DNA in plants: more than just rubbish”, *Cytogenetic and Genome Research* 146(2), 153-170, 2015.

Garrido-Ramos, M.A., “Satellite DNA: an evolving topic”, *Genes* 8(9), 230, 2017.

Galindo-González, L., Mhiri, C., Deyholos, M.K. and Grandbastien, M.A., “LTR-retrotransposons in plants: engines of evolution”, *Gene* 626, 14-25, 2017.

Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W. and Gascuel, O., “New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0”, *Systematic Biology* 59(3), 307-321, 2010.

Gupta, V., Estrada, A.D., Blakley, I., Reid, R., Patel, K., Meyer, M.D., Andersen, S.U., Brown, A.F., Lila, M.A. and Loraine, A.E., “RNA-Seq analysis and annotation of a draft blueberry genome assembly identifies candidate genes involved in fruit ripening, biosynthesis of bioactive compounds, and stage-specific alternative splicing”, *GigaScience* 4(1), 5, 2015.

Haghighi, K. and Hancock, J.F., “DNA restriction fragment length variability in the genomes of highbush blueberry”, *HortScience* 27(1), 44-47, 1992.

Hancock, J.F., Lyrene, P., Finn, C.E., Vorsa, N. and Lobos, G.A., Blueberries and cranberries. In Temperate fruit crop breeding, *Springer*, Dordrecht, 115-150, 2008.

He, Q., Cai, Z., Hu, T., Liu, H., Bao, C., Mao, W. and Jin, W., “Repetitive sequence analysis and karyotyping reveals centromere-associated DNA sequences in radish (*Raphanus sativus* L.)”, *BMC Plant Biology* 15(1), 105, 2015.

Heitkam, T., Petrasch, S., Zakrzewski, F., Kögler, A., Wenke, T., Wanke, S. and Schmidt, T., “Next-generation sequencing reveals differentially amplified tandem repeats as a major genome component of Northern Europe’s oldest *Camellia japonica*”, *Chromosome Research* 23(4), 791-806, 2015.

Heslop-Harrison, J.P. and Schwarzacher, T., “Nucleosomes and centromeric DNA packaging”, *PNAS*, 19974–19975, 2013.

Heslop-Harrison, J.S. and Schwarzacher, T., “Organisation of the plant genome in chromosomes”, *The Plant Journal* 66(1), 18-33, 2011.

Heslop-Harrison, J.S., “The molecular cytogenetics of plants”, *Journal of Cell Science* 100(1), 15-21, 1991.

Hou, D.X., “Potential mechanisms of cancer chemoprevention by anthocyanins”, *Current Molecular Medicine* 3(2), 149-159, 2003.

Hummer, K.E., Bassil, N.V., Rodríguez Armenta, H.P. and Olmstead, J.W., “August. *Vaccinium* species ploidy assessment”, In *XXIX International Horticultural Congress on Horticulture: Sustaining Lives, Livelihoods and Landscapes (IHC2014): IV 1101*, pp. 199-204, 2014.

International Barley Genome Sequencing Consortium, “A physical, genetic and functional sequence assembly of the barley genome”, *Nature* 491(7426), 711, 2012.

International Rice Genome Sequencing Project, “The map-based sequence of the rice genome”, *Nature* 436, 793-800, 2005

International Wheat Genome Sequencing Consortium, “A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome”, *Science* 345(6194), 1251788, 2014.

Ito, H., “Small RNAs and transposon silencing in plants”, *Development, Growth and Differentiation* 54(1), 100-107, 2012.

Iwata, A., Tek, A.L., Richard, M., Abernathy, B., Fonsêca, A., Schmutz, J., Chen, N.W., Thureau, V., Magdelenat, G., Li, Y. and Murata, M., “Identification and characterization of functional centromeres of the common bean”, *The Plant Journal* 76(1), 47-60, 2013.

Jagannathan, M., Cummings, R. and Yamashita, Y.M., “A conserved function for pericentromeric satellite DNA”, *eLife* 7, e34122, 2018.

Kamm, A., Galasso, I., Schmidt, T. and Heslop-Harrison, J.S., “Analysis of a repetitive DNA family from *Arabidopsis arenosa* and relationships between *Arabidopsis* species”, *Plant Molecular Biology* 27(5), 853-862, 1995.

