Genetic diversity and evolutionary relationships of *Sugarcane mosaic virus* variants

` **BY**

Hafiza Noor ul ain

MS (Plant Biotechnology)

00000171157

Supervisor

Dr. Muhammad Tahir

HoD Plant Biotechnology

Atta-Ur-Rahman School of Applied Biosciences (ASAB),

National University of Sciences & Technology (NUST),

Islamabad, Pakistan 2018

Genetic diversity and evolutionary relationships of *Sugarcane mosaic virus* variants

BY

Hafiza Noor ul ain

00000171157

A thesis submitted to partial fulfillment of the requirement for the degree of Master of Sciences

In

Plant Biotechnology

Atta-Ur-Rahman School of Applied Biosciences (ASAB),

National University of Sciences & Technology (NUST),

Islamabad, Pakistan

2018

FORM TH-4

MS THESIS WORK

We hereby recommend that the dissertation prepared under our supervision by: **Hafiza Noor ul ain, Registration No. 00000171157**, **Titled**: **Genetic diversity and evolutionary relationships of *Sugarcane mosaic virus* variants** be accepted in partial fulfillment of the requirements for the award of **MS (Plant Biotechnology)** degree with (_____grade).

Examination Committee Members

1. Name: **Dr. Najam Us Sahar Sadaf Zaidi** Signature:_____

2. Name: **Dr. Faiza Munir** Signature:_____

Supervisor**: Dr. Muhammad Tahir** Signature:_____

Date:_____

Supervisor**: Dr. Muhammad Qasim Hayat** Signature:_____

Date:_____

COUNTERSINGED

Date:_____ _____

Dean/Principal

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS/MPhil thesis written by Ms Hafiza Noor ul ain (Registration No. 00000171157, of Atta-ur-Rehman School of Applied Biosciences, NUST has been vetted by undersigned, found complete on all respects as per NUST Statute/Regulations, is free of plagiarism, errors and mistakes and is accepted as partial fulfillment for award of MS/MPhil degree.

Signature: _____

Name of Supervisor: Dr. Muhammad Tahir

Date: _____

Signature (HOD): _____

Date: _____

Signature (Dean/Principal): _____

Date: _____

Dedicated to my wonderful parents

& to my respectful teachers

ACKNOWLEDGEMENTS

My profound gratitude to Almighty ALLAH for granting me with health and opportunity to accomplish this work and for always blessing me with all that I have ever wished for. My humble gratitude for our beloved Prophet Muhammad (P.B.U.H) who has always been an inspiration to me, of being a perfect human being.

I would like to take this opportunity to express gratitude to my supervisor Dr. Muhammad Tahir, for his invaluable guidance, encouragement and support through my research time.

I am thankful to Dr. Muhammad Qasim Hayat. He gave me the opportunity to use his computer facilities and introduced me to the amazing world of molecular phylogenetics.

I am grateful to Dr. Najam us Sahar Sadaf Zaidi and Dr. Faiza Munir for their assistance, counselling and for their constant support throughout my MS program.

I am also grateful to principle ASAB Dr. Peter John who has maintained to provide excellent opportunities to students who strongly wish to carry research at the institute.

At this point I would like to specially thank my lab fellows Aisha Ashraf, Aroosa Maqsood, Maham, Khalid, Sir Fida and many significant figures who have helped me throughout my research, without their trouble shooting and skillful advice I would not have been able to complete my lab work. I am grateful to Shifa for her insightful advice during my research and for sharing her expertise.

I have been fortunate enough to have a charming group of friends. Leaving my hometown and moving to Islamabad was challenging. I express my gratitude to Maheen Javed, Mehk Arshad, Nigham Sarfaraz, Saher Shefait, Hira Sikandar, Sidra Asghar, Asma Mumtaz for their wonderful company. They always made me feel I have a home away from home. Without these people this journey would have not been this much beautiful and complete.

Finally, I would like to acknowledge my family. My beloved parents (Shah Nawaz, Bushra Gulnaz) who have always given me a free hand to achieve my dreams. Without their constant words of encouragement. I would thank my siblings Ayesha Shah Nawaz, Afsheen Fatima, Zahra Shah Nawaz, Abeera Shah Nawaz and the most beloved Abd-ur-Rehman for always being there for me and for being such an inspiration.

Hafiza Noor ul ain

Table of Contents

LIST OF ABBREVIATIONS

AIC	Akaike information criterion
PhyML	Phylogenetic Maximum Likelihood
GDP	Gross domestic product
SCMV	*Sugarcane mosaic virus*
SCSMaV	*Sugarcane striate mosaic associated virus*
VPg	Genome linked viral protein
CP	Coat protein
ML	Maximum Likelihood
HKY	Hasegawa, Kishino and Yano
PWD	Pairwise distance matrix analysis
SDT	Sequence demarcstion tool
MUSCLE	Multiple Sequence Comparison by Log Expectation
NCBI	National Center for Biotechnology Information
MCMC	Markov Chain Monte Carlo
NJ	Neighbor joining
RDP	Recombination Detection Program
ORF	Open reading frame
MDMV	*Maize dwarf mosaic virus*
MT	Million ton
SCMMV	*Sugarcane mild mosaic virus*
SCSMV	*Sugarcane streak mosaic virus*
UTR	Untranslated region
CI	Cylindrical Inclusion

LIST OF TABLES

LIST OF FIGURES AND GRAPHS

year 2008, it has reached up to 52% till year 2010, which is almost 17% increase, while in KPK starting from less percentage it has reached up to 53.75 within two years (Yasmin *et al.*, 2011).

Figure 4.4.1: A maximum likelihood phylogenetic tree estimated using CP gene of *Sugarcane mosaic virus*. Each branch is led by a virus and is denoted by following format, GenBank accession number_ gene name _Country. Five clades 1a, 1b, 1c, 1d and 1e are made. Branches are well supported by bootstrap values. CP gene of Maize dwarf mosaic virus is used as an out group.

Figure 4.6.1: A maximum likelihood phylogenetic tree estimated using P1 gene of *Sugarcane mosaic virus*. Each branch is led by a virus and is denoted by following format, GenBank accession number_ name of virus_ gene name _Country. Six clades named as 2a, 2b, 2c, 2d, 2e and 2f are made. Branches are well supported by bootstrap values. P1 gene of Maize dwarf mosaic virus is used as an out group.

Figure 4.8.1: A maximum likelihood phylogenetic tree estimated using whole genome of SCMV. Each branch is led by a virus and is denoted by following format, GenBank accession number _Country _Host _Date. Four clades A, B, C and D are made. Branches are well supported by bootstrap values.

Figure 4.9.1: Bayesian phylogeny of full genome of *Sugarcane mosaic virus* constructed for viruses with polyphyletic origin. Genbank accession numbers, countries, years and host are shown. Each clade is marked on right.

Figure 4.11.1: Recombination showing picture for Ecuador sequence. This figure is depicting recombination events. Major (JX047431) and minor (JX047410) parents are labeled below every recombination point. Recombined region in this recombination are C terminal region of 6K1, NIa and NIb.

Figure 4.11.2:Recombination analysis of Ecuador sequence. This figure is representing major and minor parents for KY006657 at recmobination points. Major and mainor parents are JX047431 and JX047410 respectively at recombined regions 6K1, NIa and Nib

Figure 4.11.3: Recombination showing picture for Ethiopian sequence. This figure is depicting recombination events. Major (AM110759) and minor (KF744391) parents are labeled below every recombination point. Recombined region in this recombination are CI, 6K1, NIa and NIb.

Figure 4.11.4:Recombination analysis of Ethopian sequence. This figure is representing major and minor parents for KP860936 at recmobination points. Major and mainor parents are AM110759 and KF744391 respectively at recombined regions 6K1, NIa and Nib

Figure 4.11.5: Recombination showing picture for Mexican sequence. This figure is depicting recombination events. Major (GU474635) and minor (AM110759) parents are labeled below every recombination point. Recombined region in this recombination are CI, 6K1, NIa and NIb

Figure 4.11.6: Recombination analysis of Mexican sequence (KP860936) This figure is representing major and minor parents for KP860936 at recmobination points. Major and mainor parents are GU474635 and AM110759 respectively at recombined regions 6K1, NIa and Nib

Figure 4.11.7: Recombination showing picture for Rwanda sequence. This figure is depicting recombination event for KF744390. Major (JX237863) and minor (KP860936) parents are labeled below every recombination point. Recombined region in this recombination are CI, 6K1, NIa and NIb.

Figure 4.11.8: Recombination analysis of Rwanda sequence. This figure is representing major and minor parents for KF744390 at recmobination points. Major and mainor parents arE JX237863 and KP860936 respectively at recombined regions 6K1, NIa and Nib

Figure 4.12.1: Pairwise distance matrix is representing by colors, range of percentage identity in numerical value is shown in scale on right side.

Figure 4.12.2 Species demarcation tool plot

Plot is representing valley (conflict free area) at 85% and at 89% to 92%. Peak represents percentage value with maximum conflict

ABSTRACT

Sugarcane mosaic virus (SCMV) belongs to genus *Potyvirus* (family *Potyviridae*). It is non-enveloped, rod shaped flexuous and filamentous positive sense RNA of approximately 750 nanometer in length. SCMV is responsible for causing mosaic disease in sugarcane, sorghum, maize and other *Poaceous* plants. Present study is aimed at to explore the genetic diversity and evolutionary relationships among various isolates of SCMV. Seventy-nine full length sequences of SCMV has been retrieved from databases and characteristics of genome were determined. Sequence features, recombination analysis, and phylogenetic relationships were performed through MatGat, RDP4, MEGA, Geneious and Bayesian matrix, respectively. Sequences size of SCMV full genome were between 9549 to 9640 nucleotides (nt) and between 3063 to 3078 amino acids. Total number of nucleotides for P1 were 698 to 732 nt while for CP nucleotides were 980 to 985 nt. The sequences of P1, CP and full genome shared nucleotide identity between 76%-100%, 58%-100% and 76%-100%, respectively while amino acid sequence identity showed between 82%- 100%, 58%-100%, and 80%-100%, respectively. The ranges of the sequence identity for P1, CP and full length are in accordance with species demarcation criteria given by ICTV. Recombination was not observed in CP region, while full genome sequences from Iran and China has shown several recombination events. Hot spots for recombination in SCMV are 6K1, Vpg, NIa-Pro and NIb. Phylogenetic trees of P1, CP and full genome sequences was constructed using Maximum likelihood method and verified through Bayesian analysis. The position of isolates in trees generated from P1, CP and full-length sequence were different the trees. Comparative analysis of individual genes and full genome shows three patterns of SCMV evolution. The first pattern reflected geographical based distribution, second pattern observed host dependent classification variation lies in presence of additional 20 nucleotide bases stretch in sequences from host maize, recombination-based pattern is observed in third pattern of evolution. The high levels of identity observed between isolates obtained from distant geographic areas could be explained by movement of maize germplasm since SCMV can be seed transmitted. Recombination plays an important role in virus evolution and emergence of new species/strains. Based on recombination and phylogenetic analysis, it is concluded that full length genome should be considered for classification and determination of evolutionary relationships. Full genome sequence data are crucial to determine biology, diversity and evolutionary history of a virus, for devising new control strategies and evaluating its risk. The results of this study will add new insights to the current knowledge of the SCMV evolution and genetic diversity in the world.

Chapter 1
INTRODUCTION

Introduction

Agriculture is the life line for economy of Pakistan and it accounts for 19.5% GDP (gross domestic product), 42.3% of labor force is linked to agriculture plus it provides raw material for various value-added sectors. It plays significant part in national development, poverty minimization and in food security. The biggest socio-economic challenge of 21st century is food security. However, over the last few years agricultural technological improvements have enhanced yield significantly, but still more productivity is needed to fulfil the growing requirement of food in coming years. One of biggest factor putting impact on agricultural productivity are plant diseases, caused by bacteria, fungi, nematodes and specifically by viruses (Safarnejad *et al.,* 2011; Liuji *et al.,* 2012).

1.1 *Saccharum officinarum*

Sugarcane is very important crop of family *Poaceae*, genus *Saccharum* and clan *Andropogoneae* (Daniels and Roach, 1987; Akbar *et al.,* 2017). Its scientific name is *Saccharum officinarium* taken from Greek word "*Sakcharon*" which originally means sucrose. It is a vital food crop of tropics and sub tropics areas. Worlds 80% sugar is produced from sugarcane. Sugarcane is majorly cultivated in Brazil followed by the countries like Mexico, India, Thailand, Pakistan and China, occupies more than 28.3 million hectares. (Saira Batool et al., 2015). Its importance is evident from the fact that, it is a food crop majorly used to produce raw and refined sugar, molasses, gurr, brown sugar, ethyl alcohol additionally it has dietary value, it can be used as food fodder, fertilizer and has industrial utilization as high biomass fiber producer, as a bagasse used for energy generation.

1.2 Sugarcane production in world

Globe wise, sugarcane production in Brazil is 739.3 million metric ton, which stands out first among sugarcane producing countries, while production quantity is 341.2, 125.5, 100.1, 61.2, 34.9, 33.7, 31.9 and 29.235 million ton in countries India, China, Thailand, Mexico, Columbia, Indonesia, Philippines and United States respectively. Overall worldwide average yield production range is 177.84- 179.64 million metric tons and sugar recovery is 9.90% (Annual Report PSMA, 2014-2015; Shokat *et al.,* 2017***).***

1.3 Sugarcane production in Pakistan

Pakistan stands 5th rank out of all sugarcane producing countries in terms of cane acreage production (Akbar *et al.,* 2017) while it occupies 8th rank in sugar production (Nazir

et al., 2013; Shokat *et al.,* 2017). Statistics showed in table 1.1, that during 2012-2017 it is grown on area ranges from 1,129 million hectares to 1,217 million hectares with production rate of range 63,750 million ton (from year 2012) to 73,607 million ton (year 2017), in case of yield its gain was 56,466 kgs/hec (2012) to 60,428 kgs/hec (2017). Statistically, province wise sugarcane production in Pakistan represents that, in Punjab its production ranges from 42,982-49,613 million ton cultivated in 76,1200-777,821 million hectares (during year 2012-2017) (PSMA, 2017; PBS, 2017). In Sindh area consumed by sugarcane production is 253,694- 320,501 million hectares which produces 15,966,224-20,208,895 million ton of sugarcane. In KPK, production of sugarcane ranges from 4,770,229-5,628,725 million ton, cultivation area ranges from 106,734-118-572 during years of 2012-2017 (PSMA, 2017; PBS, 2017).

Table 1.1: Sugarcane production and cultivation worldwide and in Pakistan

Countries	Amount of production (million metric ton)	Amount of production (million metric ton)
	Worldwide	***Pakistan***
Brazil	739.3	
India	341.2	49,613 (Punjab)
China	125.5	20,895 (Sindh)
Thailand	100.1	17,725 (KPK)
Columbia	34.9	
Indonesia	33.7	
Philippines	31.9 29.235	
USA	29.235	
Columbia	34.9	

Source: US department of agriculture annual report 2017 and Pakistan sugar mills association 2017.

1.4 Revenues, Export, Gross domestic product

Its production has enhanced agriculture's value addition by 3.4% and increased GDP (gross domestic product) by 0.7%. It is foreign exchange earner, provides revenues to the government and brings grower's prosperity (Karim, 2001). Pakistan has exported 1,064,215 –

399,309 million ton of sugar, 225,221-101,410 million ton of molasses, 142,065,426 - 358,483,301 liters of ethyl alcohol during the years of 2012-2017 at the average price rate of 48,573- 53,258 Rs/MT, 12,198-12,001 Rs/MT, 61.49-82 Rs/Liter, respectively as shown in table 1.2. In Pakistan about 1.5 million employees, directly or indirectly involved in approximately 86 sugar mills all around Pakistan put 1.9% share in Gross Domestic Product (GDP) (PSMA, 2017). Undoubtedly it is valuable crop in Pakistan but no worthy losses were identified each year, due to ailments instigated by range of several biological agents like fungi, bacteria and viruses (Bock and Baily, 1989; Akbar *et al.*, 2017).

Table 1.2: Export table for products obtained from sugarcane from the year 2012-2017

	Export	
Product	**Quantity** (2012-2017) (Million ton or Liters)	**Revenues** (2012-2017) (Rupees per million ton or Rupees per liter)
Sugar	1,064,215-399,309 (MT)	48,573- 53,258 (Rs/MT)
Molasses	225,221- 101,410 (MT)	12,198-12,001 (Rs/MT)
Ethyl alcohol	142,065,426-358,483,301 (Liters)	61.49-82 (Rs/Liter)

Source. Pakistan sugar mills association report 2017.

1.5 Diseases of Sugarcane

In Pakistan *Sachharum officinarum* (sugarcane) is extensively grown cash crop. It can be cultivated in subtropical and tropical areas, it can be grown in varied temperature, variation ranges from cool moist conditions to hot dry atmosphere. It is susceptible to diseases caused several biological agents such as viruses and fungi, that's why crop losses are observed each year (Akbar *et al.*, 2017; Bock and Baily, 1989).

Fungal diseases of sugarcane includes Black rot caused by fungus *Ceratocystis adipose*, *Cercospora atrofiliformis* and C*ercospora longipes* fungus are responsible for Black rot and Black stripe diseases respectively. Leaf blight and Red rot are induced by *Stagonospora tainanensis* and *Glomerella tucumanensis* respectively (Nasare *et al.*, 2007)

Viruses responsible for causing disease in sugarcane are *Sugarcane mild mosaic virus* (SCMMV), which infects sugarcane to cause mild mosaic, *Sugarcane streak mosaic virus* (SCSMV), it can cause streak mosaic in sugarcane, and striate mosaic is disease induced by *Sugarcane striate mosaic associated virus* (SCSMaV). However disease occur due to *Sorghum mosaic virus* and *Sugarcane mosaic virus* are usually referred as "sugarcane mosaic" (Rott *et al.,* 2002; Grisham, 2000)

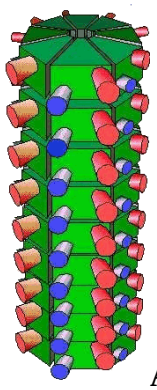
Mosaic disease caused by *Sugarcane mosaic virus* is most widely spread and it mainly affects sugarcane, maize and sorghum's crop yield (Shukla *et al.,* 1989; Liuji *et al.,* 2012), even it had caused the very first epidemic, recorded in the world in early 20th century which had almost collapsed the sugarcane industry. (Abbott, 1961; Yang and Mirkov 1997; Goncalves *et al.,* 2011). Mosaic disease in sugarcane is major cause of losses regarding cane yield (10-32%), sugar yield (6-10%) in Pakistan, it is record decline (Anwar, 2005; Haider *et al.,* 2011; Akbar *et al.,* 2017). According to statistics of 2010, KPK was more affected by *Sugar cane mosaic virus* with infection rate of 75%, comparatively to Punjab, with infection rate of 55% (Yasmin *et al.,* 2011; Akbar *et al.,* 2017). Mosaic Infectiveness more than 50% is responsible for huge yield waste. (Zia-ul-Husnnain and Afghan, 2004; Akbar *et al.,* 2017). SCMV is the only group of virus in the genus *Potyvirus* that can infect members of family *Poaceae*, presumably it emerged 7250 years ago in northern Africa and South-East Asia hence it is one of oldest member of genus. (Gibbs and Ohshima, 2010; Rosales *et al.,* 2015)**.**

1.6 *Sugarcane mosaic virus*

Sugarcane mosaic virus (SCMV) is a member of the genus *Potyvirus* and belongs to family *Potyviridae*. It is one of the largest and most economically important genera of plant viruses (Moradi *et al.,* 2016; Adam *et al.,* 2012; Handley *et al.,* 1998)**.** It is made up of flexuous filamentous rod shaped non enveloped particle. This shape is due to encapsulation by almost 2000 monomers of CP (coat protein) (Riechmann *et al.,* 1992; Adams *et al.,* 2005; Rosales *et al.,* 2015) with viral length of 680-900 nanometer with the width of 11 nanometer (Haider *et al.,* 2011; Akbar *et al.,* 2017; Addy *et al.,* 2017; Wylie *et al.,* 2017). It consists of poly A tail at 3' untranslated region and covalently linked VPg (viral protein) at 5' translated region instead of canonical cap (5'm7G).

1.7 Morphological structure of SCMV

SCMV is flexuous, filamentous, rod shaped, non-enveloped particle. Morphological structure of SCMV is shown in Figure 1.1



A

B

C

Figure 1.1: Morphological structure of *Potyviridae*. A is the schematic diagram of *potyvirus* particle. The N terminus with 30 amino acids (large rectangle) and C terminus with 19 amino acids (small rectangle) from coat protein (CP) molecules are apparent on the surface of intact virus particle (Shukla and Ward, 1989; ICTV 10th). B is stained electron micrograph of virus particles from family *Potyviridae*. C is representing non-enveloped, filamentous, flexuous particle of length 650-900 with diameter of 12-15 nm, constituting helical symmetry (Swib, 2008).

1.8 Symptoms

Common symptoms of the SCMV disease are mosaic patterns, streaks and stunting, in different hosts like sorghum, sugarcane and maize. In year 1919, Brandes described symptomatology of virus as "spots on the leaves or irregular light-colored streaks" with strikingly different patterns which distinctly differentiate between affected and unaffected areas as shown in Figure 1.2. Prominence was put on coloration, position and streaking pattern, which are host dependent, host includes corn, sorghum, sugarcane, rice, millet, foxtail/ *Panicum* and crab grass. In case of maize, infected by SCMV, plants were symptomatized as short plants with mosaic pattern or mottle at the base of young leaves, diffused. In young leaves continuous or broken narrow streaks appear along the veins as light areas conjoin, streaks are present between margins and midribs, while sides and tips are more

distinct by chlorotic or mosaic (uniformly low) area (Brandes, 1919; Williams and Alexander, 1965; Rosales *et al.*, 2015). SCMV usually produces ring-spot, mosaic pattern, reduced growth (stunting) on its natural hosts as sugarcane (Saccharum spp), sorghum (*Sorghum bicolor*), Panic grass (*Panicum* spp), foxtail (*Setaria* spp), Maize (*Zea mays*) and *Eleusine* spp (Adams and Antoniw, 2006; Rosales *et al.*, 2015).

