Analysis of Circulating Recombinant Forms (CRFs) of HIV-1 using Chaos Game Representation (CGR)

Adhikar Bansiwal

A dissertation submitted for the partial fulfilment of

BS-MS dual degree in Science

Indian Institute of Science Education and Research , Mohali

April 2014

Certificate of Examination

This is to certify that the dissertation titled " **Analysis of Circulating Recombinant Forms (CRFs) of HIV-1 using Chaos Game Representation (CGR)** " submitted by Mr. Adhikar Bansiwal (Reg. No. MS09007) for the partial fulfilment of BS-MS dual degree programme of the Institute, has been examined by the thesis committee duly appointed by the Institute. The committee finds the work done by the candidate satisfactory and recommends that the report be accepted.

Dr. Kuljeet Sandhu Dr. Shashi B. Pandit Prof. Somdatta Sinha

(Supervisor)

Declaration

The work presented in this dissertation has been carried out by me under the guidance of Prof. Somdatta Sinha at the Indian Institute of Science Education and Research Mohali. This work has not been submitted in part or in full for a degree, a diploma, or a fellowship to any other university or institute. Whenever contributions of others are involved, every effort is made to indicate this clearly, with due acknowledgement of collaborative research and discussions. This thesis is a bonafide record of original work done by me and all sources listed within have been detailed in the bibliography.

Adhikar Bansiwal

(Candidate)

MS09007.

Dated: April 25, 2014

In my capacity as the supervisor of the candidate's project work, I certify that the above statements by the candidate are true to the best of my knowledge.

Prof. Somdatta Sinha

(Supervisor)

Acknowledgement

Firstly, I would like to thank Lord for giving me an opportunity to do a MS project. I would like to thank all people who helped me and encouraged me during this project.

I am thankful to my supervisor Prof. Somdatta Sinha for her guidance all throughout this project. This thesis would not have been possible without her efforts. I sincerely want to thank her for inspiration, support and guidance.

I thank IISER Mohali for providing all infrastructure, financial support and computational facilities required in this project.

I would like to thank Aridaman Pandit, Ashutosh Srivastava and Priya VK for their help and useful discussions. I would also like to thank my other labmates - Arashdeep, Kanwal, Meenakshi, Srishti, Rivi and Preeti for their help, support and for providing a friendly environment in laboratory.

My special thanks to my friends Aditya Jhajharia, Anurag Kulshrestha, Harsh Katyayan, Abhishek Anand, Abhishek Mishra, Akash Sharma, Devender Yadav, Aditya Verma, Yatender Arya, Ankit Kukreja, Sushant Singh and Sapna Meena for their support and cheerful moments, which made me stress-free during this project.

I am very much thankful to my MS thesis committee members Dr. Kuljeet Singh Sandhu and Dr. Shashi Bhushan Pandit for their valuable suggestions regarding computational work and their moral support.

In the end, I would like to thank all my family members, who supported me all through my studies and it is because of their prayers, support and encouragement that helped me finish this project.

List of Figures

List of Tables

Notation

bp : base pairs

NJ : Neighbor Joining Algorithm

k : word-length

cpz : chimpanzees

U : unclassified region in HIV whole genome

d(i,j) : pairwise distance matrix

Contents

Abstract

Human Immunodeficiency Virus (HIV) is the causative agent for Acquired Immune Deficiency Syndrome (AIDS). It exhibits very high genetic diversity with different variants and subtypes. Classification of these subtypes is thus essential for monitoring epidemic. Current methods of classification include specific genes-based phylogenetic analysis, but these methods showed certain inconsistencies in classification of subtypes in past. However, recent alignment free methods, like Chaos Game Representation (CGR), have been shown to be successful in classification of HIV subtypes at word length k=6 (Sinha, Pandit ; 2010). This method is not only computationally less intensive, but can also analyze whole genome variations.

Problem with HIV classification becomes more complex as different HIV subtypes can recombine and form Circulating Recombinant Forms (CRFs). These CRFs continuously emerge over time and circulate into host population. They show variable susceptibility to drugs.

In my 5th year MS project, my aim was to test if these CRFs could also be classified using the CGR method. Being recombinants of subtypes, the variation in the sequences are expected to be quite low. My studies are presented in this thesis in the following sections.

Chapter -1 of this thesis is an introduction to HIV subtypes and CRFs. It also introduces basics of CGR plotting and classification using CGR method. Chapter-2 gives an overview of various software tools, algorithms and other computational methods used in this work.

In Chapter-3, the results are shown for classification of the CRFs using the CGR method. I checked the effect of lowering word-length and it is shown that again k=6 is the minimum word-length required for correct classification. In cladograms generated it was reported that CRFs clustered with those parental subtypes that have the largest length in the genome. Chapter-4 deals with reduction in word-set, and it was seen that correct clustering can still be obtained even by selecting lesser number of words.