Katoh, K. and Standley, D.M., “MAFFT multiple sequence alignment software version 7: improvements in performance and usability”, *Molecular Biology and Evolution* 30(4), 772-780, 2013.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C. and Thierer, T., “Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data”, *Bioinformatics* 28(12), 1647-1649, 2012.

Kelly, L.J., Renny-Byfield, S., Pellicer, J., Macas, J., Novák, P., Neumann, P., Lysak, M.A., Day, P.D., Berger, M., Fay, M.F. and Nichols, R.A., “Analysis of the giant genomes of *Fritillaria* (Liliaceae) indicates that a lack of DNA removal characterizes extreme expansions in genome size”, *New Phytologist* 208(2), 596-607, 2015.

Kirov, I.V., Kiseleva, A.V., Van Laere, K., Van Roy, N. and Khrustaleva, L.I., “Tandem repeats of *Allium fistulosum* associated with major chromosomal landmarks”, *Molecular Genetics and Genomics* 292(2), 453-464, 2017.

Kloet, S.V., “The taxonomy of the highbush blueberry, *Vaccinium corymbosum*”, *Canadian Journal of Botany* 58(10), 1187-1201, 1980.

Kolano, B., Bednara, E. and Weiss-Schneeweiss, H., “Isolation and characterization of reverse transcriptase fragments of LTR retrotransposons from the genome of *Chenopodium quinoa* (Amaranthaceae)”, *Plant Cell Reports* 32(10), 1575-1588, 2013.

Kron, K.A., Powell, E.A. and Luteyn, J.L., “Phylogenetic relationships within the blueberry tribe (*Vaccinieae*, *Ericaceae*) based on sequence data from matK and nuclear ribosomal ITS regions, with comments on the placement of *Satyria*”, *American Journal of Botany* 89(2), 327-336, 2002.

Kubis, S., Schmidt, T. and Heslop-Harrison, J.S., “Repetitive DNA elements as a major component of plant genomes”, *Annals of Botany* 82(suppl_1), 45-55, 1998.

Kubis, S.E., Heslop-Harrison, J.S., Desel, C. and Schmidt, T., “The genomic organization of non-LTR retrotransposons (LINEs) from three *Beta* species and five other angiosperms”, *Plant Molecular Biology* 36(6), 821-831, 1998.

Levi, A. and Rowland, L.J., “Identifying blueberry cultivars and evaluating their genetic relationships using randomly amplified polymorphic DNA (RAPD) and simple sequence repeat-(SSR-) anchored primers”, *Journal of the American Society for Horticultural Science* 122(1), 74-78, 1997.

Li, L., Zhang, H., Liu, Z., Cui, X., Zhang, T., Li, Y. and Zhang, L., “Comparative transcriptome sequencing and de novo analysis of *Vaccinium corymbosum* during fruit and color development”, *BMC Plant Biology* 16(1), 223, 2016.

Li, W., and Godzik, A., “Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences”, *Bioinformatics* 22(13), 1658-1659, 2006.

Llorens, C., Futami, R., Covelli, L., Domínguez-Escribá, L., Viu, J.M., Tamarit, D., Aguilar-Rodríguez, J., Vicente-Ripolles, M., Fuster, G., Bernet, G.P. and Maumus, F., “The Gypsy Database (GyDB) of mobile genetic elements: release 2.0”, *Nucleic Acids Research* 39(suppl_1), D70-D74, 2010.

Lobos, G.A. and Hancock, J.F., “Breeding blueberries for a changing global environment: a review”, *Frontiers in Plant Science* 6, 782, 2015.

Lyrene, P., “Breeding southern highbush blueberries”, *Plant Breeding Reviews* 30, 353-414, 2008.

Macas, J., Novak, P., Pellicer, J., Čížková, J., Koblížková, A., Neumann, P., Fukova, I., Doležel, J., Kelly, L.J. and Leitch, I.J., “In depth characterization of repetitive DNA in 23 plant genomes reveals sources of genome size variation in the legume tribe Fabeae”, *PLoS One* 10(11), e0143424, 2015.

Macas, J., Navrátilová, A. and Koblížková, A., “Sequence homogenization and chromosomal localization of VicTR-B satellites differ between closely related *Vicia* species”, *Chromosoma* 115(6), 437-447, 2006.

Macas, J., Meszaros, T. and Nouzova, M., “PlantSat: a specialized database for plant satellite repeats”, *Bioinformatics* 18(1), 28-35, 2002.