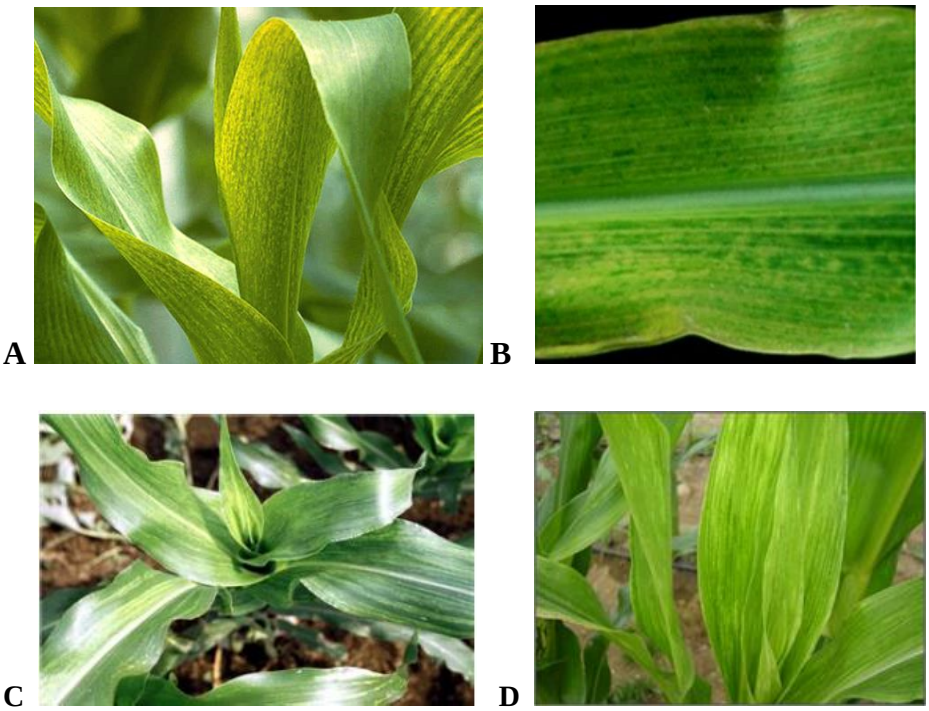


Figure 1.2: Symptoms of *Potyviruses***.** (A) Mosaic symptoms on *Zea Mays* (Menabde D.A. 1978). (B) Contrasting shades of green on sugarcane leaf. (C) Mosaic symptoms on top of maize leaves (Malisa, 2018). (D) Sorghum leaf with mosaic patterns (John, 1982).

1.9 Transmission

Since the starting of studies regarding *potyvirus* transmission through aphids more than forty years ago, significant progress has been made to understand molecular process that governs the mechanism of transmission, specifically in 1980s and 1990s for better understanding (Revers and Garcia, 2015; Blanc *et al.*, 2011; Brault *et al.*, 2010, Lopez *et al.*, 2002; Ng and Falk, 2006; Pirone and Blanc, 1996; Stafford *et al.*, 2012). Like other potyviruses (Blackman ans Eastop, 2000; Gibbs *et al.*, 2008), transmission of SCMV take place by many aphid species.

Theory proposed for and supported by practical results is “the bridge hypothesis”. According to bridge hypothesis (as represented in Figure 1.3) HC-Pro region consists of PTK motif, CP region contains DAG motif, PTK binds to DAG at exposed region of CP thereby mediates retention of virion at suitable site in the vector. Studies suggest that presence of PTK motif is essential for transmission through aphid. *Zuchhini yellow mosaic virus* binds to the aphid stylet through KITC motif present in the HC-Pro at N-terminus, and it forms an active bridge between aphid and virion CP (Peng *et al.*, 1998; Rosales *et al.*, 2015) Location of PTK and KITC motifs in SCMV are at 287 and 1634 amino acid position.

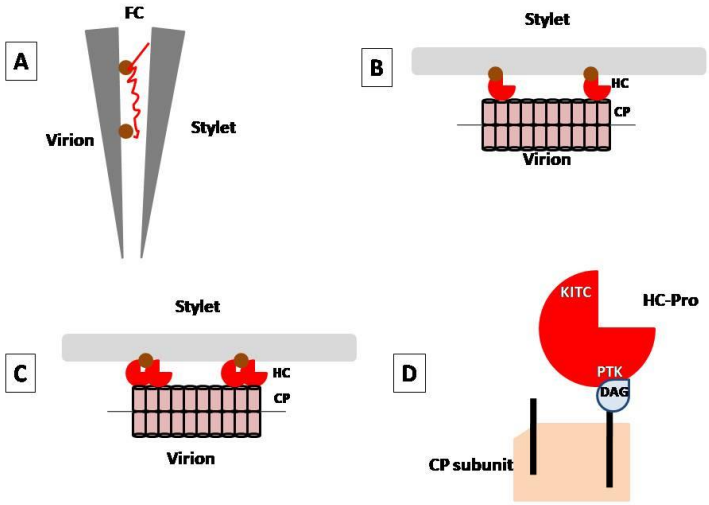


Figure 1.3: Transmission (Bridge Hypothesis). (A): Depicting stylet of aphid with virion attached at its receptor. (B) and (C) representing how virion is attached to receptor of aphid through HC-Pro and coat protein, one part (KITC) of HC-Pro is attached to receptor in stylet of aphid, on the other hand second part (PTK) of HC-Pro is attached to coat protein at DAG motif, and making bridge like attachment. (D) HC-Pro consists of KITC and PTK motif, coat protein consists of DAG motif, which is attached to HC-Pro at PTK region.

1.10 Occurrence of SCMV

Sugarcane mosaic virus has been reported in twenty-five countries including Argentina, Brazil, Cameroon, China, Australia, Congo, Denmark, Colombia, Egypt, Ethiopia, Iran, Germany, India, France, Pakistan, Mexico, Spain, Vietnam, United Kingdom, South Africa, Poland, United States and Netherlands (Liuji *et al.*, 2012). World map with countries highlighted where SCMV is present is shown in Figure 1.4. In Java during 1882 it was first discovered as causing agent of anomaly in sugarcane. In Puerto Rico during 1919 it was

characterized as viral entity in grasses and sugarcane (Liuji *et al.*, 2012). Brandes, (1919) reported, spread of virus by pointing out that source of infection originated in Java and later on spread to USA, Brazil, Peru and Argentina (Brandes, 1919; Abbot, 1929; Koike and Gillaspie, 1989; Rosales *et al.*, 2015). Since 1930s, virus was reported in maize in Sub Saharan Africa (Cronje, 2001). In USA, an outbreak was reported in corn during 1963 (Louie and Knoke, 1975; Rosales *et al.*, 2015). In 1960s proof of SCMV infection was found in cane fields of India, Taiwan and Thailand (Abbot and Stokes, 1966; Sharma *et al.*, 2002) and in fields of maize in China (Chen *et al.*, 2002). Since then virus has been reported in sorghum, sugarcane and maize in countries Australia, Kenya (Teakle and Grylls, 1973; Louie and Darrah, 1980) Italy (Tosic *et al.*, 1977) Morocco (Fischer and Lockhart, 1974) Egypt, Japan, Colombia (Gillaspie and Mock, 1979) Mexico (Delgadillo, 1987; Espejel *et al.*, 2006) Germany (Oertel *et al.*, 1997) Spain (Achon *et al.*, 2007) tropical Africa (Thottappilly *et al.*, 1993) Cameroon, Pakistan and Iran (Gillapsie *et al.*, 1978; Rosales *et al.*, 2015)

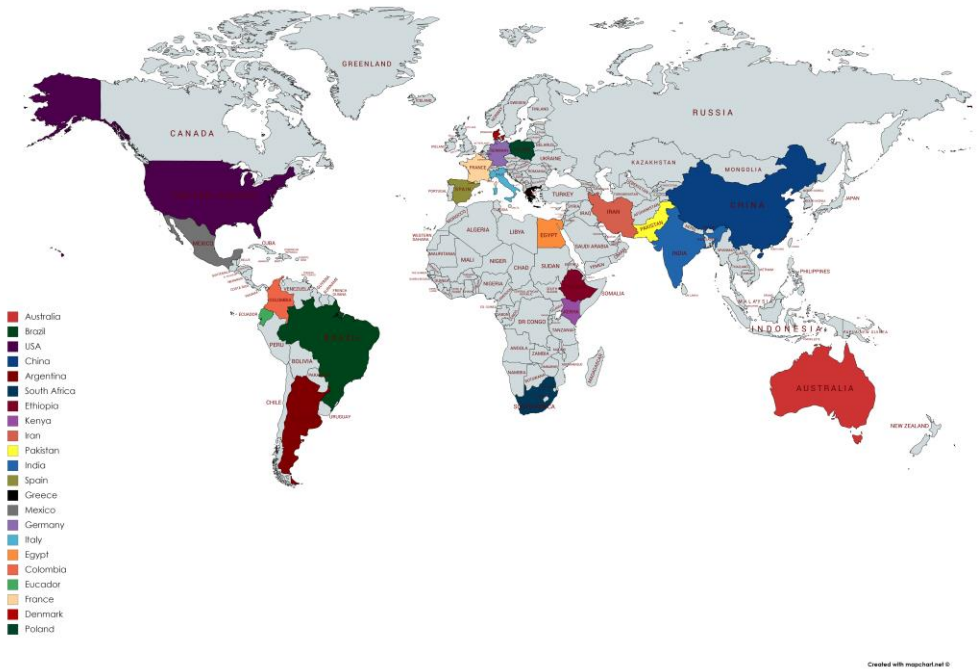
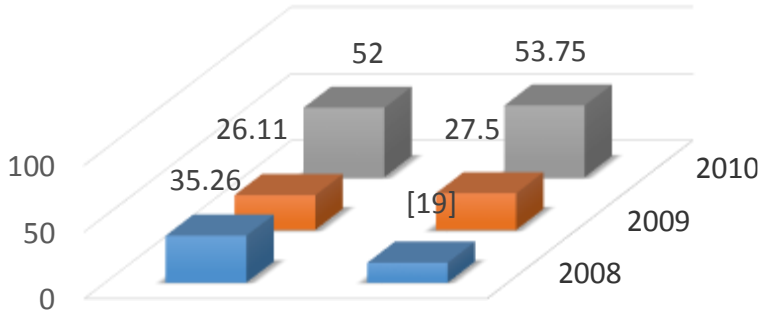


Figure 1.4: World map. This map is representing the countries where SCMV has been reported.

1.11 Prevalence of SCMV in Pakistan

Prevalence of *Sugarcane mosaic virus* in Pakistan is shown in Graph 1. Graph is depicting increasing impact of SCMV in provinces Punjab and KPK year wise. Initiating from 35.26 in year

2008, it has reached up to 52% till year 2010, which is almost 17% increase, while in KPK starting from less percentage it has reached up to 53.75 within two years (Yasmin *et al.*, 2011).



Graph 1: Graph is depicting increasing impact of SCMV in Pakistan provinces Punjab and KPK year wise (2008-2010)

1.12 Genomic Organization

SCMV genome codes for ten proteins those are P1 (first protein), HC-pro (helper component proteinase), P3 (third protein), 6K1 (first 6K protein), CI (cylindrical inclusion protein), 6K2 (second 6K protein), VPg (viral protein genome linked), NIa (nuclear inclusion a protein), NIb (nuclear inclusion b protein) and CP (coat protein), from N-terminal to C-terminal (Addy *et al.*, 2017) along with PIPO which is small frame shift derived peptide. (Adams *et al.*, 2005; Chung *et al.*, 2008; Rosales *et al.*, 2015).

1.12.1 P1

The potyviral P1 protein consists of serine-protease domain that cleaves itself at C terminal (Verchot *et al.*, 1991; Revers and Garcia, 2015) typically at position of Tyr/Phe-Ser (Valli *et al.*, 2007; Wylie *et al.*, 2017). It is highly basic protein with the potential to interact with nucleic acid *in vitro.* Genomic amplification is stimulated by P1 *in Tran's* manner (Verchot and Carrington 1995b; Revers and Gracia, 2015). P1 plays an important part in replication of virus, because it can stimulate the HC-Pro which is gene silencing suppressor (Anandalakshmi *et al.*, 1998; Pruss *et al.*, 2004; Rajamaki *et al.*, 2005, Valli *et al.*, 2007; Revers and Garcia, 2015; Wylie *et al.*, 2017). P1 increases infection in RNA silencing lacking plants which represents that it plays its part in causing virus infection independent of RNA silencing suppression (Pasin *et al.*, 2014; Revers and Garcia, 2015).

In general, it is least conserved sequence protein and most divergent in size ranges from 30-63 kDa, although its protease domain located at C-terminal is well conserved (Adams *et al.,* 2005; Valli *et al*., 2007; Yoshida *et al.,* 2012; Revers and Garcia, 2015) while variability of its N- terminal region consequently makes it most variable. This part is highly dis-arranged and negatively controls the self- cleavage of P1 (Pasin *et al.,* 2014; Revers and Garcia, 2015). P1 interacts with Rieske Fe/S protein of host, but the significance of this interaction is yet to be revealed (Shi *et al.,* 2007, Revers and Garcial, 2015).

1.12.2 Coat Protein (CP)

The fundamental function of capsid protein is to encapsulate the viral genome. To make the potyviral virions, almost two thousand coat protein units helically organize themselves around genomic RNA. CP is the most conserved region in potyviral genome, and it makes the core of virus particle (Dolja *et al.,* 1994; Jagadish *et al.,* 1993; Varrelman and Maiss, 2000; Voloudakis *et al.,* 2004; Revers and Garcia, 2015). Moreover CP inter subunit interaction is responsible for start of virus assembly, N and C terminal domains plays an important role here (Anindya and Savithri, 2003; Kang *et al.,* 2006; Seo *et al.,* 2013). CP is known to have post-translational modifications such as phosphorylation, and O-GlcNAcylation (Ivanov *et al.,* 2003; Chen *et al.,* 2005; Scott *et al.,* 2006; Subr *et al.,* 2010; Revers and Garcia, 2015).

Coat protein is involved in NTPase activity, probability exists that this activity is linked to regulatory mechanism (Rakitina *et al.,* 2005; Revers and Garcia, 2015). It also plays an important role in virus movement, amplification of genome and transmission of virus (Wylie *et al.,* 2017). It interacts with host protein Rubisco (Feki *et al.,* 2004) which is an import host element for virus infection (Bhat *et al.,* 2013; Zhao *et al.,* 2013). Thus, interaction of coat protein with Rubisco leads to probability that CP has a part in virus infection or in defensive responses of plants (Revers and Garcia, 2015**)**

1.12.3 Genome Structure

SCMV is monopartite in nature with genomic size of 9.7 kb, shown in Figure 1.5. It consists of single stranded RNA. Viral genome linked protein (VPg) is present at 5'UTR and poly A tail is present at 3' UTR. Single open reading frame is processed into poly-protein by autocatalytic cleavage.

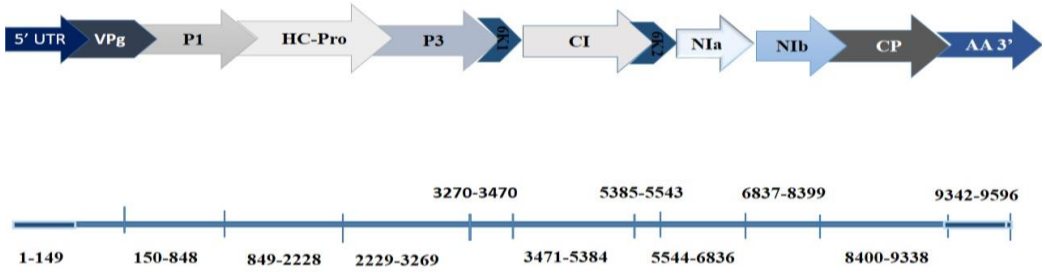


Figure 1.5: *Sugarcane mosaic virus* consists of viral genome linked protein at 5" untranslated region, Protein 1 (P1, serine-protease domain), HC-Pro (cysteine-protease domain), Protein 3 (P3, involved in virus replication), 6K1, Cylindrical inclusion (CI, RNA helicase), 6K2, Nuclear inclusion a (NIa, serine like cysteine protease), Nuclear Inclusion b (NIb, RNA polymerase), Coat protein (protects genomic RNA, involved in virus movement and aphid transmission), Poly A tail at 3" untranslated region.

1.13 Co-infection of SCMV

Co-infections of SCMV happens with *Maize chlorotic mottle virus* (MDMV) which belongs to family Tombusviridae and genus Machlomovirus which leads to maize lethal necrosis disease as a combination of mutual reinforcement of two diseases (Uyemoto *et al*., 1980; Wangai *et al.,* 2012). SCMV is linked to co infections with MDMV and *Maize white line mosaic virus* which belongs to family *Tombusviridae* and genus *Aureusvirus*. High incidence of mixed co infection has been reported by study in Southern China between SCMV and SrMV in hybrid sugarcane and in noble sugarcane (Xu *et al.,* 2008). In Brazil maize fields, which were surrounded by sugarcane fields case of mixed infection has been reported between SCMV and *Maize raya do fino virus* which belongs to family *Tymoviridae* and genus *Marafivirus* (Goncalves *et al.,* 2007). In Pakistan, Yasmin *et al.* (2011) reported mixed infection in sugarcane fields and weed species located in close proximity between SCMV, *Maize dwarf mosaic virus*, Sugarcane bacilliform virus and *Maize streak virus.* Co-infections of SCMV and *Sorghum mosaic virus* has been documented by Yahaya *et al.* (2014) in sorghum. Mixed infection of *Barley yellow dwarf virus* with SCMV in maize has been reported by Ilbagi *et al.* (2006) in Turkey. Mixed infection of MCMV and SCMV has recently caused the outbreak of maize lethal necrosis disease in Sub Saharan Africa (Wangai *et al.,* 2012).

1.14 Economic impact

SCMV is one of the most important viruses that infects maize and second is *Maize dwarf mosaic virus*. Serious yield losses can be caused by these two viruses (Louie *et al.*, 1991; Fuchs and Gruntzig, 1995; Lapierre and Signoret, 2004; Ali and Yan, 2012; Rosales *et al.*, 2015) for example in China yield lose range from 20% to 80% in maize production (Chen *et al.,* 2002; Jiang and Zhou, 2002; Silva *et al.,* 2015). Prevalence and impact of viruses vary from country to country as its effect depends on soil condition, climate and form of germplasm cultivated. More diseases are reported in Spain and Germany from Europe than in any other country of that region. Since the middle of 1990s, most cases are detected and documented in Germany (Oertel *et al.,* 1997; Silva *et al.,* 2015). Since 2007, SCMV incidence rate was increased 10% in Spain (Achon *et al.,* 2007; Rosales *et al.,* 2015). Asia and Europe are more susceptible to loses due to SCMV as compared to America, where SCMV has no such striking effect except for some sporadic out breaks (Gilbert *et al.,*2005; Silva *et al.,* 2015). Likewise, in Kenya and the Sub Saharan region mosaic disease is under control with infrequent outbreaks of economic importance (Cronje, 2001; Wangai *et al.,* 2012). SCMV has shown high level of incidence along with other plant virus co-infections in China (Xu *et al.,* 2008; Liuji *et al.,* 2012) and in Argentina (Perera *et al.,* 2009, Liuji *et al.,* 2012).

1.15 Aims and Objective

The fast evolution and huge sequence diversity found in SCMV sequence is a reason of major obstacle in development of effective management and control strategies, detection approaches, to map prevalence pattern and to develop resistant varieties. It is responsible for causing enormous destruction to crops i.e. sugarcane, maize and sorghum. Furthermore its capability of causing co-infection with other viruses is more lethal and cause huge economic damage.

The goal of this study was to determine genetic diversity and evolutionary analysis of *Sugarcane mosaic virus*. We tried to gain insight of the evolutionary aspects of SCMV full genome, coat protein and P1 region, to determine which region would be most appropriate to study when it comes to molecular evolutionary or phylogenetic analysis.

1) To explore genetic diversity of *Sugarcane mosaic virus*

- Retrieval of sequences from database i.e. NCBI, GenBank
- Create multiple sequence alignment
- Formation of pairwise distance matrix (nucleotide and amino acid based) to find percentage nucleotide identity between sequences (full genome, CP and P1) of SCMV.

2) To determine evolutionary relationships of *Sugarcane mosaic virus*

- Phylogenetic analysis (maximum likelihood based phylogenetic trees) to be performed to find evolutionary pattern of *Sugarcane mosaic virus*.
- Bayesian analysis to be done to validate the results obtained from ML tree.

Chapter 2

LITERATURE REVIEW

Literature review

2.1 SCMV Diversity

In the early era of 1960's, *Maize dwarf mosaic virus* was found in Ohio, and characterized as A, B, C, D, E and F strains (Williams and Alexander, 1965; Adams and Antoniw, 2006; Silva *et al.*, 2015). During that time virus species were got separated on the basis of host range by plant virologists. As such it was observed, MDMV was able to infect sugar cane but could not affect Johnson grass, while MDMV A was capable to infect Johnson grass but not to infect sugarcane. SCMV and MDMV were not discriminated for long time.