Base composition analysis of these selected words was performed and it was reported that these words were mostly A-rich.

Chapter-5 shows the use of certain HIV genes, instead of whole genome, to classify CRFs properly using CGR method. It shows the drawback of this method in analyzing short genomic sequences.

Lastly, Chapter-6 discusses a simple software tool created in PHP and HTML to generate CGR and to calculate base composition of the given input sequence.

Chapter -1

Introduction

1.1 Introduction to Human Immunodeficiency Virus (HIV)

Human Immunodeficiency Virus (HIV) is a retrovirus that causes Acquired Immunodeficiency Disease Syndrome (AIDS). HIV targets hosts immune cells and weakens them so that host can become susceptible to any opportunistic infections. HIV is broadly classified into two different types HIV-1 and HIV-2. Both of these subtypes have been zoonotically transferred to humans. HIV-1 originated from Simian Immunodeficiency Virus (SIV cpz) that infects chimpanzees, whereas HIV-2 originated from SIV that infects Sooty mangabey. HIV-1 is more prevalent across world and is majorly responsible for AIDS throughout the world.

HIV-1 is further divided into 3 groups , M (major) , O (outlier) , N (non-major, non-outlier). Among these Group M is globally most prevalent. Group M is further classified into 9 different subtypes (A,B,C,D,F,G,H,J,K) based on their genetic variations. (Fig. 1.1)

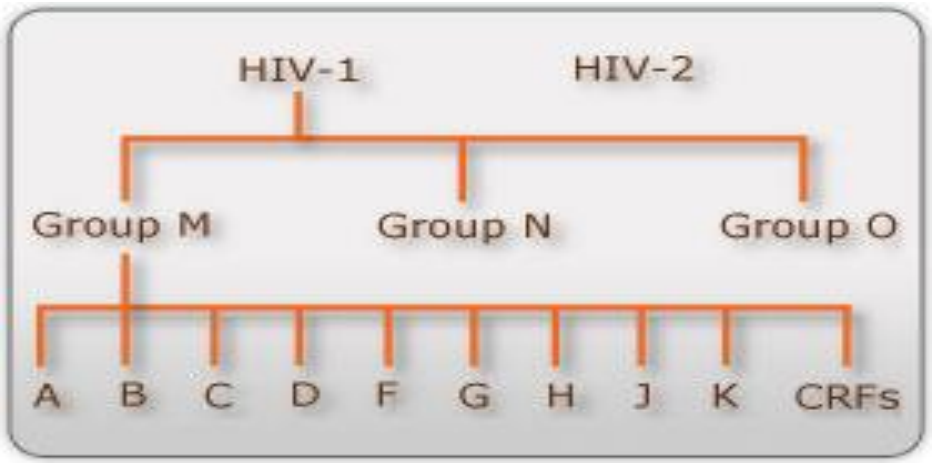


Fig. 1.1 : HIV classification into groups and subtypes

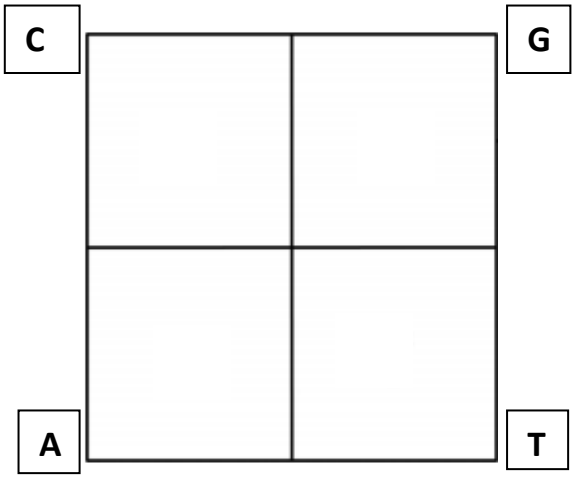
Subtypes A and F are further classified into sub-subtypes (A1,A2) and (F1,F2). However, complexity may increase further as different HIV subtypes can recombine with each other and form Circulating Recombinant Forms (CRFs). These newly formed CRFs continuously emerge and circulate in the host population. Hence, identification and classification of newly emerging subtypes and CRFs is essential for keeping track of AIDS pandemic. (Leitner *et al*. 2005)

1.2 Introduction to Chaos Game Representation(CGR) –

Chaos Game Representation (CGR) is a two-dimensional representation of genomic sequences using chaos game algorithm. This chaos game algorithm was given by M. Barnsley in his book "Fractals Everywhere". However , H.J. Jeffery was first to apply this algorithm for visual representation of genomic sequences (Jeffery; 1990). Earlier this technique was restricted for visualization purposes only but later it was applied for studying phylogenetic relationship among diverse species (Almeida, 2001).