Marín, I. and Lloréns, C., “Ty3/Gypsy retrotransposons: description of new *Arabidopsis thaliana* elements and evolutionary perspectives derived from comparative genomic data”, *Molecular Biology and Evolution* 17(7), 1040-1049, 2000.

Mariotti, B., Manzano, S., Kejnovský, E., Vyskot, B. and Jamilena, M., “Accumulation of Y-specific satellite DNAs during the evolution of *Rumex acetosa* sex chromosomes”, *Molecular Genetics and Genomics* 281(3), 249, 2009.

Mehrotra, S. and Goyal, V., “Repetitive sequences in plant nuclear DNA: types, distribution, evolution and function”, *Genomics, Proteomics and Bioinformatics* 12(4), 164-171, 2014.

Melters, D.P., Bradnam, K.R., Young, H.A., Telis, N., May, M.R., Ruby, J.G., Sebra, R., Peluso, P., Eid, J., Rank, D. and Garcia, J.F., “Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution”, *Genome Biology* 14(1), R10, 2013.

Menzel, G., Dechyeva, D., Wenke, T., Holtgräwe, D., Weisshaar, B. and Schmidt, T., “Diversity of a complex centromeric satellite and molecular characterization of dispersed sequence families in sugar beet (*Beta vulgaris*)”, *Annals of Botany* 102(4), 521-530, 2008.

Meštrović, N., Mravinac, B., Pavlek, M., Vojvoda-Zeljko, T., Šatović, E. and Plohl, M., “Structural and functional liaisons between transposable elements and satellite DNAs”, *Chromosome Research* 23(3), 583-596, 2015.

Meyers, B.C., Tingey, S.V. and Morgante, M., “Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome”, *Genome Research* 11(10), 1660-1676, 2001.

Miga, K.H., “Completing the human genome: the progress and challenge of satellite DNA assembly”, *Chromosome Research* 23(3), 421-426, 2015.

Mizuno, H., Wu, J., Kanamori, H., Fujisawa, M., Namiki, N., Saji, S., Katagiri, S., Katayose, Y., Sasaki, T. and Matsumoto, T., “Sequencing and characterization of telomere and subtelomere regions on rice chromosomes 1S, 2S, 2L, 6L, 7S, 7L and 8S”, *The Plant Journal* 46(2), 206-217, 2006.

Moyer, R.A., Hummer, K.E., Finn, C.E., Frei, B. and Wrolstad, R.E., “Anthocyanins, phenolics, and antioxidant capacity in diverse small fruits: *Vaccinium*, *Rubus*, and *Ribes*”, *Journal of Agricultural and Food Chemistry* 50(3), 519-525, 2002.

Mudd, A.B., White, E.J., Bolloskis, M.P., Kapur, N.P., Everhart, K.W., Lin, Y.C., Bussler, W.W., Reid, R.W. and Brown, R.H., “Students’ perspective on genomics: from sample to sequence using the case study of blueberry”, *Frontiers in Genetics* 4, 245, 2013.

- Muñoz-López, M. and García-Pérez, J.L., “DNA transposons: nature and applications in genomics”, *Current Genomics* 11(2), 115-128, 2010.
- Nakamura, T.M. and Cech, T.R., “Reversing time: origin of telomerase”, *Cell* 92(5), 587-590, 1998.
- Natali, L., Cossu, R.M., Mascagni, F., Giordani, T. and Cavallini, A., “A survey of *Gypsy* and *Copia* LTR-retrotransposon superfamilies and lineages and their distinct dynamics in the *Populus trichocarpa* (L.) genome”, *Tree Genetics and Genomes* 11(5), 107, 2015.
- Neumann, P., Navrátilová, A., Koblížková, A., Kejnovský, E., Hříbová, E., Hobza, R., Widmer, A., Doležel, J. and Macas, J., “Plant centromeric retrotransposons: a structural and cytogenetic perspective”, *Mobile DNA* 2(1), 4, 2011.
- Nickavar, B. and Amin, G., “Anthocyanins from *Vaccinium arctostaphylos* berries”, *Pharmaceutical Biology* 42(4-5), 289-291, 2004.
- Novák, P., Ávila Robledillo, L., Koblížková, A., Vrbová, I., Neumann, P. and Macas, J., “TAREAN: A computational tool for identification and characterization of satellite DNA from unassembled short reads”, *Nucleic Acids Research* 257, 2017.
- Novák, P., Neumann, P., Pech, J., Steinhaisl, J. and Macas, J., “RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads”, *Bioinformatics* 29(6), 792-793, 2013.
- Oliver, K.R., McComb, J.A. and Greene, W.K., “Transposable elements: powerful contributors to angiosperm evolution and diversity”, *Genome Biology and Evolution* 5(10), 1886-1901, 2013.
- Paterson, A.H., Bowers, J.E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G., Hellsten, U., Mitros, T., Poliakov, A. and Schmutz, J., “The *Sorghum bicolor* genome and the diversification of grasses”, *Nature* 457(7229), 551, 2009.