In 1989, differentiation between SCMV and MDMV happened much later through the pioneering efforts of Shukla and co-workers. They inspected MDMV and SCMV isolates affecting maize and sugarcane respectively, through cross-absorbed antisera against MDMV B and MDMV A. On the basis of cross reactions against seventeen strains of MDMV / SCMV they classified the strains as, (1) *Johnson grass* mosaic virus group consisting of two strains, SCMV JG and MDMV O, (2) *Sorghum mosaic virus* group which consists of three strains, SCMV-I, SCMV-M and SCMV-H, (3) MDMV group containing 4 strains named as MDMV-A, MDMV-F, MDMV-D and MDMV-E, (4) SCMV group which is made up of 8 strains, SCMV-A, SCMV-B, SCMV-D, SCMV-E, SCMV-SC, SCMV- BC, MDMV-B and SCMV- Sabi (Shukla *et al*., 1989; Silva *et al*., 2015)

Delimitation of virus genera and species was done by using phylogenetics and pairwise similarity values by using nucleic acid sequences. According to species demarcation criteria for genus *Potyvirus* (Adams *et al.*, 2005; Silva *et al*., 2015), the SCMV group presently made up of following species, *Sugarcane mosaic virus, Sorghum mosaic virus, Maize dwarf mosaic virus, Cocksfoot streak virus, Pennisetum mosaic virus, Zea mosaic virus* and *Johnson grass mosaic virus* (Chen *et al.*, 2002; Gibbs and Ohshima, 2010; Silva *et al.*, 2015**.** SCMV is the only group of genus *Potyvirus* that infects members of family Poaceae (Gibbs and Ohshima, 2010, Silva *et al*., 2015).

2.2 Genetics, Strains and Phylogenetic

Molecular phylogenetic analyses are broadly used to re-construct evolution-based relationships among organisms using sequences either nucleotide-based or amino acid based. To reconstruct evolution, basic step is to attain accurate alignment or correct data matrix. This represents that nucleotides shares common ancestor if they are at same location. The alignment leads us to develop phylogenetic tree for phylogenetic inference (Huda, 2016).

A phylogenetic tree contains nodes which are attached to branches. Branch length represents estimate of ratio of evolution between every two nodes of the tree while nodes describe taxa, present day taxa is represented by terminal node whereas contemporary taxa can be described by internal node. Phylogenetic tree can be of two types, (1) rooted, (2) un-rooted. A defined direction of evolution can be represented by rooting tree, in which root indicates a mutual ancestor of all individual taxa existing in the tree whereas in un-rooted tree, no one can discriminate about which node is representing ancestor of all taxa (Huda, 2016).

A tree can be rooted by multiple approaches. Out-group rooting in which distantly allied taxa is added which is called an out group. Second type is midpoint rooting, in which a root is added to a point where two most dis-similar taxa are joining in the tree (Ghori, 2016)

Molecular data is use to reconstruct phylogenetic history with various methods. These methods can be characterized on the basis of form of data used or on the basis of algorithm that is used to make the trees. Methods are classified as distance method and discrete method on the basis of data that is to be used. Whereas on the basis of algorithm, two methods exit, those are cluster method or optimality criterion method (Ghori, 2016)

Handley *et al.,* 1998 had sequenced the CP (coat protein) coding sequence of eleven field isolates of SCMV belonged to countries Australia, South Africa and USA. Nucleotide and amino acid based comparison of isolates has demonstrated 0.2% to 4.1% and 0.0% to 3.5% difference respectively. Phylogenetic analysis was performed of SCMV and other potyviruses such as SCMV-MDB, JGMV, MDMV, SrMV and PVY. SCMV made a tightly cladded match with SCMV-MDB, and made a separate branch. This represented that SCMV belong to one strain which is SCMV-A, and are not geographically distinct species. SCMV-MDB may show another potyvirus species and is un-ambiguously distinct.

National Center for Biotechnology Information contains 866 sequences related to SCMV. These isolates are from 26 countries and belongs to different hosts. Primarily, coat protein (CP) nucleotide gene sequences were analyzed to interpret relationship between geographic origin and host. Chen's group infected diverse grasses to find out phylogeny of *Potyvirus* as well as to find co-relation between CP gene and host (from which isolate was derived). Two well defined clades were made, one from maize and one from sugarcane (Chen *et al.,* 2002). On the other hand, Algeria *et al.* (2003) grouped isolates of SCMV belongs to USA, China, Africa and Germany on the phylogenetic basis into two main monophyletic

clades, maize, sugarcane and thirteen minor additional groups. In 2005, the SCMV group was re-classified into 3 groups. Group I, maize, (MZ) that was distributed into 2 sub-groups with geographic co-relations. (a) Group IA, named as Amero-European, which consists of isolates from Mexico, Germany and Spain (b) Group IB, Asian, contains isolates from china. Group II comprised of sugarcane isolates from USA, Australia and China, whereas Group III is constituted of novel isolates from Thailand, mixed isolates from sugarcane and maize and a unique MDB-SCMV isolate from USA (Gemechu *et al.*, 2005).

Succeeding analysis, in 2008, described five clades and two unique strains of SCMV isolates affecting sugarcane, *Sachharum officinarum* (noble) and *Sachharum* interspecific hybrids (hybrid cultivars). The SCE clade consists of isolates from China, Australia, Brazil, Iran, India, USA, Africa and Pakistan belongs to hosts like noble sugarcane, maize, weed and hybrid cultivars of sugarcane. The MZ group was belonged to hosts sorghum and maize while they were from countries China, Spain, Argentina, USA and Mexico. The third group, named as SCE/MZ was solely compromised of isolate from Thailand. Furthermore, geographically distributed (into Southern and Eastern China isolates) new group was formed, consisting of isolates from noble sugarcane. In the last, Brazilian isolates made the fifth clade. Two distinctive strains, SCMV Abaca was isolated from *Musa textilis* in Philipines and SCMV-MDB (now known as SCMV-Ohio) was isolated from maize in USA were reported (Xu *et al.*, 2008; Silva *et al.*, 2015).

Ha *et al*., 2008 reported phylogenetic analysis of SCMV belonged to Vietnam and sampled from maize, arrow root and sugarcane. Analysis showed diversity of SCMV and were differentiated into three clades. Clade constituted of isolate sampled from arrowroot was named as SCMV-VN=AR1. Nucleotide-based (CP) similarity group represented with other isolates of SCMV was 75.1% to 80.7%, and maximum identity (89.7%) was shown with isolates belonged to china from host maize. Second clade consists of two SCMV isolates from host maize, named as SCMV VN=M1, -VN=M2 and one from host sugarcane named as –VN=SCI). CP nucleotide-based identity isolates shared together was 93.7% to 98%, percentage similarity with other SCMV isolates was 72.7% to 80.7%, with the highest percentage resemblance (98.5%) to isolate from Thailand. Third clade was made up of sugarcane isolates named SCMV-VN=SC2, -VN=SC4 and –VN=SC3. CP nucleotide similarity they shared with each other was 87.3% to 91.8%, with other isolates (from Vietnam) it was 72.7% to 76.7%. Low similarity has been shown by these three isolates with

published SCMV isolates, highest identity they shared with Chinese isolates, from host maize, and it was 79.1%.

Viswanthan *et al.*, 2009 constructed phylogenetic tree from hyper variable region and CR part from the CP coding part. Forty-six isolates were used to form the phylogenetic tree from hyper variable region and divided into nine clusters (HVR1 – HVR9), except type strains, those are SCMV A, B, D and E from country USA and some isolates (CB617, CBNHG671 AND CBC92061) from country India. In both phylogenetic trees (HVR based and CR region of CP nucleotides based), thirty Indian isolates were grouped into five different phylogenetic clades. Out of which four are made from CR (CR1, CR2, CR3 and CR9) part of CP coding sequences plus HVR (HVR1 – HVR4) were specifically taken by Indian isolates.

Mainly the viral isolates were distributed on bases of their origin in the phylogenetic tree except SCMV-IND (from India) and SCMV-SC (from Australia), these two isolates cladded together with sharing of 99% nucleotide identity. Isolates from Argentina, Republic of Congo, Australia, Iran and Cameroon were placed in different groups in both trees. Analysis concluded that core region of CP is best choice to observe sequence relationships instead of HVR which is not very reliable region to study.

CP nucleotide-based phylogenetic analysis has divided all available SCMV sequences in database into four well supported clusters. Cluster I being most diverse can be differentiated in term of three factors, one, host, which includes maize, banana, arrowroot and sugarcane, second geographic origin such as Asia, Europe, America, third factor includes sequence distances. Vietnam sequences from Clade 1 and 2 fell within this group. Moreover, sequences from Vietnam (clade 3 isolates) from host sugarcane combined together to form Cluster 4. Cluster 2 constituted of isolates from Brazil only. Cluster 3 made up of isolates (host, sugarcane) from different continents.

In 2011 phylogenetic study of SCMV was performed that rendered 5 clades, A, B, C, D and E, analysis was based on full genome (for the first time) and CP gene (Gao *et al.*, 2011). Isolates of maize from Argentina, China, Mexico, Spain and Germany was characterized as group A named as MZ, isolates of noble sugarcane was belonged to group B with the name SCE, mixture of isolates of sugarcane and maize from countries Thailand, China and Vietnam classified as group C and named as SCE/MZ, isolates of maize from USA/Mexico and isolates of sugarcane from USA, South Africa, Iran, China and Australia

were present in the group D, group E constituted of most divergent isolates of sugarcane belonged to China and Vietnam. Moreover, fourteen complete genomes were observed five from sugarcane and nine from maize, a consistent cluster was analyzed. Group I, maize isolates distributed into two subgroups, IA with isolates of maize form China, IB isolates of maize from America and Europe, group II sugarcane isolates consisted of four Chinese sequences, Group III, cluster of sugar cane isolates with unique sequences, belonged to country Australia.

Full length CP gene was observed by using maximum likelihood through PhyML (Guindon *et al.,* 2010), JModelTest 2 was used to identify the model to apply (GTR+G+I) (Darriba *et al.,* 2012), and represented three SCMV lineages. Sugarcane's ancestral group was observed in lineage III, constituted of isolates from Yunnan (province of China) and Vietnam, till now, the nearest isolates to Java island, that was discussed as origin of mosaic disease in sugarcane crop (Brandes, 1919). Lineage II was classified as sugarcane cluster, it was solely made up of isolates from sugarcane and belonged to countries USA, Argentina, China, Iran, India, Australia and Cameroon. Maize isolates made the lineage I, subdivided into 4 maize groups named as MZa-MZd and the hybrid group with name H, which consisted of interspecific isolates from *Marantha arundinacea*, sugar cane, maize and sugarcane. Group MZa clade consists of isolates from Europe and America, group MZb consists of maize isolates from China. MZc the most divergent clade belonged to Brazil, North America and Rwanda. Combination of maize and sugarcane isolates from South East Asia and isolates from China, Thailand and Vietnam observed affecting *Setaria* spp. and *Musa* spp. were classified into MZd group. Overall, the evolutionary history of SCMV from ancient isolates can be represented by this grouping. Lineage III contains the isolates that solely affects sugarcane, isolates that continued to infect sugarcane host were arose from this clade (lineage III) and after that spread into other hosts, including lineage I and H those consists of isolates that infects maize. Two main clades depending on maize and sugarcane hosts were analyzed upon reconstruction of phylogenetic history of SCMV by using Bayesian and maximum likelihood analysis. Twenty-two complete genome sequences were used to construct the phylogenetic tree. Wide spread events of SCMV genome was found in five continents through recombination and phylogenetic analysis.

Moradi *et al.*, 2016 performed phylogenetic analysis of 23 whole genome sequences along with two naturally existing recombinant isolates, (NRA; KT895080, ZRA; KT895081)

of *Sugarcane mosaic virus* from Iran, revealed that SCMV can be classified into three groups in the Neighbor joining tree. Clade 1, consists of twelve isolates from host maize and countries China, Spain, Mexico, Ethiopia and Germany. This clade is sub divided into two groups based on geographical origin. Clade II consists of five isolates, from hosts maize and sugarcane, all originated from China. Clade III consists of six isolates, from hosts maize and sugarcane both, from countries USA, Australia and Argentina (Moradi *et al.*, 2016). Out of those three clusters both Iranian isolates can be added into group III which consists of SCMV isolates from Australia and Argentina, but recombinant isolates form separate sub-lineage. Iranian isolates genomic percentage identity is 92% and 91.7% respectively, with isolate of Australia (AJ278405), while Iranian isolates share maximum poly-protein amino acid-based identity with isolate from Argentina (JX237862). 5' untranslated region and P1 portion of recombinant isolates were more identical to isolate form United States (host; maize), HC-Pro, NIa- VPg, NIb, coat protein (CP) and P3 regions were more identical to isolate from Australia, furthermore 3' untranslated region with remaining five proteins were more identical to isolate from Argentina.

Deng *et al.*, 2016 aimed to study phylogenies of *Sugarcane mosaic virus*. Two SCMV isolates (whole genome) was found at Fuzhou, named as FZ-C1 and FZ-C2. Population structure and genetic diversity were observed on the basis of full genome of mentioned two sequences along with eighteen other SCMV isolates by using MEGA 4.1 and Simplot software. Results represented that FZ-C1 consists of 9570 nucleotides, FZ-C2 constituted of 9573 nucleotides, with single open reading frame of 9189 nucleotides which encodes poly protein of 3063 amino acids. FZ-C1 and FZ-C2 shared high percentage nuclear and amino acid based identity and restriction sites with SCMV A. On phylogenetic basis SCMV could be differentiated into three clades. SSG 1, 2 and 3. Strains taken from maize were cladded into SSG 1, strains isolated from sugarcane were clustered into SSG 2 and 3, and SSG 3 were specifically taken from chewing cane. Both FZ isolates were cladded into SSG 2. This study analyzed the whole genome of SCMV and examined the phylogenies of SCMV.

Cheng *et al.*, 2017 Constructed neighbor joining tree from the complete genome sequences of KU171814 (DWK1), KU171814 (DKW2) from China (Shandong: host; maize) and fourteen other *Sugarcane mosaic virus* represents that isolates KU171814 and KU171815 shared nucleotide identity of 81.7%, while KU171814 (DWK1) shared maximum identity (90.9%) with isolate AY569692 (SX) and KU171815 (DKW2) shared maximum identity

(99.4%) with isolate JN021933 (BD8). Phylogenetic tree depicts that KU171814 (DWK1) and KU171815 (DWK2) are classified into groups 1 and 4 respectively.

Wang *et al.,* 2017 reported SCSMV phylogenetic tree, according to which all Chinese SCSMV isolates were grouped into one China cluster, isolates from different countries clades into different group. This predicts that SCSMV has definite geographic specificity. This result contradicts analysis of Viswanthan *et al.,* 2008, who stated that 14 phylogenetic groups were made from isolates belonged to India, Australia, South Africa and USA, and the clustered pattern represented that viral isolates couldn't be differentiated on geographical origin bases of host varieties. Surprisingly in China group one Indian SCMV isolate was found, which could be due to same geographical origin or due to common ancestral strain of Indian and Chinese isolate. Indian SCSMV isolates were present in each group, predicting various population types or high genetic diversity.

Akbar *et al.,* 2017 has investigated evolutionary history of SCMV through viral coat protein gene. DNA sequences were used to make the phylogenetic tree. Sequences were taken from NCBI database depending upon high sequence identity in Basic Local Alignment Search Tool (BLAST). Alignment was done using ClustalW and Neighbor Joining tree was made by using MEGA 6 software. A single clade was formed of all CP nucleotide sequences showing that virus affecting sugarcane in several areas of Pakistan have similar sequence and viruse is originated from same source. In present study, three clusters are formed from selected sequences (present in database). Clade 1 constituted of isolates belonged to countries Pakistan, China, Argentina, Australia, Iran, France and South Africa. Cluster 2 solely consists of isolates from India. Isolates from China, Germany, Thailand, Mexico and Spain were classified into cluster 3, and this clade represented divergence from other two groups (cluster 1 and 2).

Complete genomes of Spanish isolates of SCMV-Sp and MDMV-Sp were sequenced completely. SCMV-Sp shared 79% to 90% nucleotide identity with other isolates of SCMV. Phylogenetic analysis (maximum likelihood tree, with model HKY, 1000 bootstrap value) depicted that *Sugarcane mosaic virus* isolates clustered on host basis instead of geographical location, it is remarkable that 3' untranslated regions of SCMV isolates from the host maize consists of an additional stretch of twenty nucleotides. Two recombinant signals were also observed in NIa and NIb regions of SCMV-Sp genome.

Akbar *et al.,* 2017 reported minor sequence differences were depicted for SCMV isolates by phylogenetic tree throughout world. CP sequences from database were subjected to analysis to construct phylogenetic tree of which three clades are formed. Pakistani SCMV CP sequences were grouped into one clade depicting very less differences in sequences among each other. Formerly discovered two Pakistani SCMV CP sequences, AM040436 and DQ648195, were clustered with recent study isolates in combination with sequences gained from Australia, France, South Africa, China, Iran and Argentina. India is neighbor country as China, but Indian SCMV CP sequences attained a distinct place in the phylogenetic tree with bootstrap value of 98%. Two Sequences from USA did not fit in any clade and taken independent place in phylogenetic tree, probable reason could be due to their sampling from geographically apart areas of USA. Similarly, isolates from China were also distributed in separate clades, cluster 1 consisted two sequences and four sequences were lied into cluster 3, depicting their departed geographical origins. Study concluded that, Pakistani sugarcane crop has got infected by SCMV imported from Australia. Furthermore analysis represented that no noticeable variation was found among Pakistani isolates.

Several biologists agree that phylogenetic tree of relationships should be main underpinning of research in various fields of biology (Hall *et al.,* 2002a; Doyle *et al.,* 2003) furthermore phylogenetic framework can reveal the evolutionary patterns of various chemical and morphological characters (Soltis and Soltis, 2000; Daly *et al.,* 2001; imp of phylogeny 4). Moreover, origin, migration patterns of viruses and spread of viral infection are widely inferred by phylogenies (Qasim *et al.,* 2016; Guindon *et al.,* 2003; Drummond *et al.,* 2012).

2.3 Recombination

RNA viruses have enhanced mutation rate because of error prone replication accredited to absence of proof reading phenomena in RNA dependent RNA polymerase, big size of population and reduced generation time (Li *et al.,* 2012, Domingo and Holland, 1997). RNA recombination among closely linked or distantly related viruses, and even with host, RNA's has been predicted to be one of the special evolutionary power contributing towards divergence and evolution of genome for plant RNA viruses (Li *et al.,* 2012; Sztuba *et al.,* 2011). Consequently, high ability of diversity and genetic difference exits in RNA viruses. Plant RNA viruses have potential to acclimatize with changing environment counting new or resistant hosts due to their genetic diversity (Li *et al.,* 2012; Holmes, 2009).

Analysis of recombination, population demography, selection pressure and phylogeny lead us to know about genetic structure, samples selected were large number of host and geographically distinct SCMV isolates (3' UTR or CP cistron). Non-recombinant CP cistron was used to construct Neighbor Joining tree with bootstrap value of 1000. Six clades of SCMV population was formed, based on their genetic distances, which were associated with genuine hosts, according to previous research different groups were present within the common host (Li*et al.,* 2012; Wang *et al.,* 2010). Maize, *Musa textiles,* and sugarcane made one group except for group 6, which was combination of sugarcane and maize isolates. European and Asian isolates came from host maize form separate clade (group 5 and 6).

Achon *et al.,* 2007 reported about recombination of SCMV-Sp. It was found at location 5604 which is NIa-VPg portion and at 7866 base point which is NIb region with 0.32 and 0.46 co relation coefficient which is low value respectively, recombination was also observed at point 6432 which is NIa-Pro and 8560 which is CP part with r (coefficient value) greater than 0.6. Furthermore, two recombination events were observed by Zhong *et al* (Zhong *et al.,* 2005), for isolate GD. Recombination events were confirmed by using RDP2 program (Martin *et al.,* 2005). RDP, SISTERSCAN, BOOTSCAN, CHIMAERA, GENECONV, and MAXIMUM CHI SQUARE are the methods implemented in RDP program. Parameters selected were P value with cutoff at 0.05, thousand boot scan replicates, and recommended parameters were selected overall. Two recombination events were verified by RDP2 method. Recombination at location 5604 from the part 5151 to 5865 (NIa-VPg), with high value of confidence (P value 8.9518_19 from RDP). Second recombination event was observed at point 7866, from coding sequence NIb (7107 to 7953).

Maximum likelihood phylogenetic analysis was performed to co-relate recombination events with variation in tree topologies. Analysis was performed on either side of recombination break points observed in NIa and NIb, with aligned sequences of SCMV. The recombined point 5604 (of SCMV-Sp) at 5 prime terminal, found to be originated from isolates of maize closely matched with sequences of isolates SD and HN. The middle 2280 nucleotide were identical to isolates of sugarcane, such as LP, XGS and YH plus isolate of maize GD, moreover the 1730 base pair point located at 3 prime terminal matched to isolate of maize GD.

These results give proofs for presence of SCMV natural recombinants (Zhong *et al.,* 2005). Analysis depicted that isolates of SCMV are more linked to hosts compared to

geographical location with one exception which is Bris isolate. CP variation was linked to host from which isolates were taken (Achon *et al.,* 2003). Algeria *et al.,* 2003 showed that identity levels in CP's of SCMV isolates differ more among hosts relatively to geographic location. This study explained presence of recombination in SCMV and predicted a same origin of Asian and European SCMV isolates from sugarcane and maize.

Cheng *et al.,* 2017 performed recombination analysis by using RDP4 software. The methods implemented were RDP, GENECONV, BootScan, MaxChi, Chimaera, SiScan, 3Seq. Recombinant regions found for sequence DWK1 was from 2047 to 5138 nucleotides and 8177 nt to 8496 nt. Concluding that KU171814 (DWK1) was recombinant isolate of AF494510 (HN), AJ310105 (Guandong) and BD8 isolate.