- Pellicer, J., Hidalgo, O., Dodsworth, S. and Leitch, I.J., “Genome size diversity and its impact on the evolution of land plants”, *Genes* 9(2), 88, 2018.
- Piégu, B., Bire, S., Arensburger, P. and Bigot, Y., “A survey of transposable element classification systems—a call for a fundamental update to meet the challenge of their diversity and complexity”, *Molecular Phylogenetics and Evolution* 86, 90-109, 2015.
- Plohl, M., Meštrović, N. and Mravinac, B., “Satellite DNA evolution”, *Genome Dyn* 7,126–152, 2012.
- Plohl, M., “Those mysterious sequences of satellite DNAs”, *Periodicum Biologorum* 112(4), 403-410, 2010.
- Polashock, J., Zelzion, E., Fajardo, D., Zalapa, J., Georgi, L., Bhattacharya, D. and Vorsa, N., “The American cranberry: first insights into the whole genome of a species adapted to bog habitat”, *BMC Plant Biology* 14(1), 165, 2014.
- Powell, E.A. and Kron, K.A., “Hawaiian blueberries and their relatives—a phylogenetic analysis of *Vaccinium* sections *Macropelma*, *Myrtillus*, and *Hemimyrtilus* (Ericaceae)”, *Systematic Botany* 27(4), 768-779, 2002.
- Pray, L.A., “Transposons: The jumping genes”, *Nature Education* 1(1), 204, 2008.
- Price, M.N., Dehal, P.S. and Arkin, A.P., “FastTree 2—approximately maximum-likelihood trees for large alignments”, *PloS one* 5(3), e9490, 2010.
- Puterova, J., Razumova, O., Martinek, T., Alexandrov, O., Divashuk, M., Kubat, Z., Hobza, R., Karlov, G. and Kejnovsky, E., “Satellite DNA and transposable elements in seabuckthorn (*Hippophae rhamnoides*), a dioecious plant with small Y and large X chromosomes”, *Genome Biology and Evolution* 9(1), 197-212, 2017.
- Qu, L. and Hancock, J.F., “Nature of 2n gamete formation and mode of inheritance in interspecific hybrids of diploid *Vaccinium darrowi* and tetraploid *V. corymbosum*”, *Theoretical and Applied Genetics* 91(8), 1309-1315, 1995.

Qu, L. and Vorsa, N., “Desynapsis and spindle abnormalities leading to 2 n pollen formation in *Vaccinium darrowii*”, *Genome* 42(1), 35-40, 1999.

Retamales, J.B. and Hancock, J.F., Blueberries (Vol. 21), *CABI*, 2012.

Rice, P., Longden, I. and Bleasby, A., “EMBOSS: the European molecular biology open software suite”, *Trends in Genetics* 16(6), 276-277, 2000.

Rosato, M., Kovařík, A., Garilleti, R. and Rosselló, J.A., “Conserved organisation of 45S rDNA sites and rDNA gene copy number among major clades of early land plants”, *PloS one* 11(9), e0162544, 2016.

Robledillo, L.Á., Koblížková, A., Novák, P., Böttinger, K., Vrbová, I., Neumann, P., Schubert, I. and Macas, J., “Satellite DNA in *Vicia faba* is characterized by remarkable diversity in its sequence composition, association with centromeres, and replication timing”, *Scientific Reports* 8(1), 5838, 2018.

Rowland, L.J., Alkharouf, N., Darwish, O., Ogden, E.L., Polashock, J.J., Bassil, N.V. and Main, D., “Generation and analysis of blueberry transcriptome sequences from leaves, developing fruit, and flower buds from cold acclimation through deacclimation”, *BMC Plant Biology* 12(1), 46, 2012.

Rowland, L.J., Bell, D.J., Alkharouf, N., Bassil, N.V., Drummond, F.A., Beers, L., Buck, E.J., Finn, C.E., Graham, J., McCallum, S. and Hancock, J.F., “Generating genomic tools for blueberry improvement”, *International Journal of Fruit Science* 12(1-3), 276-287, 2012.

Rowland, L.J., Ogden, E.L. and Ehlenfeldt, M.K., “EST-PCR markers developed for highbush blueberry are also useful for genetic fingerprinting and relationship studies in rabbiteye blueberry”, *Scientia Horticulturae* 125(4), 779-784, 2010.

Ruiz-Ruano, F.J., López-León, M.D., Cabrero, J. and Camacho, J.P.M., “High-throughput analysis of the satellitome illuminates satellite DNA evolution”, *Scientific Reports* 6, 28333, 2016.

Singer, M.F., “Highly repeated sequences in mammalian genomes”, *International Review of Cytology* 76, 67-112, 1982.

Schlautman, B., Covarrubias-Pazaran, G., Diaz-Garcia, L., Iorizzo, M., Polashock, J., Grygleski, E., Vorsa, N. and Zalapa, J., “Construction of a high-density American cranberry (*Vaccinium macrocarpon* Ait.) composite map using genotyping-by-sequencing for multi-pedigree linkage mapping”, *G3: Genes, Genomes, Genetics* 7(4), 1177-1189, 2017.

Schmidt, T. and Heslop-Harrison, J.S., “Genomes, genes and junk: the large-scale organization of plant chromosomes”, *Trends in Plant Science* 3(5), 195-199, 1998.

Schmidt, T., “LINEs, SINEs and repetitive DNA: non-LTR retrotransposons in plant genomes”, *Plant Molecular Biology* 40(6), 903-910, 1999.

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T.A. and Minx, P., “The B73 maize genome: complexity, diversity, and dynamics”, *Science* 326(5956), 1112-1115, 2009.

Sharma, S. and Raina, S.N., “Organization and evolution of highly repeated satellite DNA sequences in plant chromosomes”, *Cytogenetic and Genome Research* 109(1-3), 15-26, 2005.

Sharpe, R.H. and Darrow, G.M., “Breeding blueberries for the Florida climate”, *Florida State Horticultural Society* 72, 308-311, 1959.

Shapiro, J.A. and Von Sternberg, R., “Why repetitive DNA is essential to genome function”, *Biological Reviews* 80(2), pp.227-250, 2005.

Skaletsky, H., Kuroda-Kawaguchi, T., Minx, P.J., Cordum, H.S., Hillier, L., Brown, L.G., Repping, S., Pyntikova, T., Ali, J., Bieri, T. and Chinwalla, A., “The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes”, *Nature* 423(6942), 825, 2003.

Spiers, J.M., Gupton, C.L. and Draper, A.D., “'Jubilee', 'Magnolia', and 'Pearl River' southern highbush blueberries”, *In VI International Symposium on Vaccinium Culture* 446, 155-158, August, 1996.

Steflova, P., Tokan, V., Vogel, I., Lexa, M., Macas, J., Novak, P., Hobza, R., Vyskot, B. and Kejnovsky, E., “Contrasting patterns of transposable element and satellite distribution on sex chromosomes (XY1Y2) in the dioecious plant *Rumex acetosa*”, *Genome Biology and Evolution* 5(4), 769-782, 2013.

Sultana, N., Serçe, S., Menzel, G., Heitkam, T. and Schmidt, T., “Comparative analysis of repetitive sequences reveals genome differences between two common cultivated *Vaccinium* species (*V. corymbosum* and *V. macrocarpon*)”, *Journal of Molecular Biology and Biotechnology* 1(2), 7-15, 2017

Smith, T.W., Walinga, C., Wang, S., Kron, P., Suda, J. and Zalapa, J., “Evaluating the relationship between diploid and tetraploid *Vaccinium oxycoccos* (Ericaceae) in eastern Canada”, *Botany* 93(10), 623-636, 2015.

Smit, A.F.A., Hubley, R., and Green, P., RepeatMasker Open-4.0. 2013-2015 <<http://www.repeatmasker.org>>.

Swofford, D.L., PAUP*: phylogenetic analysis using parsimony (* and other methods), *Sunderland MA*, 2002.