Moradi *et al.,* 2016 had done recombination analysis NRA (KT895080) and ZRA (KT895081) Iranian isolates. Three and two intraspecific recombination events were detected by software RDP4 for NRA and ZRA respectively. For isolate NRA, total three recombination events were taken place, for recombination event number 1, major parent predicted by software was Bris A, isolate from Australia (AJ278405), minor parent Ohio, isolate from USA (JX188385), for recombination event number 2 and 3 major parent determined was ZRA (KT895081), with minor parents JX237863 and AJ278405 respectively. For isolate ZRA, two recombination events were observed, major and minor parents for recombination event number 1 was AJ278405 and NRA while for event number 2 Ohio (JX188385) and ARG-915 (JX237863) were observed as major and minor parents respectively.

2.4 Specie demarcation criteria methods comparison

In past, before discovery of sequence data, species and strains were discriminated with difficulty, criteria usually used were symptomatology, serology and host range. Accessibility of sequence data regarding full genome, individual genes has aided to clarify the taxonomy. Molecular criteria have been set for genus and species discrimination within the family *Potyviridae*. It involved two-way comparisons between full genome sequences and for every gene amongst all other complete sequences from the family. Comparisons (nucleotide and amino acid based) between sequences (full genome and gene wise) approach combined with phylogenetic analysis was used to re-examine the characterization of related strains and species.

Method involved complete set of pairwise comparison by using the amino acid and nucleic acid sequences of complete open reading frame from the family *Potyviridae*. GCG program was used for comparison, parameters applied for nucleotide-based comparison were GAP creation penalty of 50, gap extension penalty of 3, while for amino acid based comparison gap creation penalty used was 8 and gap extension penalty was 2. This program tends to align the selected sequences and find the similarity or percentage identity between sequences. Complete open reading frames were compared, individual genes taken from complete open reading frames were compared, and CP genes of all available sequences were compared.

Afterwards phylogenetic analysis was performed. Peptide sequences were firstly aligned by using ClustalX, alignment was trimmed to adjust poly protein cleavage sites and other known motifs. Corresponding nucleotide alignment was made by using former alignment as template, software used to generate corresponding alignment was trans align in EMBOSS. DNADIST was further used to generate phylogenetic tree with bootstrap value 1000.

Second method is solely based on true pairwise sequence alignments rather than global alignment based pairwise identities, with no gaps (sites with gaps should be excluded before). In this method full length genomic sequences of *Begomovirus* were downloaded from NCBI- Gen Bank. Using the data, Neighbor joining tree was made not to classify the *Begomoviruses*, but just to identify groups of most closely related sequences that could be joint to do pairwise sequence comparisons and to perform ML (maximum likelihood) phylogenetic analysis. Thirty eight groups were made in NJ phylogenetic tree, this strategy was applied to outline distinct groups. These thirty-eight groups (each) were subjected to pairwise sequence comparison separately by using SDT (Sequence Demarcation Tool). MUSCLE Alignment was used in SDT, maximum likelihood tree was constructed for each group by using PHYML 3.0 method applied in MEGA 5.2. Nucleotide substitution method used was GTR+I+G, with boot strap value of 3000.

Pairwise sequence comparison results were put into simulations by using varied cut off values to delineate capable species, to evaluate belonging of isolate, weather it belongs to same species or to distinct species. Parameters used for comparison are critical. SDT is software freely available to perform true pairwise sequence comparison excluding the sites with gap characters.

Chapter 3
MATERIALS & METHODS

MATERIALS AND METHODS

3.1 Data Collection

A total of 80 sequences (Complete genome) of *Sugarcane mosaic virus* were retrieved from publically available database National Center for Biotechnology Information (NCBI, GenBank). These sequences belonged to world's different countries and were renamed in a pattern as: Accession number: Country: Date.

Inclusion Criteria which was followed to select sequences included in the study was (1) sequences origin country should be known and clearly established; (2) sequences date of isolation or publication should be clearly present in genuine publication; (3) sequences can be recombinants. Sequences not following the criteria were removed.

3.2 Alignment

Data retrieved from NCBI was aligned using MEGA 7 software (https://www.megasoftware.net/, Kumar S, Stecher G, and Tamura K, 2016). Option Align was chosen from menu panel. Edit/Build Alignment was selected from drop down menu, FASTA file will be uploaded on new appeared window, MUSCLE alignment was chosen to perform analysis.

3.3 Pairwise Distance Matrix Analysis

Pairwise distance matrix were calculated by using software MatGat v2.01 (http://www.angelfire.com/nj2/arabidopsis/MatGAT.html , Campanella *et al.*, 2003). FASTA files were uploaded of full genome, CP and P1 sequences. Nucleotide-based and amino acid based pairwise distance matrix tables were calculated. Output file was saved as an excel spreadsheet.

3.4 Phylogenetic Reconstruction

Whole genome sequences retrieved from NCBI were aligned by Multiple Sequence Comparison by Log Expectation (MUSCLE alignment) in MEGA 7 software (https://www.megasoftware.net/, Kumar S, Stecher G, and Tamura K, 2016). Maximum Likelihood trees were inferred from MEGA 7 software by using Tamura-Nei model method, (https://www.megasoftware.net/, Kumar S, Stecher G, and Tamura K, 2016). Maize dwarf mosaic virus was used as an out group. Boot strap analysis was done by using 1000 replicates.

3.5 Bayesian Analysis

Bayesian analysis was conducted by using MrBayes 3.2.6 (Bayesian Inference of Phylogeny) (Guindons *et al*., 2012) program plugin available in Geneious v 8.1.5 program (http://www.geneious.com, Kearse *et al*., 2012) by using all the available full genome sequences of *Sugarcane mosaic virus*. Approach used during analysis was MCMC (Markov chain Monte Carlo). Default parameters were used as 1,100,000 chain length, 4 heated chains and gamma categories, gamma rate variation. The substitution model found fit in J Model Test v2.1.6 (Darriba *et al*., 2012; Guindon and Gascuel, 2003) according to Akaike information criterion was HKY85.

3.6 Recombination

Sequences were aligned by using MEGA software then recombination of aligned sequences were determined by using various methods implemented in the RDP4 (Martin *et al.*, 2015; http://web.cbio.uct.ac.za/~darren/rdp.html) software which includes RDP, CHIMERA, GENECONV, MAXCHI, SISCAN, and 3SEQ. Default parameters were applied throughout the analysis.

3.7 Sequence Demarcation Tool Analysis

Full length genome sequences retrieved from NCBI database and saved in FASTA file format were analyzed in SDT v 1.2 program (www.cbio.uct.ac.za/SDT, Muhire *et al*., 2014). Sub genome length sequences was already removed from the FAST file which was intended to use during analysis. In the dataset, identities between every pair of sequence was calculated by using MUSCLE alignment algorithm.

Chapter 4
RESULTS & DISCUSSION

Results and Discussion

4.1 Study Population

Full genome SCMV sequences were retrieved from NCBI. Whole genome sequences were checked, there was no need of manual editing. The study was performed using whole genome. To avoid unambiguous alignment other genes were not used because they were either highly conserved or too variable. There are total 80 sequences out of which one belongs to country Ecuador, two sequences from Ethiopia, Mexico and Iran, three nucleotide sequences belongs to Rwanda, while USA, Argentina, Australia, Spain and Germany shared one sequence each, all the remaining 65 sequences are from China. Sequences with information about country, accession number and year they isolated (descending order) are summarized in Table 4.1

Table 4.1: Sequences used for ML and Bayesian analysis of whole genome

S/No	Accession number	Country	Year	S/No	Accession number	Country	Year
1	KY548507	China	2018	41	JX047395	China	2010
2	KY548506	China	2018	42	GU474635	Mexico	2010
3	JX047383	China	2017	43	JX047410	China	2010
4	KY006657	Ecuador	2016	44	JX047409	China	2010
5	KP860935	Ethiopia	2014	45	JX047408	China	2010
6	KP860936	Ethiopia	2014	46	JX047407	China	2010
7	KR108212	China	2014	47	JX047388	China	2010
8	KR108213	China	2014	48	JX047413	China	2010
9	KT895081	Iran	2013	49	JX047398	China	2010
10	KT895080	Iran	2013	50	JX047389	China	2010
11	KF744391	Rwanda	2013	51	JX047387	China	2010
12	KF744390	Rwanda	2013	52	JX047412	China	2010
13	KF744392	Rwanda	2013	53	JX047411	China	2010
14	JX185303	Germany	2012	54	JX047386	China	2010
15	JX047418	China	2011	55	JX047385	China	2010
16	JX047420	China	2011	56	JX047381	China	2010

17	JX047421	China	2011	57	JX047390	China	2010
18	JX047427	China	2011	58	JX047382	China	2010
19	JX047419	China	2011	59	JX047396	China	2010
20	JX047422	China	2011	60	JX047392	China	2010
21	JX047417	China	2011	61	JX047403	China	2010
22	JX047423	China	2011	62	JX047402	China	2010
23	JX047431	China	2011	63	JX047401	China	2010
24	JX047424	China	2011	64	JX047405	China	2010
25	JX047414	China	2011	65	JX047400	China	2010
26	JX047425	China	2011	66	JX047391	China	2010
27	JX047429	China	2011	67	JX237862	Argentina	2010
28	JN021933	China	2011	68	JX047406	China	2010
29	JX047416	China	2011	69	AM110759	Spain	2008
30	JX047426	China	2011	70	JX237863	Argentina	2007
31	JX047415	China	2011	71	AJ297628	China	2005
32	JX047430	China	2011	72	AY569692	China	2005
33	JX047428	China	2011	73	AJ278405	Australia	2005
34	EU091075	Mexico	2011	74	AJ310105	China	2005
35	JX047393	China	2010	75	AJ310104	China	2005
36	JX047384	China	2010	76	AJ310103	China	2005
37	JX047397	China	2010	77	AJ310102	China	2005
38	JX047399	China	2010	78	AY042184	China	2004
39	JX047394	China	2010	79	AF494510	China	2003
40	JX047395	China	2010	80	JX188385	USA	1965

4.2 Analysis of *Sugarcane mosaic virus* sequences

All information about *Sugarcane mosaic virus* genome is gathered and organized in tabulated form. Table 4.2 contains information about all 79 available sequences on NCBI, full genome, start and end position of each gene, total number of nucleotides and amino acid present in *Sugarcane mosaic virus*. Total number of nucleotide bases present is SCMV varies from 9487 to 9640 nucleotides. Maximum number of SCMV consists of 9596 nucleotides, excepted few sequences which makes range from 9574 to 9578 and 9600 to 9640 nucleotides.

Untranslated region nucleotide bases ranges from 1 to 149 (5' UTR), 9342 to 9596 nucleotides for 3' UTR. As SCMV is poly protein so its starting and ending position varies from 125 to 9374 nucleotide bases, most of the sequences comes in a range of 149- 9341 except few sequences KP860936, JX185303, KYOO6657 and KF744390 with range 146 to 9382, 125-9316, 170 to 9385, 138 to 9374 respectively.

Gene wise distribution represented that P1 ranges from 149-847, except one sequence (KY006657) which shows variation and ranges from 170 to 868 nucleotide bases. Nucleotide bases starts and ends from 847 to 2229 for HC-Pro, 2228 to 3269 for P3, 3269 to 3470 for 6K1 and 3471 to 5384 for CI. However nucleic acid bases range varies from 5385 to 5543 for 6K2, 5544 to 6836 for NIa. NIb initiates from 6837 and ends at 8399 bases while coat proteins starting location is 8400 nucleotide base and ending point is at 9338 bases.

Total number of amino acid present in most of the SCMV is 3063 with few exceptions. Sequences KF744391, GU474635, and KP860936 have 3078 amino acids in total, while KP860935 and KY006657 have 3065 and 3071 amino acids respectively.

Table 4.2 Information about *Sugarcane mosaic virus* genome retrieved from databases.

S/No	Accession number	Start-stop position												Nucleotide	Amino acid
		Poly protein	5'UTR	P1	HC-Pro	P3	6K1	CI	6K2	NIa	NIb	CP	3'UTR		
1	AJ297628	150-9341	1-149	150-848	849-2228	2229-3269	3270-3470	3471-5384	5385-5543	5544-6836	6837-8399	8400-9338	9342-9596	9596bp	3063aa
2	NC_003398	150-9341	1-149	150-848	849-2228	2229-3269	3270-3470	3471-5384	5385-5543	5544-6836	6837-8399	8400-9338	9342-9596	9596bp	3063aa
3	AY569692	150-9341	1-149	153-848	849-2228	2229-3269	3270-3470	3471-5384	5385-5543	5544-6836	6837-8399	8400-9338	9342-9596	9596bp	3063aa
4	JX047393	149-9340	1-148	149-847	848-2227	2228-3268	3269-3469	3470-5383	5384-5542	5543-6835	6836-8398	8399-9337	9341-9595	9595bp	3063aa
5	JX047384	150-9341	1-149	150-848	849-2228	2229-3269	3270-3470	3471-5384	5385-5543	5544-6836	6837-8399	8400-9338	9342-9596	9596bp	3063aa
6	AY042184	149-9340	1-149	153-848	849-2228	2229-3269	3270-3470	3471-5384	5385-5543	5544-6836	6837-8399	8400-9338	9342-9596	9596bp	3063aa
7	JX047394	149-9340	1-148	149-847	848-2227	2228-3268	3269-3469	3470-5383	5384-5542	5543-6835	6836-8398	8399-9337	9341-9595	9595bp	3063aa

8	JX047417	149-9340	1-148	149-847	848-2227	2228-3268	3269-3469	3470-5383	5384-5542	5543-6835	6836-8398	8399-9337	9341-9574	9574bp	3063aa
9	JX047428	149-9340	1-148	149-847	848-2227	2228-3268	3269-3469	3470-5383	5384-5542	5543-6835	6836-8398	8399-9337	9341-9574	9574bp	3063aa
10	JX047418	149-9340	1-148	149-847	848-2227	2228-3268	3269-3469	3470-5383	5384-5542	5543-6835	6836-8398	8399-9337	9341-9574	9574bp	3063aa
11	JX047422	149-9340	1-148	149-847	848-2227	2228-3268	3269-3469	3470-5383	5384-5542	5543-6835	6836-8398	8399-9337	9341-9574	9574bp	3063aa
12	JX047419	149-9340	1-148	149-847	848-2227	2228-3268	3269-3469	3470-5383	5384-5542	5543-6835	6836-8398	8399-9337	9341-9574	9574bp	3063aa
13	JX047427	149-9340	1-148	149-847	848-2227	2228-3268	3269-3469	3470-5383	5384-5542	5543-6835	6836-8398	8399-9337	9341-9574	9574bp	3063aa
14	JX047423	149-9340	1-148	149-847	848-2227	2228-3268	3269-3469	3470-5383	5384-5542	5543-6835	6836-8398	8399-9337	9341-9576	9576bp	3063aa
15	AF494510	150-9341	1-149	150-848	849-2228	2229-3269	3270-3470	3471-5384	5385-5543	5544-6836	6837-8399	8400-9338	9342-9596	9596bp	3063aa
16	JX047420	150-9341	1-149	150-848	849-2228	2229-3269	3270-3470	3471-5384	5385-5543	5544-6836	6837-8399	8400-9338	9342-9575	9575bp	3063aa
17	GU474635	150-9386	1-149	150-848	849-2228	2229-3269	3270-3470	3471-5384	5385-5543	5544-6836	6837-8399	8400-9383	9387-9628	9628bp	3078aa
18	JX047395	149-9340	1-148	149-847	848-2227	2228-3268	3269-3469	3470-5383	5384-5542	5543-6835	6836-8398	8399-9337	9341-9595	9595bp	3063aa
19	AM110759	150-9341	1-149	153-848	849-2228	2229-3269	3270-3470	3471-5384	5385-5543	5544-6836	6837-8399	8400-9338	9342-9596	9596bp	3063aa
20	JX047421	159-9350	1-158	159-857	858-2237	2238-3278	3279-3479	3480-5393	5394-5552	5553-6845	6846-8408	8409-9347	9351-9584	9584bp	3063aa
21	EU091075	150-9341	1-149	150-848	849-2228	2229-3269	3270-3470	3471-5384	5385-5543	5544-6836	6837-8399	8400-9338	9342-9583	9583bp	3063aa
22	KP860936	146-9382	1-145	146-847	848-2217	2228-3258	3269-3459	3470-5373	5384-5532	5543-6825	6836-8388	8399-9327	9383-9640	9640bp	3078aa
23	KP860935	145-9342	1-144	145-847	848-2217	2228-3258	3269-3459	3470-5373	5384-5532	5543-6825	6836-8388	8399-9327	9343-9600	9600bp	3065aa
24	JX047399	150-9341	1-149	150-848	849-2228	2229-3269	3270-3470	3471-5384	5385-5543	5544-6836	6837-8399	8400-9338	9342-9596	9596bp	3063aa
25	JX047397	149-9340	1-148	149-847	848-2227	2228-3268	3269-3469	3470-5383	5384-5542	5543-6835	6836-8398	8399-9337	9341-9595	9595bp	3063aa
26	AY149118	150-9341	1-149	150-848	849-2228	2229-3269	3270-3470	3471-5384	5385-5543	5544-6836	6837-8399	8400-9338	9342-9596	9596bp	3063aa
27	JX185303	125-9316	1-124	125-823	824-2203	2204-3244	3245-3445	3446-5359	5360-5518	5519-6811	6812-8518	8519-9313	9342-9596	9549bp	3063aa
28	JX047409	150-9341	1-149	150-848	849-2228	2229-3269	3270-3470	3471-5384	5385-5543	5544-6836	6837-8399	8400-9338	9342-9577	9577bp	3063aa
29	JX047408	151-9342	1-150	151-849	850-2229	2230-3270	3271-3471	3472-5385	5386-5544	5545-6837	6838-8400	8401-9339	9343-9578	9578bp	3063aa
30	JX047381	149-9340	1-148	149-847	848-2227	2228-3268	3269-3469	3470-5383	5384-5542	5543-6835	6836-8398	8399-9337	9341-9576	9576bp	3063aa
31	JX047382	149-9340	1-148	149-847	848-2227	2228-3268	3269-3469	3470-5383	5384-5542	5543-6835	6836-8398	8399-9337	9341-9576	9576bp	3063aa
32	JX047383	149-9340	1-148	149-847	848-2227	2228-3268	3269-3469	3470-5383	5384-5542	5543-6835	6836-8398	8399-9337	9341-9576	9576bp	3063aa
33	JX047407	149-9340	1-148	149-847	848-2227	2228-3268	3269-3469	3470-5383	5384-5542	5543-6835	6836-8398	8399-9337	9341-9576	9576bp	3063aa
34	JN021933	149-9340	1-148	149-847	848-2227	2228-3268	3269-3469	3470-5383	5384-5542	5543-6835	6836-8398	8399-9337	9341-9576	9576bp	3063aa
35	JX047389	149-9340	1-148	149-847	848-2227	2228-3268	3269-3469	3470-5383	5384-5542	5543-6835	6836-8398	8399-9337	9341-9576	9576bp	3063aa
36	JX047387	149-9340	1-148	149-847	848-2227	2228-3268	3269-3469	3470-5383	5384-5542	5543-6835	6836-8398	8399-9337	9341-9576	9576bp	3063aa
37	JX047385	149-9340	1-148	149-847	848-2227	2228-3268	3269-3469	3470-5383	5384-5542	5543-6835	6836-8398	8399-9337	9341-9576	9576bp	3063aa
38	JX047388	149-9340	1-148	149-847	848-2227	2228-3268	3269-3469	3470-5383	5384-5542	5543-6835	6836-8398	8399-9337	9341-9576	9576bp	3063aa
39	JX047413	149-9340	1-148	149-847	848-2227	2228-3268	3269-3469	3470-5383	5384-5542	5543-6835	6836-8398	8399-9337	9341-9576	9576bp	3063aa
40	JX047400	149-9340	1-148	149-847	848-2227	2228-3268	3269-3469	3470-5383	5384-5542	5543-6835	6836-8398	8399-9337	9341-9576	9576bp	3063aa
41	KY006657	170-9385	1-169	170-868	869-2248	2249-3289	3290-3490	3491-5404	5405-5563	5564-6856	6857-8419	8420-9382	9386-9632	9632bp	3071aa
42	JX047398	149-9340	1-148	149-847	848-2227	2228-3268	3269-3469	3470-5383	5384-5542	5543-6835	6836-8398	8399-9337	9341-9576	9576bp	3063aa
43	JX047402	149-9340	1-148	149-847	848-2227	2228-3268	3269-3469	3470-5383	5384-5542	5543-6835	6836-8398	8399-9337	9341-9576	9576bp	3063aa
44	JX047401	149-9340	1-148	149-847	848-2227	2228-3268	3269-3469	347-5383	5384-5542	5543-6835	6836-8398	8399-9337	9341-9576	9576bp	3063aa
45	JX047403	149-9340	1-148	149-847	848-2227	2228-3268	3269-3469	3470-5383	5384-5542	5543-6835	6836-8398	8399-9337	9341-9576	9576bp	3063aa
46	JX047430	149-9340	1-148	149-847	848-2227	2228-3268	3269-3469	3470-5383	5384-5542	5543-6835	6836-8398	8399-9337	9341-9576	9576bp	3063aa
47	JX047391	149-9340	1-148	149-847	848-2227	2228-3268	3269-3469	3470-5383	5384-5542	5543-6835	6836-8398	8399-9337	9341-9576	9576bp	3063aa
48	JX047386	149-9340	1-148	149-847	848-2227	2228-3268	3269-3469	3470-5383	5384-5542	5543-6835	6836-8398	8399-9337	9341-9576	9576bp	3063aa
49	JX047416	149-9340	1-148	149-847	848-2227	2228-3268	3269-3469	3470-5383	5384-5542	5543-6835	6836-8398	8399-9337	9341-9576	9576bp	3063aa
50	JX047429	149-9340	1-148	149-847	848-2227	2228-3268	3269-3469	3470-5383	5384-5542	5543-6835	6836-8398	8399-9337	9341-9576	9576bp	3063aa
51	JX047424	149-9340	1-148	149-847	848-2227	2228-3268	3269-3469	3470-5383	5384-5542	5543-6835	6836-8398	8399-9337	9341-9576	9576bp	3063aa