Tek, A.L., Song, J., Macas, J. and Jiang, J., “Sobo, a recently amplified satellite repeat of potato, and its implications for the origin of tandemly repeated sequences”, *Genetics* 170(3), 1231-1238, 2005.

Torres, G.A., Gong, Z., Iovene, M., Hirsch, C.D., Buell, C.R., Bryan, G.J., Novák, P., Macas, J. and Jiang, J., “Organization and evolution of subtelomeric satellite repeats in the potato genome”, *G3: Genes, Genomes, Genetics* 1(2), 85-92, 2011.

Treangen, T.J. and Salzberg, S.L., “Repetitive DNA and next-generation sequencing: computational challenges and solutions”, *Nature Reviews Genetics* 13(1), 36, 2012.

Trehane, J., Blueberries, Cranberries and other *Vacciniums*, *Royal Horticultural Society, Plant Collector Guide* 19, 2004.

Ugarkovic, D., “Functional elements residing within satellite DNAs”, *EMBO Reports* 6(11), 1035-1039, 2005.

Ustyantsev, K., Blinov, A. and Smyshlyaev, G., “Convergence of retrotransposons in oomycetes and plants”, *Mobile DNA* 8(1), 4, 2017.

Ustyantsev, K., Novikova, O., Blinov, A. and Smyshlyaev, G., “Convergent evolution of ribonuclease H in LTR retrotransposons and retroviruses”, *Molecular Biology and Evolution* 32(5), 1197-1207, 2015.

Vander Kloet, S.P. and Dickinson, T.A., “A subgeneric classification of the genus *Vaccinium* and the metamorphosis of *V.* section *Bracteata* Nakai: more terrestrial and less epiphytic in habit, more continental and less insular in distribution”, *Journal of Plant Research* 122(3), 253-268, 2009.

Vander Kloet, S.P., The genus *Vaccinium* in North America (No. 1828), *Agriculture Canada*, 1988.

Vicient, C.M. and Casacuberta, J.M., “Impact of transposable elements on polyploid plant genomes”, *Annals of Botany* 120(2), 195-207, 2017.

Van de Peer, Y., Mizrachi, E. and Marchal, K., “The evolutionary significance of polyploidy”, *Nature Reviews Genetics* 18(7), 411, 2017.

Wang, L.J., Su, S., Wu, J., Du, H., Li, S.S., Huo, J.W., Zhang, Y. and Wang, L.S., “Variation of anthocyanins and flavonols in *Vaccinium uliginosum* berry in Lesser Khingan Mountains and its antioxidant activity”, ***Food Chemistry*** 160, 357-364, 2014.

Waring, M. and Britten, R.J., “Nucleotide sequence repetition: a rapidly reassociating fraction of mouse DNA” ***Science*** 154(3750), 791-794, 1966.

Wang, M., Yu, Y., Haberer, G., Marri, P.R., Fan, C., Goicoechea, J.L., Zuccolo, A., Song, X., Kudrna, D., Ammiraju, J.S. and Cossu, R.M., “The genome sequence of African rice (*Oryza glaberrima*) and evidence for independent domestication”, ***Nature Genetics*** 46(9), 982, 2014.

Wendel, J.F., Jackson, S.A., Meyers, B.C. and Wing, R.A., “Evolution of plant genome architecture”, ***Genome Biology*** 17(1), 37, 2016.

Weber, B., Wenke, T., Frömmel, U., Schmidt, T. and Heitkam, T., “The Ty1-copia families SALIRE and Cotzilla populating the *Beta vulgaris* genome show remarkable differences in abundance, chromosomal distribution, and age”, ***Chromosome Research*** 18(2), 247-263, 2010.

Weber, B., Heitkam, T., Holtgräwe, D., Weisshaar, B., Minoche, A.E., Dohm, J.C., Himmelbauer, H. and Schmidt, T., “Highly diverse chromoviruses of *Beta vulgaris* are classified by chromodomains and chromosomal integration”, ***Mobile DNA*** 4(1), 8, 2013.

Wicker, T. and Keller, B., “Genome-wide comparative analysis of copia retrotransposons in Triticeae, rice, and *Arabidopsis* reveals conserved ancient evolutionary lineages and distinct dynamics of individual copia families”, ***Genome Research*** 17(7), 1072-1081, 2007.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O. and Paux, E., “A unified classification system for eukaryotic transposable elements”, ***Nature Reviews Genetics*** 8(12), 973, 2007.

Wollrab, C., Heitkam, T., Holtgräwe, D., Weisshaar, B., Minoche, A.E., Dohm, J.C., Himmelbauer, H. and Schmidt, T., “Evolutionary reshuffling in the Errantivirus lineage Elbe within the *Beta vulgaris* genome”, *The Plant Journal* 72(4), 636-651, 2012.

Xu, Z. and Wang, H., “LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons”, *Nucleic Acids Research* 35(suppl_2), W265-W268, 2007.

Xue, S. and Barna, M., “Specialized ribosomes: a new frontier in gene regulation and organismal biology”, *Nature Reviews Molecular Cell Biology* 13(6), 355, 2012.

Yin, H., Du, J., Wu, J., Wei, S., Xu, Y., Tao, S., Wu, J. and Zhang, S., “Genome-wide annotation and comparative analysis of long terminal repeat retrotransposons between pear species of *P. bretschneideri* and *P. communis*”, *Scientific Reports* 5, 17644, 2015.

Yang, X., Zhao, H., Zhang, T., Zeng, Z., Zhang, P., Zhu, B., Han, Y., Braz, G.T., Casler, M.D., Schmutz, J. and Jiang, J., “Amplification and adaptation of centromeric repeats in polyploid switchgrass species”, *New Phytologist* 218(4), 1645-1657, 2018.

Zakrzewski, F., Weber, B. and Schmidt, T., “A molecular cytogenetic analysis of the structure, evolution, and epigenetic modifications of major DNA sequences in centromeres of *Beta* species”, *Plant Centromere Biology*, 39-55, 2013.

Zakrzewski, F., Wenke, T., Holtgräwe, D., Weisshaar, B. and Schmidt, T., “Analysis of ac 0 t-1 library enables the targeted identification of minisatellite and satellite families in *Beta vulgaris*”, *BMC Plant Biology* 10(1), 8, 2010.

Zdepski, A., Debnath, S.C., Howell, A., Polashock, J., Oudemans, P., Vorsa, N. and Michael, T.P., Cranberry. In: Genetics, genomics and breeding of berries, Folta, K., Kole, C., editor. Boca Raton, F.L., *CRC Press* 200, 2011

Zhang, G., Liu, X., Quan, Z., Cheng, S., Xu, X., Pan, S., Xie, M., Zeng, P., Yue, Z., Wang, W. and Tao, Y., “Genome sequence of foxtail millet (*Setaria italica*) provides insights into grass evolution and biofuel potential”, *Nature Biotechnology* 30(6), 549, 2012.

Zhang, H., Koblížková, A., Wang, K., Gong, Z., Oliveira, L., Torres, G.A., Wu, Y., Zhang, W., Novák, P., Buell, C.R. and Macas, J., “Boom-bust turnovers of megabase-sized centromeric DNA in *Solanum* species: rapid evolution of DNA sequences associated with centromeres”, *The Plant Cell* 26(4), 1436-1447, 2014.

Zoratti, L., Palmieri, L., Jaakola, L. and Häggman, H., “Genetic diversity and population structure of an important wild berry crop”, *AoB Plants* 7, 2015



CURRICULUM VITAE

Nusrat Sultana was born on 12 January 1989 in Barisal, Bangladesh. She completed her higher secondary education from Dhaka, Bangladesh. She graduated from Botany Department, University of Dhaka, Bangladesh in February 2011. She accomplished M. Sc. specialized in “Plant Cytogenetics” from the same Department, (branch Plant Biotechnology) in March 2013. From 2013-2014 she worked as a project assistant and Lecturer in Cytogenetics laboratory, Botany Department, University of Dhaka and Jagannath University respectively. She started her doctoral studies in Agricultural Genetic Engineering Department, Niğde Ömer Halisdemir University in September 2014. Since then she has been continuing her Ph.D. studies with a research interest in plant genomics.

PUBLICATION PRODUCED FROM THIS THESIS WORK

Sultana, N., Serçe, S., Menzel, G., Heitkam, T. and Schmidt, T., “Comparative analysis of repetitive sequences reveals genome differences between two common cultivated *Vaccinium* species (*V. corymbosum* and *V. macrocarpon*)”, *Journal of Molecular Biology and Biotechnology* 1(2), 7-15, 2017