52	JX047392	149-9340	1-148	149-847	848-2227	2228-3268	3269-3469	3470-5383	5384-5542	5543-6835	6836-8398	8399-9337	9341-9576	9576bp	3063aa
53	JX047412	149-9340	1-148	149-847	848-2227	2228-3268	3269-3469	3470-5383	5384-5542	5543-6835	6836-8398	8399-9337	9341-9576	9576bp	3063aa
54	JX047425	149-9340	1-148	149-847	848-2227	2228-3268	3269-3469	3470-5383	5384-5542	5543-6835	6836-8398	8399-9337	9341-9576	9576bp	3063aa
55	JX047415	149-9340	1-148	149-847	848-2227	2228-3268	3269-3469	3470-5383	5384-5542	5543-6835	6836-8398	8399-9337	9341-9576	9576bp	3063aa
56	JX047414	149-9340	1-148	149-847	848-2227	2228-3268	3269-3469	3470-5383	5384-5542	5543-6835	6836-8398	8399-9337	9341-9576	9576bp	3063aa
57	JX047410	149-9340	1-148	149-847	848-2227	2228-3268	3269-3469	3470-5383	5384-5542	5543-6835	6836-8398	8399-9337	9341-9576	9576bp	3063aa
58	JX047431	149-9340	1-148	149-847	848-2227	2228-3268	3269-3469	3470-5383	5384-5542	5543-6835	6836-8398	8399-9337	9341-9576	9576bp	3063aa
59	JX047390	150-9341	1-149	150-848	849-2228	2229-3269	3270-3470	3471-5384	5385-5543	5544-6836	6837-8399	8400-9338	9342-9577	9577bp	3063aa
60	AJ310105	150-9341	1-149	153-848	849-2228	2229-3269	3270-3470	3471-5384	5385-5543	5544-6836	6837-8399	8400-9338	9342-9596	9596bp	3063aa
61	JX047411	149-9340	1-148	149-847	848-2227	2228-3268	3269-3469	3470-5383	5384-5542	5543-6835	6836-8398	8399-9337	9341-9576	9576bp	3063aa
62	JX047405	149-9340	1-148	149-847	848-2227	2228-3268	3269-3469	3470-5383	5384-5542	5543-6835	6836-8398	8399-9337	9341-9576	9576bp	3063aa
63	JX047426	149-9340	1-148	149-847	848-2227	2228-3268	3269-3469	3470-5383	5384-5542	5543-6835	6836-8398	8399-9337	9341-9576	9576bp	3063aa
64	JX047396	149-9340	1-148	149-847	848-2227	2228-3268	3269-3469	3470-5383	5384-5542	5543-6835	6836-8398	8399-9337	9341-9576	9576bp	3063aa
65	AJ310104	150-9341	1-149	153-848	849-2228	2229-3269	3270-3470	3471-5384	5385-5543	5544-6836	6837-8399	8400-9338	9342-9575	9575bp	3063aa
66	JX047406	149-9340	1-148	149-847	848-2227	2228-3268	3269-3469	3470-5383	5384-5542	5543-6835	6836-8398	8399-9337	9341-9576	9576bp	3063aa
67	AJ310103	150-9341	1-149	153-848	849-2228	2229-3269	3270-3470	3471-5384	5385-5543	5544-6836	6837-8399	8400-9338	9342-9575	9575bp	3063aa
68	KF744390	138-9374	1-137	138-837	848-2217	2228-3258	3269-3459	3470-5373	5384-5532	5543-6825	6836-8388	8399-9327	9341-9566	9552bp	3078aa
69	AJ310102	150-9344	1-149	153-848	849-2228	2229-3269	3270-3470	3471-5384	5385-5543	5544-6836	6837-8399	8400-9341	9345-9578	9578bp	3064aa
70	KF744391	134-9370	1-133	138-833	848-2213	2228-3254	3269-3455	3470-5369	5384-5528	5543-6821	6836-8384	8399-9323	9341-9562	9607bp	
71	KR108212	148-9339	1-148	149-847	848-2227	2228-3268	3269-3469	3470-5383	5384-5542	5543-6835	6836-8398	8399-9337	9341-9576	9570bp	3063aa
72	JX188385	150-9386	1-149	150-848	849-2228	2229-3269	3270-3470	3471-5384	5385-5543	5544-6836	6837-8546	8547-9383	9342-9575	9613bp	3078aa
73	KF744392	143-9379	1-142	143-841	847-2221	2227-3262	3268-3463	3469-5377	5383-5536	5542-6829	6835-8392	8398-9333	9343-9568	9487bp	3078aa
74	AJ278405	148-9342	1-147	151-846	847-2226	2227-3267	3268-3468	3469-5382	5383-5541	5542-6834	6835-8397	8398-9339	9343-9573	9589bp	3064aa
75	JX237862	148-9339	1-147	148-846	847-2226	2227-3267	3268-3468	3469-5382	5383-5541	5542-6834	6835-8397	8398-9336	9340-9571	9571bp	3063aa
76	KT895081	150-9341	1-149	153-848	849-2228	2229-3269	3270-3470	3471-5384	5385-5543	5544-6836	6837-8399	8400-9338	9000-9341	9572bp	3063aa
77	JX237863	148-9339	1-147	148-846	847-2226	2227-3267	3268-3468	3469-5382	5383-5541	5542-6834	6835-8397	8398-9336	9340-9576	9576bp	3063aa
78	KT895080	150-9341	1-149	153-848	849-2228	2229-3269	3270-3470	3471-5384	5385-5543	5544-6836	6837-8399	8400-9338	9000-9341	9571bp	3063aa
79	KR108213	148-9339	1-147	151-846	847-2226	2227-3267	3268-3468	3469-5382	5383-5541	5542-6834	6835-8397	8398-8999	9000-9339	9573bp	3063aa

4.3 Pairwise Distance Matrix Analysis

4.3.1 Nucleotide-based Pairwise Distance Matrix Analysis of CP Gene

Nucleotide-based pairwise distance matrix of CP gene analysis has shown sequence identity among sequences belongs to same and different countries.

Nucleotide-based study has revealed sequence identities between sequences of different countries which is as follows, USA sequences has high similarity with sequences of Germany (85.7%), Mexico (79.8% to 81.8%), Ethiopia (77.7% to 81.5%), Rwanda (69% to 81.1%), Ecuadorian, Iranian, Spain and Argentine sequences shares identity of 76% to 78.3%. Mexican sequences have percentage similarity of 73.7% to 80.7% with German sequences, 83.6% to 87.5% identity with Ethiopian sequences, Germany shares similarity of 71.1% to 74.5% with Ethiopian sequences, 74.8% with Argentine nucleotide bases, 79% with

Spain, 74.5% identity with sequences of Ecuador, 73% and 62.4% to 73.6% resemblance with bases of Iran and Rwanda. Ethiopian sequences shared highest similarity with sequences of Rwanda 82% to 94.5%, it shares high percentage identity with Ecuador (80.9% to 82.1%), Iran 79.9% to 81.1%, Argentina 79.7% to 82.5%, Spain 79.8% to 83.3%. Iran is similar to Argentina 89.9% to 94.1%, Ecuador resembles Spain 81.5% and Iran, Rwanda, and Argentina have percentage similarity of 79.5% to 80.3%, 71.5% to 81.3% and 81% with Spain respectively. Mexico and Spain sequences are highly identical (91.5%), Ecuador and Mexico nucleotide bases are 81.8% similar, Mexico and Argentina resemble 78.7% to 82.4%, Mexico shows low similarity with Iranian sequences (76.3% to 77), Ecuador shared percentage identity of 76.4% to 77.1% with Iran, 71.3% to 81.9% with Rwanda.

Chinese sequences show maximum identity with sequences of Spain (99.6%), Iran (97.7%), Argentina (95.5%), Ecuador (92%), Ethiopia (91.1%), Mexico (91.7%), while for other countries (Rwanda, USA, Germany) percentage varies from 72.3% 83.8%. Nucleotide-based minimum and maximum sequence identities between sequences of same countries ranges as follows, nucleotide bases from Ethiopia shares identity of 91.1% to 100% with each other, and Iranian sequences 97.7% to 100%. Sequences from Rwanda were 82% to 100% identical to each other, Argentina, Mexico and China sequences were 92.2% to 100%, 83.7% to 100% and 70% to 100% identical respectively. Percentage similarity analysis matrix of SCMV CP sequences is given at the end.

4.3.2 Amino acid Based Pairwise Distance Matrix Analysis of CP Gene

Amino acid based pairwise distance matrix of CP gene analysis has shown sequence identity among sequences belongs to same and different countries.

Amino acid-based study has revealed sequence identities between sequences of different countries which are as follows, Germany highly resembled USA (91%), Mexico, Ethiopia, Ecuador, Iran and Argentina shared identity with sequence of USA in a range of 81.2% to 86.3%. Mexico has shown maximum identity with Spain (97.4%), Rwanda (93.6%) Ethiopia (91.5%), Iran (90.7%) and Argentina (90.4%) while its sequence similarity with Germany, Ecuador ranges 77.7% to 87.7%, Resemblance range of German sequences lies from 77.1% to 83.1% with countries Ethiopia, Argentina, Ecuador, Iran, Spain and Rwanda to 80.6%. Ethiopian sequences showed 86% to 87.3% identity with sequences of Argentina, Ecuadorian and Argentine sequences are 84.4% identical, Iranian and Argentine sequences shared highest resemblance (94.6% to 98.1%), Spain sequences have percentage similarity of

86.6% to 88.6% with sequences of Ecuador, Spain and Ethiopia. Iran has shown maximum resemblance with sequences of Spain (90.7%), Rwanda and Spain are 84.7% to 89.5% alike, USA and Spain share relatively low percentage of similarity (85.6%), Argentina and Spain share high percentage of identity (90.7% to 91.4%). Ethiopian nucleotide bases resembled 86.3% to 87.9%, with Ecuador, Ethiopia and Iran sequences. Ethiopia and Rwanda showed maximum resemblance 88.9% to 94.9.

Chinese sequences showed highest identity with sequences of Argentina (79% to 98%), Spain (87.5% to 97.4%), Mexico (87% to 96%), they show high sequence similarity with sequences of Ecuador (81% to 92%), Ethiopia (82% to 90%), Rwanda (86% to 90%), Iran (87% to 90%) and relatively low resemblance with sequences of USA 83% to 85% and Germany 80% to 82.7%,

Nucleotide-based minimum and maximum sequence identities between sequences of same countries ranges as follows, sequences from Ethiopia country shows resemblance of 91.1% to 100%, from Iran they show similarity of 97.1 to 100%, from country Rwanda sequences shows similarity of 92.1% to 100%, Argentina sequences are highly identical (96.8% to 100%) to each other, nucleotide bases from Mexico resembled 85.4% to 100%. Percentage similarity analysis matrix of SCMV CP sequences is given at the end.

4.4 Maximum Likelihood Phylogenetic Analysis of CP Gene

According to phylogenetic analysis of CP gene of *Sugarcane mosaic virus* tree as shown in Figure 4.4.1 has been divided into four clades, named as clade 1, 2, 3 and 4. Clade 1 consists of sequences belong to China except one which belongs to Ecuador present in clade 1. Clade 2 consists one sequence each from Mexico, Germany and Spain while rest of sequences in respective clade belongs to China. Clade 3 consists of seven sequences, three from Rwanda, one sequence from Mexico, USA, China, each while two sequences from Ethiopia. Clade 4 consists of sequences from Argentina, Iran and China. CP gene of Maize dwarf mosaic virus is used as an out group.

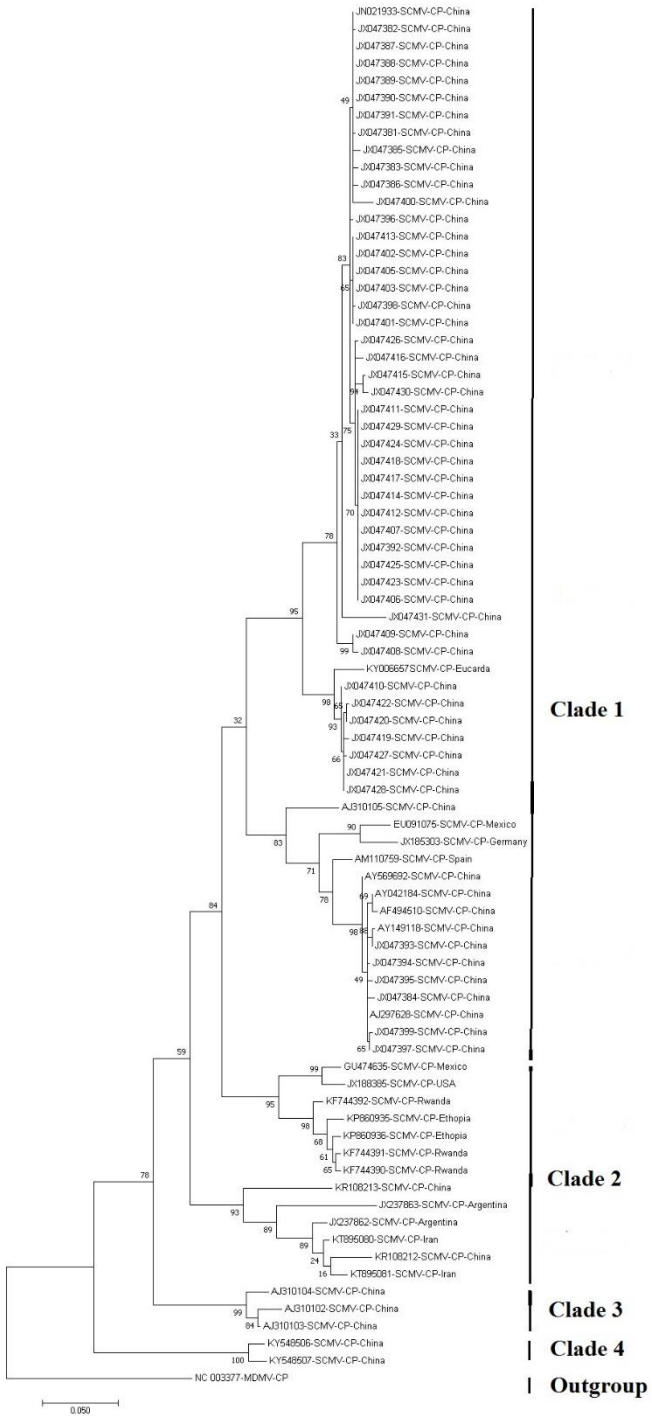


Figure 4.4.1. A maximum likelihood phylogenetic tree estimated using CP gene of *Sugarcane mosaic virus*. Each branch is led by a virus and is denoted by following format, GenBank accession number_ gene name _Country. Four clades 1, 2, 3 and 4 are made. Branches are well supported by bootstrap values. CP gene of Maize dwarf mosaic virus is used as an out group.

4.5 Pairwise Distance Matrix Analysis of P1 Gene

4.5.1 Nucleotide-based PWD Analysis of P1

Nucleotide-based pairwise distance matrix of P1 analysis has shown sequence identity among sequences belongs to same and different countries.

Nucleotide-based study has revealed sequence identities between sequences of different countries which are as follows, USA and Mexico shares the maximum similarity percentage 90.7%, sequences of countries Rwanda and Iran shares identity ranges from 87.4% to 86.8%,USA sequences shows minimum identity with nucleotide bases of Germany (70%), Ethiopia (68%) and Argentina (67%), P1 sequences from Mexico shows high identity with Ethiopian sequences (93.6%) and Germany sequences (87.8% to 89.8%), Spain (89%) while it has shown less identity with sequences of Argentina (69.8%) and Australia (70%). Germany resembles Ethiopia 88%, 70% with Argentina, Ethiopia has percentage identity of 69.3% to 70.1% with Argentina. Iran is similar to Argentina 66.2% to 66.7%, Germany, Ethiopia, Iran, Rwanda and USA shares alike-ness of 66% to 69% with Australia. Resemblance of Spain with countries Germany, Ethiopia, Australia, Iran, Rwanda, USA and Argentina ranges from 67% to 92.1%. Germany is 69.7% with Iran, Mexico shared identity of 70.2% to 71.1% with Iran, and Ethiopia percentage identity is 70.7% with Iran. Germany is likely to have less similarity with sequences of Rwanda (69.1% to 69.8%), Mexico and Rwanda were appeared to have 68.4% to 70.1% resemblance, and Ethiopia and Argentina similarity with Rwanda sequences were 68.2% to 69% and 67.8% to 69.5% respectively.

Chinese sequences have shown highest identity with sequences of Australia (96%), in case of other countries it has similarity, range varies as for Ethiopia it is 78.4% to 83%, for Spain resemblance is 62.7% to 89.6%, for Mexico similarity is 62.9% to 92.7%, for Argentina range differs from 59.8% to 94.1%, for USA alike ness is 60% to 70%, for Ecuador percentage identity is 79% to 95.4%.

Nucleotide-based minimum and maximum sequence identities between sequences of same countries ranges as follows, sequences belongs to Ethiopia have similarity 98.1% to 100%, Iranian sequences have resemblance of 98% to 100%, Rwanda sequences match to each other 94.7% to 100%, Argentina with little bit differences have resemblance 98% to 100%, Mexico sequences have identity range of 97.7% to 100%, Chinese sequences range is broad, starting from 55% till 100%. Percentage similarity analysis matrix of SCMV P1 sequences is given at the end.

4.5.2 Amino acid Based Pair Wise Distance Matrix Analysis of P1

Amino acid based pairwise distance matrix of P1 analysis has shown sequence identity among sequences belongs to same and different countries.

Nucleotide-based study has revealed sequence identities between sequences of different countries which are as follows, Germany has shown maximum similarity with Ethiopia (95.3%), Spain (95.7%), Ethiopia (95.3%), for other countries identity ranges from 79.4% to 81.5%USA 80.7%. Mexican sequence shows alike-ness of 82% to 82.4% with sequences of USA, Ethiopia, Iran, Argentina shares identity of 79.8%, 93.1% and 83.7% to 84.1% with USA respectively. Argentina sequences shares identity of 80.7%, 80.6% and 83.2% to 84.1% with Mexico, Ethiopia and Iran sequences. Australian sequences shows maximum resemblance with sequences of Australia (98.3% to 99.1%), in case of other countries it shares median percentage identity Mexico (81.1%), Ethiopia (81%), Iran (83.2% to 83.6), USA (84.1%). Sequences from Mexico is similar to Spain sequences 93.6% to 95.3%, Ethiopia shows resemblance of 94.8% with Spain, Australia shares identity with Spain 82.8%, Spain shares similarity of 80.6% with Iran, 81.5% with USA, 83% with Argentina, Mexican sequences shares percentage similarity of 79.8% to 80.7% with Iran.

Chinese sequences share identity ranges of 72.4% to 95%, 74.2% to 96.6%, 74% to 80%, 70.8% to 98.7%, 71.2% to 97.5%, 72.5% to 84.1%, 96% to 99% and 78% to 80% with Ethiopia, Spain, Mexico, Australia, Argentina, USA, Rwanda and Iran respectively. Nucleotide-based minimum and maximum sequence identities between sequences of same countries ranges as follows, Mexican sequences with each other 98.3% to 100%, Chinese sequences within country 72% to 100%. Percentage similarity analysis matrix of SCMV P1 sequences is given at the end.

4.6 Maximum Likelihood Phylogenetic Analysis of P1 Gene

Phylogenetic analysis of P1 gene has estimated tree by using maximum likelihood method. Tree as shown in Figure 4.6.1, is divided into three clades named as1, 2 and 3. Clade 1 consists of sequences belongs to China, Ecuador, Iran, USA, Rwanda and Argentina. Clade 2 consists of sequences from various countries, including China, Germany, Spain, Ethiopia and Mexico. Clade 3 last and the smallest clade which includes two just two sequences both from China. Maize dwarf mosaic virus P1 gene is used as an out group

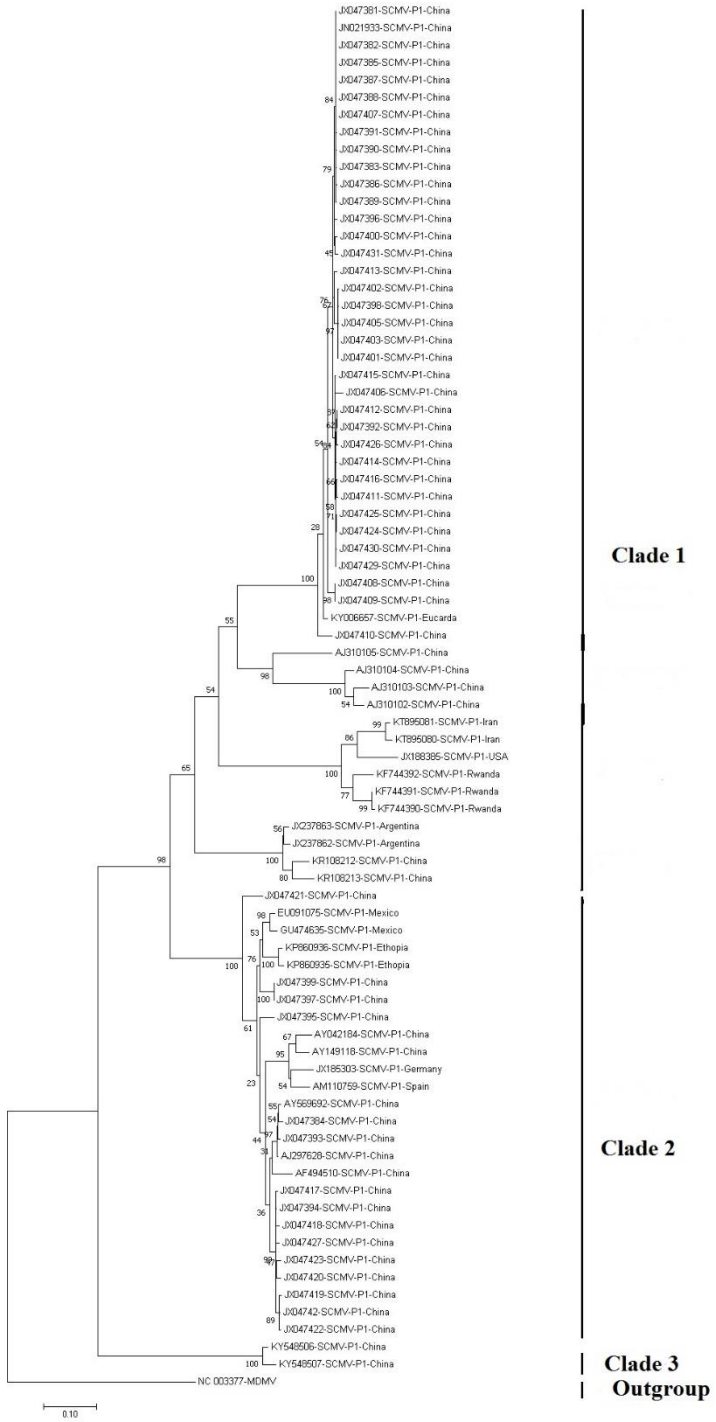


Figure 4.6.1. A maximum likelihood phylogenetic tree estimated using P1 gene of *Sugarcane mosaic virus*. Each branch is led by a virus and is denoted by following format, GenBank accession number_ name of virus_ gene name _Country. Three clades named as 1, 2 and 3 are made. Branches are well supported by bootstrap values. P1 gene of Maize dwarf mosaic virus is used as an out group.

4.7 Pairwise Distance Matrix Analysis of Full Genome

4.7.1 Nucleotide-based Pairwise Distance Matrix Analysis of Full Genome

Nucleotide-based pairwise distance matrix of full genome analysis has shown sequence identity among sequences belongs to same and different countries.

Nucleotide-based study has revealed sequence identities between sequences of different countries which are as follows, Germany has shown similarity of 78.7% with sequences of USA, Mexico and USA are 79% identical, identity percentage between Ethiopia and USA ranges from 80.1% to 82.6%, similarity percentage of sequences belongs to USA with other countries as Ecuador, Iran, Argentina is 79% and 91% with sequences of Rwanda. Range of percentage between Mexico and Germany ranges from 92.1% to 96.2%. SCMV sequences from Ethiopia has shown minimum to maximum identity of 78.8% to 87% with the sequences of countries, Mexico, Australia, Argentina, Spain, Ecuador, Iran, Germany and Rwanda. Nucleotide-based resemblance of sequences belongs to Germany, Mexico, Ecuador, Spain, Rwanda is from 78.3% to 79.6% with the sequences of Argentina, while sequences from Australia and Iran have shown high identity with nucleotide bases of Argentina with percentages of 95.4% and 91.5% respectively. Australian sequences have shown maximum similarity with sequences of Iran which is 91.9%, other identity range is between 78.1 % to 79.2% with sequences of countries Germany, Mexico, Ecuador, Rwanda and USA. Resemblance between sequences of Germany- Spain, Mexico-Spain, Australia- Spain, Ecuador- Spain, Iran- Spain, Rwanda- Spain, and USA- Spain is 93.7%, 92.7%, 79.3%, 79.4%, 79.6%, 78.2% to 79% and 79% respectively. Germany nucleotide bases has shown 78.4% to 79.4% identity with sequences belongs to Iran and Rwanda. Range of Mexican sequences similarity varies from 78.9% to 79.7% with Iranian, Rwanda and Ecuadorian nucleotide bases. Sequence of Ecuador has shown identity of 79.3% with Iranian sequences and 78.9% to 80% similarity with nucleic acid bases of Rwanda.

Chinese sequences have shown high similarity with sequences of Iran (93.2% to 98%), Ecuador (93 % to 99%), identity with Ethiopia, Mexico, Australia, Rwanda, Spain, and USA sequences varies from 76.1% to 81%. Nucleotide-based minimum and maximum sequence identities between sequences of same countries ranges as follows, Ethiopian sequences 96.7% to 100%, nucleotide bases of Iran sequences 97.3 to 100%, within Rwanda

sequences 95.1% to 100%, percentage similarity shown by Argentina sequences with each other 95.7% to 100%, Chinese sequences resemblance ranges from 76% to 100%.

Table 4.7.1: Percentage identities present between full genome sequences of SCMV

	Group A (39 sequences)	Group B (23 sequences)	Group C (15 sequences)	Group D (2 sequences)
Group A	87% to 100%	79% to 84%	79% to 83%	76% to 77%
Group B		84% to 100%	79% to 84%	76% to 77%
Group C			80% to 100%	76% to 77%
Group D				98% to 100%

. Four groups with percentage identity they show within same group and with different groups.

4.7.2 Amino Acid based PWD Analysis of Full Genome

Amino acid based pairwise distance matrix of full genome analysis has shown sequence identity among sequences belongs to same and different countries.

Amino acid-based study has revealed sequence identities between sequences of different countries which are as follows, USA nucleotide bases has shown high similarity with sequences of Germany (95.1%), Mexico (95%), Ethiopia (96.2% to 97.2%), Ecuador (95.4%), Iran (96%) and Rwanda (98.4%), Mexico and Germany sequences resembled 97.7% to 98.9%, Germany and Ethiopian sequences 97.4%, Germany and Argentina nucleotide bases resembles 95.8%. Percentage similarity of Mexican sequences with sequences of Ethiopia and Argentina is 97.2% and 95.5%. Argentina sequences range of similarity lies from 95% to 99% with sequences of Ethiopia, Australia, Spain, Rwanda, Ecuador and Iran. Australian sequences have shown maximum identity with sequences of Iran (98.1%) and with other sequences range differ from 94.9% to 95.2%. Identity as high as 99.2% exists between sequences of Germany and Spain, 98.9% between Mexico and Spain, 97.3% among nucleotide bases of Ethiopia and Spain and lower range is present between sequences of Australia versus Spain 95.7%, Ecuador and Spain 95.8%, Iran versus Spain 95.7%, Rwanda and Spain 95.6%, USA versus Spain 95.4%. Resemblance of Ecuadorian sequences with other nucleic acid bases (Germany, Mexico and Ethiopia) ranges from 95% to 95.6%. Identity of German, Mexican, Ethiopian, Ecuadorian sequences with Iran sequences is 95.7%, 95.6%, 95%, 95% respectively. Range for sequences of Rwanda is 95% to 96.6% for sequences belongs to country Germany, Mexico, Ethiopia, Ecuador.

Nucleotide-based similarity among Chinese sequences and sequences of Ethiopia, Spain, Mexico, Australia, Rwanda, USA, Iran, Ecuador, and Germany lies between 92.8% to 99%, with maximum similarity with Ecuador (99%) and lowest resemblance with Rwanda (92.8%). Nucleotide-based lowest and highest sequence identities between sequences of same countries ranges as follows, similarity exists within sequences of Ethiopia 99.8% to 100%, Iran 99.4 to 100%, China 93% to 99.6% and Mexico,98.1% to 100%. Percentage similarity analysis matrix of SCMV full genome sequences is given at the end.

4.8 Phylogenetic Inference

Maximum likelihood analysis was performed to construct the phylogeny of *Sugarcane mosaic virus* using full genome. The evolutionary model used was Tamura Nei model. Figure 4.8.1 shows the resulting phylogenetic tress of full genome. Branches of tree are well supported by bootstrap values. Tree divides *Sugarcane mosaic virus* into four different clades named as A, B, C and D.

Clade A consists of most sequences, all belongs to China except one which belongs to Ecuador. Sequence of Ecuador shared identity of 94% to 100% with Chinese sequences, however with all other sequences (belonging to different countries) its percentage similarity ranges from 79 % to 80% which is relatively very less. Clade B consists of six sequences, *Sugarcane mosaic virus* in this clade belongs to different geographical origins/countries, like Mexico, Spain, Germany, and Ethiopia. Clade C is second largest group after clade A, it shows that SCMV is polyphyletic in origin, as sequences in this group belongs to China, Rwanda, USA, Iran and Argentina. Clade D is made up of just two sequences (KY548507, KY548506) belongs to country China.

Clade B consists of isolates related to different countries but taken from same host (maize), it reflects that co-relation between SCMV more depends on host rather than geographical location validated by Achon *et al.,* 2007 and Moradi *et al.,* 2016, Bo *et al.,* 2011. The 5’ untranslated region and P1 part of Iranian isolates from Clade C were very similar to strain from USA and isolated from host maize, the HC-Pro, NIa-Vpg, P3, CP and NIb portion was more identical to Australian isolate, remaining 5 proteins and 3’ untranslated region showed resemblance to strains from Argentina. In case of Iranian isolates strains from different countries as from Australia and Argentina are cladded together, secondly isolates taken from different hosts (maize and sugarcane) were clustered together, which shows that type of host or geographical location was not classifying factor in Iranian isolates, rather it

was recombination, which is also a main driving power in virus evolution (Moradi *et al.*, 2016). Clade D two sequences shared 98% to 100% nucleotide base percentage identity with each other, in case of all other sequences they show very less nucleotide-based similarity which is 76% to 77%.

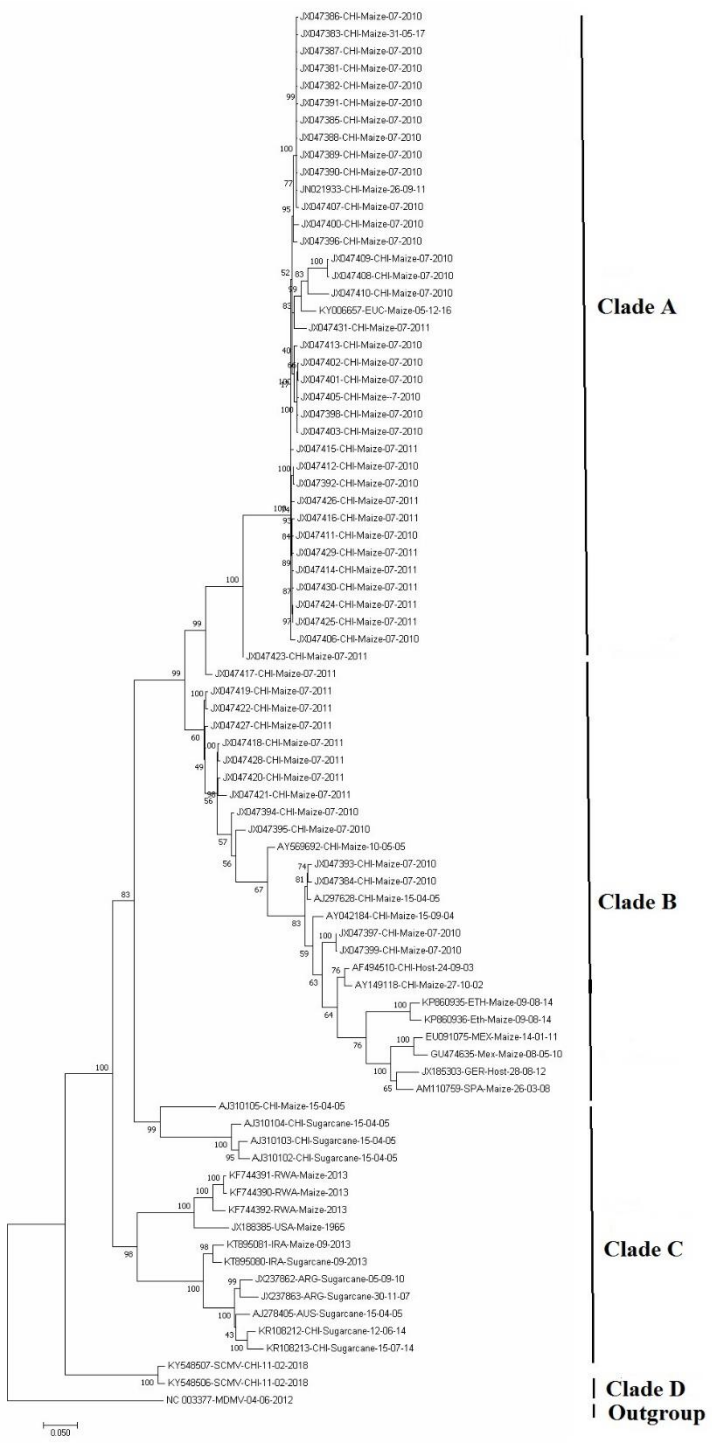


Figure 4.8.1. A maximum likelihood phylogenetic tree estimated using whole genome of SCMV. Each branch is led by a virus and is denoted by following format, GenBank accession number _Country _Host _Date. Four clades A, B, C and D are made. Branches are well supported by bootstrap values.

4.9 Bayesian Analysis

Phylogenetic Bayesian analysis for *Sugarcane mosaic virus* has divided phylogenetic tree as shown in Figure 4.9.1 into four clades, named as A, B, C and D. Clade number A consists of sequences belongs to China except one sequence KY006657 which is from country Ecuador. Clade number B consists of viruses polyphyletic in nature and belongs to Mexico, Germany, Spain and Ethiopia. Clade number C consists of eleven sequences, two from Iran, China and Argentina each, one sequence from Australia and USA each while three sequences belongs to country Rwanda.

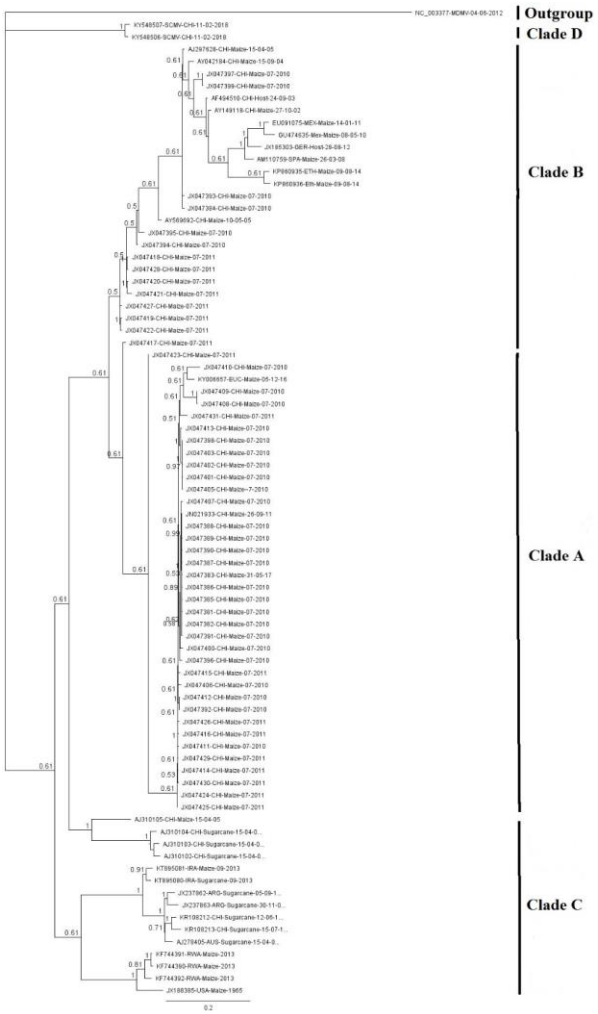


Figure 4.9.1: Bayesian phylogeny of full genome of *Sugarcane mosaic virus* constructed for viruses with polyphyletic origin. Genbank accession numbers, countries, years and host are shown. Each clade is marked on right.

4.10 Comparison between Phylogenetic trees

Maximum likelihood and Bayesian analysis based phylogenetic trees (SCMV, full genome) were compared and they both appeared to comprehend each other. Isolates clustered in different clades were same in both trees. As sequences belong to Ethiopia, Mexico, Germany and Spain were cladded together, similarly SCMV from Rwanda, USA, Australia, Argentina, Iran were clustered together with the exception of two Chinese SCMV sequences (KR108212, KR108213) which were also present in same group. All other Chinese isolates were grouped together with one exceptional sequence, which belonged to country Ecuador.

Comparison between ML based phylogenetic tree made with CP sequences, ML tree and Bayesian analyzed tree constructed from full genome of SCMV showed a little difference. In full genome based both trees had sequences belonged to Ethiopia, Mexico, Germany, Spain and sequences from Iran, Rwanda, Argentina, Australia and USA were clustered into two different clades, in case of ML CP based phylogenetic tree isolates from Mexico, USA, Rwanda, Ethiopia, Argentina and Iran cladded together. Sequences AJ310104, AJ310103, AJ310102 isolated from host sugarcane (country, China) were grouped together with sequence AJ310105 (from host maize and country china) present on different sub-branch in both trees from full genome, but were not clustered together in CP sequences-based tree. In CP based tree three sequences from host sugarcane cladded together while AJ310105 from host maize grouped separately by making sub-branch with sequences from Mexico, Germany and Spain.

P1 based maximum likelihood tree showed more similarities and less differences when compared with full genome-based ML and Bayesian phylogenetic tree. Mexican, German, Ethiopian and Spanish isolates were cladded together in whole genome-based tree but were not clustered together in P1 based tree. Mexican and Ethiopian sequences were scattered from sequences belonged to Germany and Spain. Isolates from Iran, USA, Rwanda, Argentina grouped together in all trees, maximum Chinese sequences clustered together except with one sequence from different country (from Ecuador).

All three sequences based analysis is validated by sequence demarcation tool (SDT), it is tool specified for *Geminiviruses* and it tend to show pair wise sequence comparison without any GAP character, output obtained in form of colored matrix instead of percentage given in numbers, instead color matrix can be compared with percentage scale given (Brown *et al.,* 2015; Brown *et al.,* 2012).The cut off value given by SDT can also be implemented on characterization of SCMV. To proof the argument two sequences (KY548507, KY548506) from China can be exemplified. Color matrix had shown that these two sequences have minimum identity with all other sequences (78% to 77%), this output is supported by the fact that these two sequences were cladded separately from all other sequences.

SCMV is made up of ten proteins, justification for why just CP, P1 and full genome is used to perform phylogenetic analysis is given classification regarding family *Potyviridae* which is dependent on these three sequences. Genus demarcation criteria for family *Potyviridae* depends on molecular data (Adams *et al.,* 2005; Gibbs and Ohshima, 2010; ICTV 10th report, 2017). For the complete ORF (open reading frame) criteria is <46% nucleotide identity, which doesn't discriminate between *rymovirus* and *potyvirus*, these two are discriminated by their vectors (ICTV 10th report, 2017) while in family *potyviridae*, specie demarcation criteria for entire open reading frame (ORF) is nucleotide identity <76%, amino acid identity <82%. The threshold demarcation criteria for individual coding regions vary from 58% nucleotide identity for P1 to 74-78% for other coding regions, while criteria for CP (coat protein) is 80% amino acid identity, 76-77% nucleotide-based identity (Adams *et al.,* 2005; ICTV 10th report, 2017).

4.11 Recombination Analysis

Recombination analysis was performed for Ecuador, Ethiopia, Mexico and Rwanda using different methods implemented in RDP4 software is shown in Figure 4.11.1. In case of Ecuador (KY006657), major and minor parents determined were JX047431 and JX047410 respectively. For recombination event number one, recombined region was within 7009 nt to 9557 nt and in event 2 within the part 6526 (C-terminal) and 7196 shown in Figure 4.11.1 and 4.11.2 and in Table 4.11.1

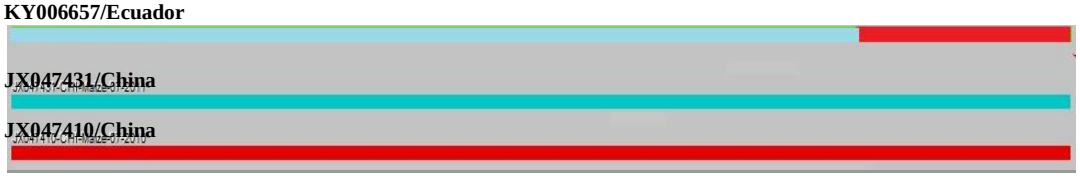


Figure 4.11.1: Recombination showing picture for Ecuador sequence**.** Figure 4.11.1 is depicting recombination events. Major (JX047431) and minor (JX047410) parents are labeled below every recombination point. Recombined region in this recombination are C terminal region of 6K1, NIa and NIb.

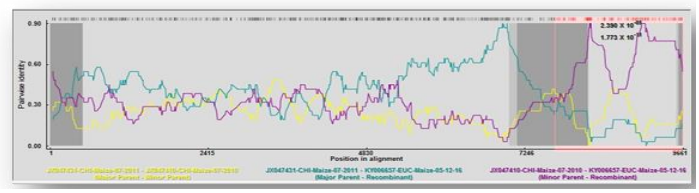


Figure 4.11.2: Recombination analysis of KY006657**.** Figure 4.11.2 is representing major and minor parents for KY006657 at recmobination points. Major and mainor parents are JX047431 and JX047410 respectively at recombined regions 6K1, NIa and Nib

Table 4.11.1 Recombination table for full genome sequence from Ecuador

KY006657/Ecuador Recombination Event Number		
	1	2
Major parent (site and host)	JX047431/Maize/China	JX047431/Maize/China
Minor parent (site and host)	JX047410/Maize/China	JX047410/Maize/China
P values determined by seven different programs		
RDP	$1.773\ 10^{-11}$	$1.307\ 10^{-04}$
GENECONV	$4.596\ 10^{-12}$	$8.179\ 10^{-06}$
BOOTSCAN	$1.834\ 10^{-13}$	$3.398\ 10^{-07}$
MAXCHI	$2.406\ 10^{-15}$	$3.078\ 10^{-02}$
CHIMERA	$1.575\ 10^{-13}$	$6.035\ 10^{-03}$
SISCAN	-	-
3SEQ	$3.219\ 10^{-14}$	-
Beginning breakpoint (nt)	7009	6526
Ending breakpoint(nt)	9557[a]	7196[a]

Table 4.11.1 consists of P values given by different methods implemented in RDP4 software. Recombination events are confirmed by five methods (minimum) and six methods (maimum)

Recombination analysis number two, In for Ethiopian sequence (KP860936), major and minor parents were AM110759 (Spain) and KF744391 (Rwanda) respectively. For recombination event number one, recombined region was within 5508 (CI) nt to 9055 (Nib) nt as shown in Figure 4.11.3, 4.11.4 and table 4.11.2.

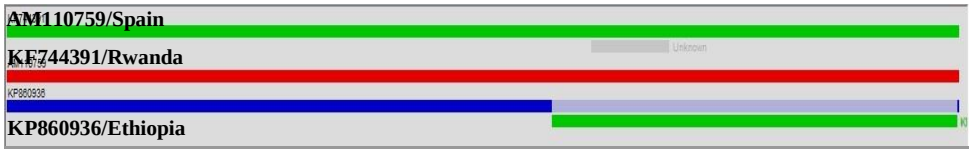


Figure 4.11.3: Recombination showing picture for Ethiopian sequence**.** This figure is depicting recombination events. Major (AM110759) and minor (KF744391) parents are labeled below every recombination point. Recombined region in this recombination are CI, 6K1, NIa and NIb.

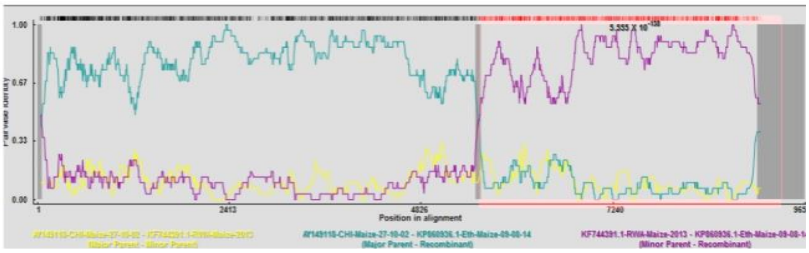


Figure 4.11.4: Recombination analysis of Ethiopia (KP860936) This figure is representing major and minor parents for KP860936 at recmobination points. Major and mainor parents are AM110759 and KF744391 respectively at recombined regions 6K1, NIa and Nib

Table 4.11.5 Recombination table for full genome sequence from Ethiopia (KP860936)

KP860936/Ethiopia	Recombination Event Number
	1
Major parent (site and host)	AM110759/Maize/Spain
Minor parent (site and host)	KF744391/Maize/Rwanda
P values determined by seven different programs	
RDP	5.55510^{-138}
GENECONV	5.648 10-126
BOOTSCAN	4.281 10-133
MAXCHI	$3.135\ 10^{-54}$
CHIMERA	$3.727\ 10^{-58}$
SISCAN	-
3SEQ	$3.330\ 10^{-14}$
Beginning breakpoint (nt)	5508
Ending breakpoint(nt)	9055[a]

Table 4.11.5 consists of P values given by different methods implemented in RDP4 software. Recombination events are confirmed by five methods (minimum) and six methods (maimum)

Recombination analysis number three, in case of Mexican sequence (EU091075), major and minor parents were GU474635 (Mexico) and AM110759 (Spain) respectively. For recombination event number one, recombined region was within 8037 nt to 9341 nt as shown in Figure 4.11.5, 4.11.6 and table 4.11.3

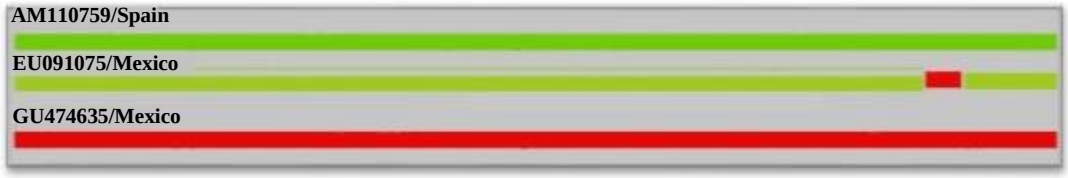


Figure 4.11.5: Recombination showing picture for Mexican sequence. Figure 4.11.5 depicting recombination events. Major (GU474635) and minor (AM110759) parents are labeled below every recombination point. Recombined region in this recombination are CI, 6K1, NIa and NIb.

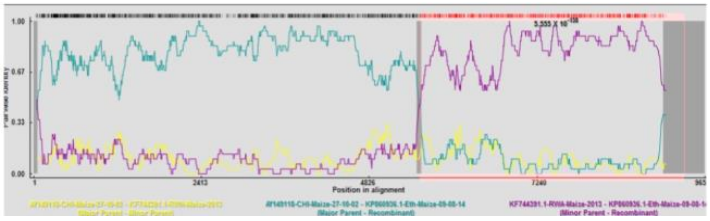


Figure 4.11.6: Recombination analysis of Mexican sequence (KP860936**). F**igure 4.11.6 is representing major and minor parents for KP860936 at recmobination points. Major and mainor parents are GU474635 and AM110759 respectively at recombined regions 6K1, NIa and Nib

Table 4.11.3: Recombination table for full genome sequence from Mexico (EU091075)

EU091075/Mexico Recombination Event Number	
	1
Major parent (site and host)	GU474635/MaizeMexico/
Minor parent (site and host)	AM110759/Maize/Spain
P values determined by seven different programs	
RDP	$9.577\ 10^{-24}$
GENECONV	$7.485\ 10^{-23}$
BOOTSCAN	$4.780\ 10^{-22}$
MAXCHI	$6.557\ 10^{-16}$
CHIMERA	$3.596\ 10^{-18}$
SISCAN	-
3SEQ	$2.220\ 10^{-16}$
Beginning breakpoint (nt)	8037
Ending breakpoint(nt)	9431[a]

Table 4.11.3 consists of P values given by different methods implemented in RDP4 software. Recombination events are confirmed by five methods (minimum) and six methods (maimum). Beginning and ending breaking points are at 8037 nt to 9431 nt

Recombination analysis number four, in case of Rwanda sequence (KF744390), major and minor parents were JX237863 (Argentina) and KP860936 (Ethiopia) respectively. For recombination event number one, recombined region was within 5574 nt to 9174 nt as shown in Figure 4.11.7, 4.11.8 and table 4.11.4.

Figure 4.11.7: Recombination showing picture for Rwanda sequence. This figure is depicting recombination events of KF744390. Major (JX237863) and minor (KP860936) parents are labeled below every recombination point. Recombined region in this recombination are CI, 6K1, NIa and Nib.

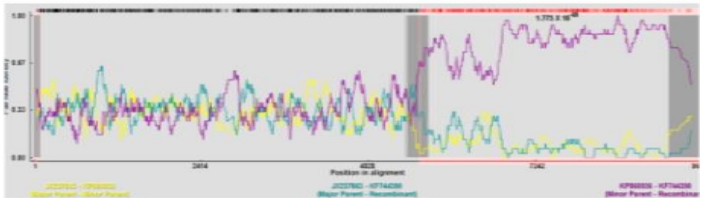


Figure 4.11.8: Figure 4.11.8 representing major and minor parents for KF744390 at recmobination points. Major and mainor parents arE JX237863 and KP860936 respectively at recombined regions 6K1, NIa and Nib

Table 4.11.4: Recombination table for full genome sequence from Mexico (EU091075)

KF744390/Rwanda Recombination Event Number	
	1
Major parent (site and host)	JX237863/Saugarcanae/Argentina
Minor parent (site and host)	KP860936/Maize/Ethiopia
P values determined by seven different programs	
RDP	$1.775\ 10^{-69}$
GENECONV	1.753 10-81
BOOTSCAN	$9.842\ 10^{-81}$
MAXCHI	$9.840\ 10^{-16}$
CHIMERA	$1.566\ 10^{-03}$
SISCAN	6.284 10-32
3SEQ	$1.099\ 10^{-02}$
Beginning breakpoint (nt)	5574
Ending breakpoint(nt)	9174

Table 4.11.4 consists of P values given by different methods implemented in RDP4 software. Recombination events are confirmed by five methods (minimum) and six methods (maimum). Beginning and ending breaking points are at 5574 nt to 9174 nt

It concludes that sequences geographical based or host-based distribution is just not enough to classify the viruses as recombination is the main driving in evolution and emergence of new SCMV variants. Recombination paired with mutations could form viruses with transformed biological properties.

4.12 Sequence Demarcation Tool Analysis

4.12.1 Matrix Analysis

Sequence demarcation criteria tool analysis representing all retrieved sequences of *Sugarcane mosaic virus* from NCBI is showing pairwise identity from 77 % to 100 %. All sequences are showing different shades of red in Figure 4.12.1 except last two sequences, in addition to shades of red they are showing shades of blue with pairwise identity ranging from 76% to 100% and SDT plot is shown in Figure 4.12.2.

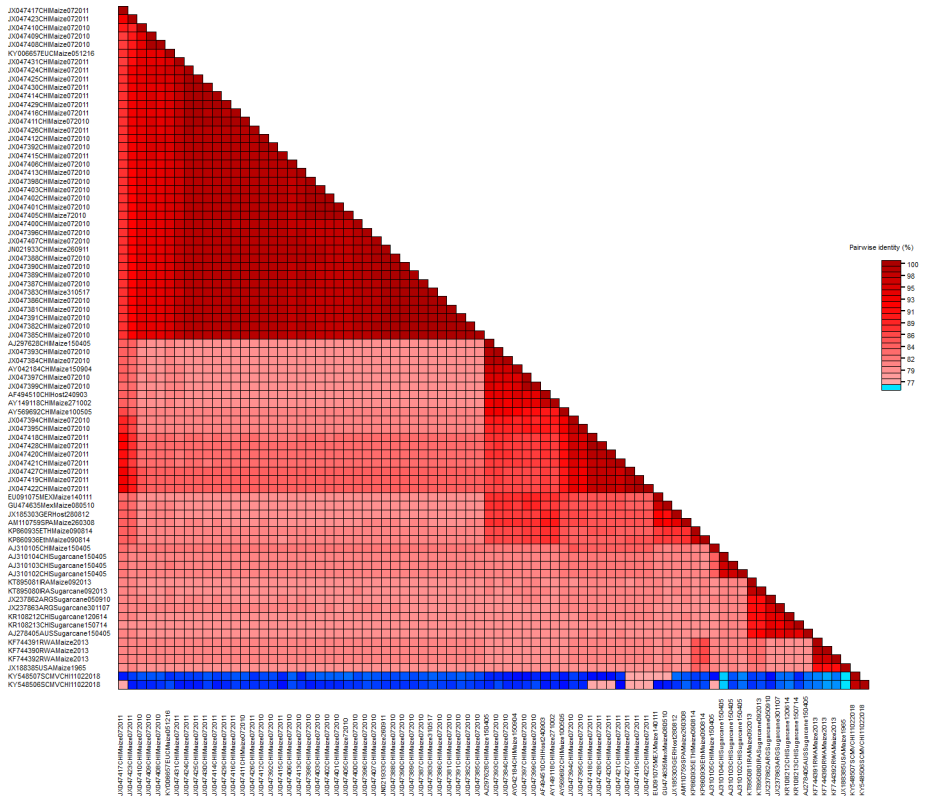


Figure 4.12.1: Pairwise distance matrix is representing by colors, range of percentage identity in numerical value is shown in scale on right side.

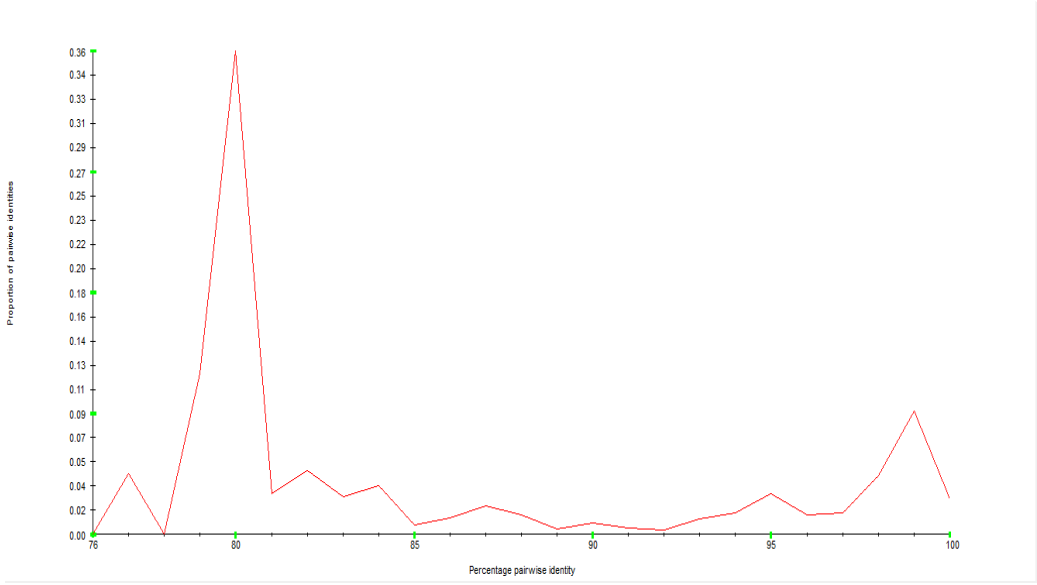


Figure 4.12.2: Species demarcation tool plot. Plot is representing valley (conflict free area) at 85% and at 89% to 92%. Peak represents percentage value with maximum conflict

According to recent studies SCMV is distinctly classifying into four groups. Cut off values proposed by ICTV for species demarcation criteria given is debatable on the basis of SDT (species demarcation tool) .Valley (conflict free area) of SDT plot may be considered to determine cut of value which is between between 85% (minimum) to 92% (maximum),but recent cut off value is 76% (by ICTV). The criteria of classification may be revised or SCMV could be divided into strains. Since sufficient data is available, classification of SCMV may be upgraded to strains level

References

Achon, M.A., Medina, V.S.M., Markham, P., Lomonossoff, G.P. (1994). Characterization of a maize infecting *potyvirus* from Spain. Eur J Plant Pathol, 100:157–165.

Achon, M.A., Sobrepere, M., Minguell, R. (2003). Molecular and biological properties of a *Sugarcane mosaic potyvirus* isolate from Spain. J Plant Dis Protect, 110:324–331

Adams, M.J., Antoniw, J.F., Fauquet, C.M., (2005). Molecular criteria for genus and species discrimination within the family *Potyviridae*. Arch Virol, 150: 459–479.

Adams, I.P. (2012). Use of next-generation sequencing for the identification and characterization of *Maize chlorotic mottle virus* and *Sugarcane mosaic virus* causing maize lethal necrosis in Kenya. Plant Pathol, 10.1111/j.1365-3059.2012.02690.

Chare, E.R., Holmes, E.C., (2006). A phylogenetic survey of recombination frequency in plant RNA viruses. Arch Virol, 151: 933–946.

Chen, J., Chen. J., Adams, M.J., (2002). Characterization of *potyviruses* from sugarcane and maize in China. Arch Virol, 147: 1237–1246.

Fan, Z.F., Cheng, H.Y., Liang, X.M., Li, H.F. (2003). Complete sequence of the genomic RNA of the prevalent strain of a *potyvirus* infecting maize in China. Arch Virol 148:773–782.

Gough, K.H., Shukla, D.D., (1993). Nucleotide sequence of *Johnson grass mosaic potyvirus* genomic RNA. Inter virology, 36: 181–192.

Guindon, S., Gascuel, O., (2003). A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol, 52: 696–704.

Kong, P., Steinbiss, H.H., (1998). Complete nucleotide sequence and analysis of the putative polyprotein of *maize dwarf mosaic virus* genomic RNA (Bulgarian isolate). Arch Virol, 143:1791–1794

Kumar, S., Tamura, K., Nei, M., (2004). MEGA3: integrated software for molecular evolutionary genetic analysis and sequence alignment. Brief Inform, 5: 150–163

Li, L., Wang, X., Zhou, G. (2007). Analyses of maize embryo invasion by *Sugarcane mosaic virus*. Plant Sci, 172: 131–138

Matthews, R.E.F. (1982). Classification and nomenclature of viruses. Fourth report of the international committee on classification of viruses. Intervirology, 17, 1e199.

Safarnejad, M.R., Jouzani, G.S., Tabatabaie, M., Twyman, R.M., Schillberg, S. (2011). Antibody-mediated resistance against plant pathogens. Biotechnol Adv, 29, 961e971.

Shukla, D.D., Tosic, M., Jilka, J., Ford, R.E., Toler, R.W., Langham, M.A.C. (1989). Taxonomy of *potyviruses* infecting maize, sorghum, and sugarcane in Australia and the United States as determined by reactivities of polyclonal antibodies directed towards virus-specific N termini of coat proteins. Phytopathology, 79,223e229.

Xu, D.L., Park, J.W., Mirkov, T.E., Zhou, G.H. (2008). Viruses causing mosaic disease in sugarcane and their genetic diversity in southern. China. Arch. Virol, 153,1031e1039.

Deng, Y.Q., Yang, Y.Q., Zhai, Y.S. (2016). Genome cloning of two *Sugarcane mosaic virus* isolates from Fuzhou and phylogenetic analysis of SCMV. Acta Phytopath Sinica, 46(6): 775-782.

Ha, C., Revill, P., Harding, R.M., Vu, M., Dale, J.L. (2008). Identification and sequence analysis of *potyviruses* infecting crops in Vietnam. Arch. Virol, 153, 45e60.

Handley, J.A., Smith, G.R., Dale, J.L., Harding, R.M. (1996). Sequence diversity in the NIb coding region of eight sugarcane mosaic potyvirus isolates infecting sugarcane in Australia. Arch. Virol, 141, 2289e2300.

Handley, J.A., Smith, G.R., Dale, J.L., Harding, R.M. (1998). Sequence diversity in the coat protein coding region of twelve sugarcane mosaic potyvirus isolates from Australia, USA and South Africa. Arch. Virol, 143, 1145e1153.

Brandes, E.W. (1919). The mosaic disease of sugarcane and other grasses. Tech. Bull. US Dept. Agr, 829, 1e26.

Harrison, B.D., Finch, J.T., Gibbs, A.J., Hollings, M., Shepherd, R.J., Valenta, V., Wetter, C. (1971). Sixteen groups of plant viruses. Virology, 45, 356e363.

Perera, M.F., Filippone, M.P., Noguera, A.S., Cuenya, M.I., and Castagnaro, A.P. (2012). An Overview of the Sugarcane Mosaic Disease in South America. CAB International. 1974. *Sugarcane mosaic virus*. Distribution Maps of Plant Diseases. Edition 4 (October), Map 229. Wallingford, UK: CAB International.

Jensen, S.G. (1992) A viewpoint on the taxonomy of potyviruses infecting sugarcane, maize, and sorghum. In: Barnett O.W. Potyvirus Taxonomy. Archives of Virology (Supplementum 5), vol 5. Springer, Vienna.

Achon, M. A., Serrano, L., N., A. D and Porta, C. (2007). Complete genome sequences of *Maize dwarf mosaic* and *Sugarcane mosaic virus* isolates coinfecting maize in Spain. Arch Virol, 152(11), 2073-2078

Adams, I., Harju, V., Hodges, T., Hany, U., Skelton, A., Rai, S. and Boonham, N. (2014).

First report of maize lethal necrosis disease in Rwanda. New Dis. Rep, 29, 22-22.

Martin, D.P., Williamson, C., Posada, D. (2005). RDP2: Recombination detection and analysis from sequence alignments. Bioinformatics, 21: 260–262

Nicolas, O., Laliberte, J.F. (1992). The complete nucleotide sequence of turnip mosaic *potyvirus* RNA. J Gen Virol, 73: 2785–2793

Oertel, U., Schubert, J.E., Fuchs, E. (1997). Sequence comparison of the 30-terminal parts of the RNA of four German isolates of sugarcane mosaic *potyvirus* (SCMV). Arch Virol, 142: 675–687

Oertel, U., Fuchs, E., Hohmann, F. (1999). Differentiation of isolates of sugarcane mosaic *potyvirus* (SCMV) on the basis of molecular, serological and biological investigations. J Plant Dis Protect, 106: 304–313

Revers, F., Le, G.O., Le, R. M., Dunez, J. (1996). Frequent occurrence of recombinant *potyvirus* isolates. J Gen Virol, 77: 1953–1965

Revers, F., Le, G.O., Candresse, T., Maule, A.J. (1999). New advances in understanding the molecular biology of plant *potyvirus* interactions. Mol Plant Microbe Interact, 12: 367–376

Rodriguez, E., Gamble, P., Shaw, J.G. (1991). A determinant of disease symptom severity is located in the 30-terminal noncoding region of the RNA of a plant virus. Proc Natl Acad Sci, 88:9863–9867

Seifers, D.L., Salomon, R., Marie, V., Alliot, B., Signoret, P., Haber, S., Loboda, A., She Y.M., Standing, K.G. (2000). Characterization of a novel *potyvirus* isolated from maize in Israel. Phytopathology, 90:505–513

Shukla, D.D., Ward, C.W., Brunt, A.A. (1994). The *Sugarcane mosaic virus* subgroup in the *potyviridae*. CAB International, Wallingford, pp 360–371

Urcuqui, I, S., Haenni, A.L., Bernadi, F. (2001). *Potyvirus* proteins: a wealth of functions. Virus Res, 74: 157–175.

Louie, R. (1980). *Sugarcane mosaic virus* in Kenya. Plant Dis, 64: 944–947.

Weiller, G.F. (1998). Phylogenetic profiles: a graphical method for detecting genetic recombination in homologous sequences. Mol Biol Evol, 15: 326–335

Yang, Z.N., Mirkov., T.E. (1997). Sequence and relationships of sugarcane mosaic and sorghum mosaic virus strains and development of RT-PCR-based RFLPs for strain discrimination. Phytopathology, 87: 932–939

Zhong, Y., Guo, A., Li, C., Zhuang, B., Lai, M., Wei, C., Luo, J., Li, Y. (2005). Identification of a naturally occurring recombinant isolate of *Sugarcane mosaic virus* causing maize dwarf mosaic disease. Virus Genes, 30:75–83

Cheng, D.J.,Yan, Z,Y., Huang, X. D. (2017). Complete genomic sequence and analysis of two *Sugarcane mosaic virus* isolates from Shandong Province, China. Acta Phytopathologica Sinica, 47(3): 357-363.

Morris, M.L., Heisey, P.W. (2003). Estimating the benefits of plant breeding research: methodological issues and practical challenges. Agricultural Economics, 29, 241–252.

USDA. (2014). Maize Lethal Necrosis – The growing challenge in Eastern Africa. GAIN Report. Global Agricultural Information Network, Washington DC.

USDA. (2O18). Grain Production Better than Expected in Ethiopia. Global Agricultural Information Network, Washington DC, GAIN Report, ET1813.

USDA. (2O18). Grain and feed annual in Argentina. Global Agricultural Information Network, Washington DC, GAIN Report.

USDA. (2O18). Grain and feed annual in Australia. Global Agricultural Information Network, Washington DC, GAIN Report, AS1811.

USDA. (2O18). Grain and feed annual in Pakistan. Global Agricultural Information Network, Washington DC, GAIN Report, PK1810.

USDA. (2O18). Grain and feed annual in Mexico, slight changes in production as grain imports continue upward trend. Global Agricultural Information Network, Washington DC, GAIN Report, MX8010.

USDA. (2O18). Grain: World markets and trade. Foreign Agricultural Services.

PSMA. (2017) Pakistan Sugar mills association, Annual report.

Achon, M.A., Serrano, L., Alonso, N., Porta, C. (2007). Complete genome sequences of m*aize dwarf mosaic* and *Sugarcane mosaic virus* isolates co-infecting maize in Spain. Archives of Virology, 152, 2073–2078.

Alegria, O.M., Royer, M., Bousalem, M., Chatenet, M., Peterschmitt, M., Girard, J.C., Rott, J. (2003). Genetic diversity in the coat protein coding region of eighty-six *Sugarcane mosaic virus* isolates from eight countries, particularly from Cameroon and Congo. Archives of Virology, 148, 357–372.

Bousalem, M., Douzery, E.J.P., Fargette, D., (2000). High genetic diversity, distant phylogenetic relationships and intra-species recombination events among natural populations of Yam mosaic virus: a contribution to understanding *potyvirus* evolution. Journal of General Virology, 81, 243 255.

Cervera, M.T., Riechmann, J.L., Martin, M.T., Garcia, J.A. (1993). 3-terminal sequence of the Plum pox virus PS and o6 isolates: evidence for RNA recombination within the *potyvirus* group. Journal of General Virology, 74, 329–334.

Chung, B.Y.W., Miller, W.A., Atkins, J.F., Firth, A.E. (2008). An overlapping essential gene in the *Potyviridae*. Proceedings of the National Academy of Sciences of the United States of America, 105, 5897–5902.

Domingo, E., Holland, J.J. (1997). RNA virus mutations and fitness for survival. Annual Review of Microbiology, 51, 151–178.

García-Arenal, F., Fraile, A., Malpica, J.M. (2001). Variability and genetic structure of plant virus populations. Annual Review of Phytopathology, 39, 157–186.

Gao, B., Cui, X.W., Li, X.D., Zhang, C.Q., Miao, H.Q. (2011). Complete genomic sequence analysis of a highly virulent isolate revealed a novel strain of *Sugarcane mosaic virus*. Virus Genes, 43, 390–397.

Holmes, E.C. (2009). The evolutionary genetics of emerging viruses. Annual review of ecology and systematics, 40, 353–372.

Hudson, R.R. (2000). A new statistic for detecting genetic differentiation. Genetics, 155, 2011–2014.

Kosakovsky, P. S.L., Frost, S.D.W. (2005). Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. Bioinformatics, 21, 2531–2533.

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

Moreno, I.M., Malpica, J.M., Díaz, J.A., Moriones, E., Fraile, A., Garcia, F. (2003). Variability and genetic structure of the population of *watermelon mosaic virus* infecting melon in Spain. Virology, 318, 451–460.

Padhi, A., Ramu, K. (2011). Genomic evidence of intraspecific recombination in *Sugarcane mosaic virus*. Virus Genes, 42, 282–285.

Pirone, T.P., Blanc, S. (1996). Helper-dependent vector transmission of plant viruses. Annual Review of Phytopathology, 34, 227–247.

Revers, F., Le, G. O., Candresse, T., Le, R. M., Dunez, J., (1996). Frequent occurrence of recombinant *potyvirus* isolates. Journal of General Virology, 77, 1953–1965.

Revers, F., Le, G. O., Candresse, T., Maule, A.J. (1999). New advances in understanding the molecular biology of plant/*potyvirus* interactions. Mol Plant Microbe Interact 12, 367–376.

Shukla, D.D., Frenkel, M.J., Ward, C.W. (1991). Structure and function of the *potyvirus* genome with special reference to the coat protein coding region. Canadian J. Plant Pathol, 13,178–191.

Shukla, D.D., Strike, P.M., Tracy, S.L., Gough, K.H., Ward, C.W., 1988. The N and C termini of the coat proteins of *potyviruses* are surface-located and the N terminus contains the major virus-specific epitopes. J. Gen Virol, 69, 1497–1508.

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

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evolution, 28, 2731–2739.

Thompson, J.D., Higgins, D.G., Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acid Res, 22, 4673–4680.

Tomitaka, Y., Ohshima, K. (2006). A phylogeographical study of the *Turnip mosaic virus* population in East Asia reveals an 'emergent' lineage in Japan. Mol Eco, 15, 4437–4457.

Tugume, A.K., Cuellar, W.J., Mukasa, S.B., Valkonen, J.P.T. (2010). Molecular genetic analysis of virus isolates from wild and cultivated plants demonstrates that East Africa is a hotspot for the evolution and diversification of *sweet potato feathery mottle virus*. Mol Eco, 19, 3139–3156.

Urcuqui, I.S., Haenni, A.L., Bernardi, F. (2001). *Potyvirus* proteins: a wealth of functions. Virus Res, 74, 157–175.

Wang, J.G., Zheng, H.Y., Chen, H.R., Adams, J.M., Chen, J.P. (2010). Molecular diversities of *Sugarcane mosaic virus* and *Sorghum mosaic virus* isolated from Yunnan province, China. J. Phytopathol, 158, 427–432.

Zhong, Y.W., Guo, A.Y., Li, C.B., Zhuang, B.Q., Lai, M., Wei, C.H., Luo, J.C., Li, Y. (2005). Identification of a naturally occurring recombinant isolate of *Sugarcane mosaic virus* causing maize dwarf mosaic disease. Virus Genes, 30, 75–83.

Adams, I. P., Miano, D. W., Kinyua, Z. M., Wangai, A., Kimani, E., Phiri, N., and Boonham, N. (2013). Use of next-generation sequencing for the identification and characterization of *Maize chlorotic mottle virus* and *Sugarcane mosaic virus* causing maize lethal necrosis in Kenya. (62), Wiley-Blackwell, Oxford. CAB Direct database.

Addy, H., Nurmalasari, Wahyudi, A., Sholeh, A., Anugrah, C., Iriyanto, F., and Sugiharto, B. (2017). Detection and Response of Sugarcane against the Infection of *Sugarcane mosaic virus* (SCMV) in Indonesia. Agronomy, 7(3), 50.

Ahmed, S., Rehman, O.-U., Ahmad, F., Asam Riaz, M., and Hussain, A. (2007). Varietal resistance in maize against chemical control of stem-borer, shoot-fly and termites in sahiwal, punjab, Pakistan. Pak. J. Agri. Sci, 44(3), 120-123.

Akbar, S., Tahir, M., and Afghan, S. (2017). Characterization of coat protein (CP) gene of *Sugarcane mosaic virus* (SCMV) from isolates of Pakistan and its phylogenetic relationships. J. Anim. Plant Sci, 27(1).

Andrew, W. K. (2015). The polymerase slips and PIPO exists. EMBO reports, 16(8), 885-886.

Brown, J., Zerbini, F., Navas-Castillo, J., Moriones, E., Ramos, R., Silva, C., and Varsani, A. (2015). Revision of *Begomovirus* taxonomy based on pairwise sequence comparisons. Arch Virol, 160(6), 1593-1619.

Chare, E. R., and Holmes, E. (2006). A phylogenetic survey of recombination frequency in plant RNA viruses. Arch virol, 151(5), 933-946.

Chave, J., Chust, G., and Thébaud, C. (2007). The importance of phylogenetic structure in biodiversity studies. Scaling biodiversity, 151-167.

Chaves, B.G., Espejel, F., Alcala, B. R., Hernández, V. J., and Silva-Rosales, L. (2011). Short distance movement of genomic negative strands in a host and non-host for *Sugarcane mosaic virus* (SCMV). Virology J, 8(1), 15.

Chaves-Bedoya, G., and Ortiz-Rojas, L. Y. (2012). Evidence of different phylogenetic origins of two mexican *Sugarcane mosaic virus* (SCMV) isolates. Acta Agronómica, 61(1), 79-87.

Chen, H., Cao, Y., Li, Y., Xia, Z., Xie, J., Carr, J., and Zhou, T. (2017). Identification of differentially regulated maize proteins conditioning *Sugarcane mosaic virus* systemic

infection. New Phytol, 215(3), 1156-1172.

Chen, J., Chen, J., and J Adams, M. (2002). Characterisation of *potyviruses* from sugarcane and maize in China. Arch virol, 147(6), 1237-1246.

Cheng, Y., Chen, J., and Chen, J. (2002). The complete sequence of a *Sugarcane mosaic virus* isolate causing maize dwarf mosaic disease in China. Sci China Series C Life Sci, 45(3), 321-330.

Cracraft, J. (1983). The significance of phylogenetic classifications for systematic and evolutionary biology. In Numerical taxonomy (pp. 1-17). Springer, Berlin, Heidelberg.

F Perera, M., P Filippone, M., Ramallo, J., Cuenya, M., Garcia, M. L., Ploper, L., and P Castagnaro, A. (2009). Genetic Diversity Among Viruses Associated with Sugarcane Mosaic Disease in Tucumán, Argentina. Phytopathology, 99(1), 38-49.

Fan, Z., Chen, H. Y., M Liang, X., and F Li, H. (2003). Complete sequence of the genomic RNA of the prevalent strain of a *potyvirus* infecting maize in China. Arch virol, 148(4), 773-782.

Gao, B., Cui, X.-W., Li, X., Zhang, C.-Q., and Miao, H.-Q. (2011). Complete genomic sequence analysis of a highly virulent isolate revealed a novel strain of *Sugarcane mosaic virus.* Virus genes, 43(3), 390.

Gemechu, A. L., Chiemsombat, P., Attathom, S., Lersrutaiyotin, R., and Reanwarakorn, K. (2004). The variations among isolates of *Sugarcane mosaic virus*in Thailand as determined by virus-host interaction. Kasetsart Journal (Natural Science), 38, 369-379.

Ghori, N.-u.-H., Hayat, M. Q., and Davino, S. (2016). Genetic diversity and evolutionary analysis of *Citrus Tristeza Virus* p20 gene in Pakistan: insights into the spread and epidemiology. Ad Life Sci, 3(3), 75-82 (Vol. 3).

Ghori, N.-u.-H., Shafique, A., Hayat, M. Q., and Anjum, S. (2016). The phylogeographic and spatiotemporal spread of HCV in Pakistani population. PloS one, 11(10), e0164265.

Grisham, M., and Pan, Y.-B. (2007). A Genetic Shift in the Virus Strains that Cause Mosaic in Louisiana Sugarcane. Plant Disease, 91(4), 453-458.

H Berger, P., D Wyatt, S., J Shiel, P., J Silbernagel, M., Druffel, K., and I Mink, G. (1997). Phylogenetic analysis of the *Potyviridae* with emphasis on legume-infecting *potyviruses.* Arch Virol, 142(10), 1979-1999.

Haider, M., Afghan, S., Riaz, H., Tahir, M., Javed, M. A., Rashid, N., and Iqbal, J. (2011).

Identification of two *Sugarcane mosaic virus* (SCMV) variants from naturally infected sugarcane crop in Pakistan. Pak. J. Bot, 43(2), 1157-1162.

Handley, J. A., Smith, G., Dale, J., and M Harding, R. (1998). Sequence diversity in the coat protein coding region of twelve sugarcane mosaic *potyvirus* isolates from Australia, USA and South Africa. Arch Virol, 143(6), 1145-1153.

Hayat, M. Q., Basra, S., Khan, M., Yasmin, G., Shaheen, and Shazia Jabeen, N. (2009). Phylogenetic Relationships in *Artemisia spp.* (*Asteraceae*) Based on Distribution of Foliar Trichomes. Int. J. Agric. Biol, 11, 553-558.

Hohmann, F., Fuchs, E., Grüntzig, M., and Oertel, U. (1999). A contribution to the ecology of *Sugarcane mosaic potyvirus* (SCMV) and *Maize dwarf mosaic potyvirus* (MDMV) in Germany. J Plant Dis Protect, 314-324.

Kong, P., and H Steinbiss, H. (1998). Complete nucleotide sequence and analysis of the putative polyprotein of *maize dwarf mosaic virus* genomic RNA (Bulgarian isolate). Arch Virol, 143(9), 1791-1799.

Lakshmi, B., Balasubramaniam, P., Chinnadurai, C., S, B., and Viswanathan, R. (2014). Complete genome sequence of Indian *Sugarcane mosaic virus*and its genetic diversity with other country isolates. In Proceedings of the symposium on bioenergy for sustainable development-the potential role of sugar crops (pp. 23-25).

Leng, P., Ji, Q., Tao, Y., Ibrahim, R., Pan, G., Xu, M., and Lübberstedt, T. (2015). Characterization of *Sugarcane mosaic virus*scmv1 and scmv2 resistance regions by regional association analysis in maize. PloS one, 10(10), e0140617.

Li, Y., Liu, R., Zhou, T., and Fan, Z. (2013). Genetic diversity and population structure of *Sugarcane mosaic virus*. Virus Research, 171(1), 242-246.

López-Moya, J. J., Wang, R. Y., and P Pirone, T. (2000). Context of the coat protein DAG motif affects *potyvirus* transmissibility by aphids. J Gen Virol, 80(12), 3281-3288.

M Alegria, O., Royer, M., Bousalem, M., Chatenet, M., Peterschmitt, M., Girard, J. C., and Rott, P. (2003). Genetic diversity in the coat protein coding region of eighty-six *Sugarcane mosaic virus*isolates from eight countries, particularly from Cameroon and Congo. Arch virol, 148(2), 357-372.

M. Martinez, H. (1983). An efficient method for finding repeats in molecular sequences. Nucleic acid res, 11(13), 4629-4634.

Mekureyaw, M. F. (2017). Journal of plant physiology and pathology maize lethal necrosis disease: an emerging problem for maize production in Eastern Africa.

Phytopathology, 105(7), 956-965.

Mohammadi, M., and Hajieghrari, B. (2011). *Sugarcane mosaic virus*: The causal agent of mosaic disease on sorghum (Sorghum bicolor L.) in Tehran province of Iran. Afr. J. Biotechno, 8(20).

Moradi, Z., Mehrvar, M., Nazifi, E., and Zakiaghl, M. (2016). The complete genome sequences of two naturally occurring recombinant isolates of *Sugarcane mosaic virus* from Iran. Virus genes, 52(2), 270-280.

Moradi, Z., Nazifi, E., and Mehrvar, M. (2017). Occurrence and Evolutionary Analysis of Coat Protein Gene Sequences of Iranian Isolates of *Sugarcane mosaic virus.* Plant Pathol. J, 33(3), 296.

Muhammad, B., Alegbejo, M., Kashina, B., and Banwo, O. (2016). Occurrence and distribution of *potyviruses* infecting sorghum in Kaduna and Kano States, Nigeria. Arch Phytopathology Plant Protect, 49(11-12), 281-292.

Muhire, B., Martin, D., Brown, J., Navas-Castillo, J., Moriones, E., Zerbini, F., and Varsani, A. (2013). A genome-wide pairwise-identity-based proposal for the classification of viruses in the genus *Mastrevirus* (family *Geminiviridae*). Arch virol, 158(6), 1411-1424.

Muhire, B., Varsani, A., and Martin, D. (2014). SDT: A virus classification tool based on pairwise sequence alignment and identity calculation. PloS one, 9(9), e108277.

Oana, D., Ziegler, A., Torrance, L., Gasemi, S., and Danci, M. (2009). *Potyviridae* family-short review. J. Hortic. For. Biotechnol, 13, 410-420.

Ohshima, K. (2013). Studies on the molecular evolution of *potyviruses.* J. Gen Plant Pathol, 79(6), 448-452.

Padhi, A., and Ramu, K. (2010). Genomic evidence of intraspecific recombination in *Sugarcane mosaic virus.* Virus genes, 42(2), 282-285.

Puchades Izaguirre, Y., La O, M., Montalván, J., Carvajal, O., Martínez, Y., Zardon, M., and Arencibia, A. (2015). Genetic and Symptomatic Characterization of *Sugarcane mosaic virus*(SCMV) in Cuba. Sugar tech, 18(2), 184-191.

Revers, F., and Antonio García, J. (2015). Molecular Biology of *Potyviruses.* In Advances in virus research (Vol. 92, pp. 101-199). Academic Press. \

Rybicki, E., and Pietersen, G. (1999). Plant virus disease problems in the developing world. In Advances in virus research (Vol. 53, pp. 127-175). Academic Press.

Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J., Neyer, F., and Aken, M. (2013). A gentle

introduction to bayesian analysis. Applications to developmental research. Child development, 85(3), 842-860.

Silva-Rosales, L., Alcala-Briseño, R., and Espejel, F. (2015). Sugarcane mosaic. Virus Diseases of Tropical and Subtropical Crops, 131.

Soltis, D., and Soltis, P. (2003). The role of phylogenetics in comparative genetics. Plant Physiol, 132(4), 1790-1800.

Souza, I. R. P., Giolitti, F., Carneiro, N., Lenardon, S., Oliveira, E., Gomes, E. A., and de Souza, F. (2012). Sequence diversity in the coat protein of SCMV infecting maize and sorghum in Brazil. Revista Brasileira de milho e Sorgo, 11(2), 120-136.

Spetz, C., Taboada, A. M., Darwich, S., Ramsell, J., F Salazar, L., and P T Valkonen, J. (2003). Molecular resolution of a complex of *potyviruses* infecting *solanaceous* crops at the centre of origin in Peru. J. Gen Virol, 84(9), 2565-2578.

Varsani, A., Martin, D., Navas-Castillo, J., Moriones, E., Hernández-Zepeda, C., Idris, A., and Brown, J. (2014). Revisiting the classification of *Curtoviruses* based on genome-wide pairwise identity. Arch virol, 159(7), 1873-1882.

Viswanathan, R. (2017). Detection of three major RNA viruses infecting sugarcane by Multiplex - Reverse Transcription - Polymerase Chain Reaction (Multiplex-RT-PCR). Australas Plant Path, 39(1), 79-84.

Viswanathan, R., and Balamuralikrishnan, M. (2005). Viswanathan, R. and M. Balamuralikrishnan (2005). Impact of Mosaic Infection on Growth and Yield of Sugarcane. Sugar Tech, 7 (1):61-65.

Viswanathan, R., Ramar, K., and Balamuralikrishnan, M. (2009). Identification of new variants of SCMV causing sugarcane mosaic in India and assessing their genetic diversity in relation to SCMV type strains. Virus Genes, 39(3), 375.

Wang, X.-Y., Li, W.-F., Huang, Y.-K., Zhang, R.-Y., Shan, H.-L., Yin, J., and Luo, Z.M. (2017). Molecular detection and phylogenetic analysis of viruses causing mosaic symptoms in new sugarcane varieties in China. European J. Plant Path, 148(4), 931-940.

Wu, L., Zu, X., Wang, S., and Chen, Y. (2012). *Sugarcane mosaic virus*– Long history but still a threat to industry. Crop protect, 42, 74-78.

Wylie, S., Adams, M., Chalam, C., Kreuze, J., López-Moya, J. J., Ohshima, K., and Sanfaçon, H. (2017). ICTV virus taxonomy profile: *Potyviridae.* J. Gen Virol, 98(3), 352-354.

X Jiang, J., and P Zhou, X. (2003). Maize dwarf mosaic disease in different regions of China is caused by *Sugarcane mosaic virus.* Archives of virology, 147(12), 2437-2443.

Xie, X., Chen, W., Fu, Q., Zhang, P., An, T., Cui, A., and An, D. (2016). Molecular Variability and Distribution of *Sugarcane mosaic virus*in Shanxi, China. PloS one, 11(3), e0151549.

Xu, D., Park, J.-W., Mirkov, T., and Zhou, G. H. (2008). Viruses causing mosaic disease in sugarcane and their genetic diversity in southern China. Arch virol, 153(6), 1031.

Yasmin, T., Iqbal, S., Farooq, M. A., Zubair, M., and Riaz, A. (2011). Prevalence, distribution and incidence of major sugarcane infecting viruses in NWFP and Punjab. Pak. J. Phytopathol, 23(1), 24-30.

Zhu, X.-G., P Lynch, J., Lebauer, D., Millar, A., Stitt, M., and P Long, S. (2015). Plants in silico: Why, Why Now and What? An integrative platform for plant systems biology research. Plant cell environ, 39(5), 1049-1057.