This method was also applied to study intra-species variability by classifying HIV-1 subtypes using CGR method at k=6. (Sinha, Pandit ; 2010)**.**

CGR is plotted by taking a square box (800 x 800) and labeling its vertices as A,T,G and C (as shown below).



The mid-point of square box is taken as starting point and the first nucleotide is read. Position of next point is calculated as mid-point between previous point and vertex representing the next

nucleotide read. This process is repeated until all nucleotides are plotted and a CGR is obtained (Fig. 1.2 shows the CGR of the HIV-HXB2 genome – the first one to be sequenced).

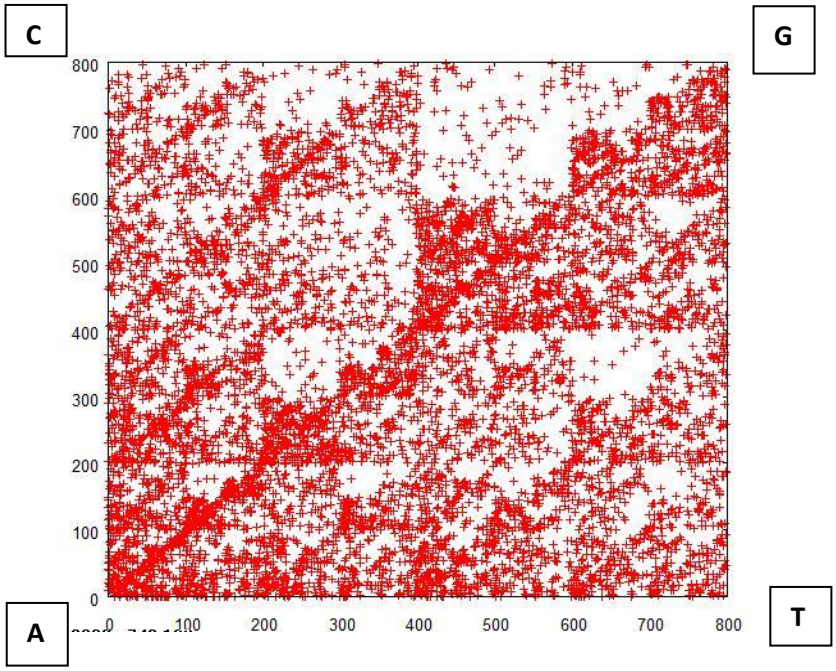


Fig. 1.2 : CGR of HIV-HXB2 (9719bp)

As Chaos game is a fractal representation, each square box is self-replicating image and thus can be divided into di-, tri- and higher nucleotides. (Fig. 1.3)

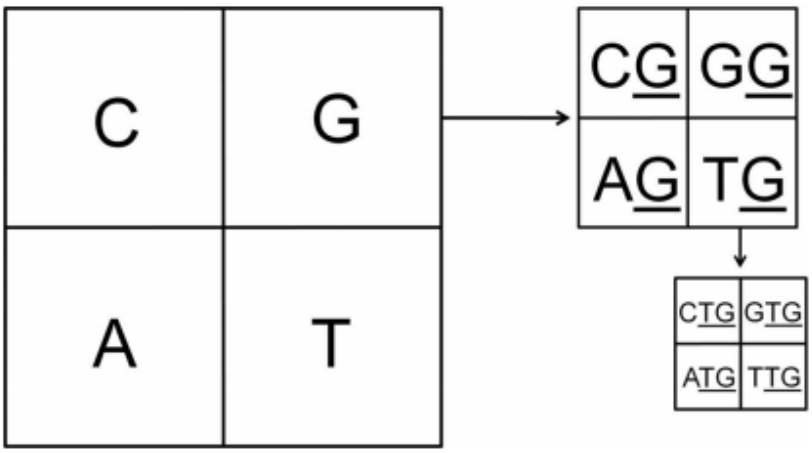


Fig 1.3 : Division of CGR to obtain di-, tri- and higher nucleotide frequencies

Now, distance between two CGR is calculated using Euclidean distance.

$$d(A,B) = \sqrt{\sum_{i=1}^{2^k}\sum_{j=1}^{2^k}(a_{i,j} - b_{i,j})^2}$$

a and *b* are the frequencies of a given word in two different CGRs, A and B, while d(A,B) is the Euclidean distance between genomes. These distances between genomes were then used to construct cladograms. Same approach was used by (Sinha , Pandit ; 2010) to classify HIV subtypes using CGR method at k=6.

1.3 Objective of the work –

The objective of this work is to check if CRFs of HIV-1 could be classified and analyzed effectively using the CGR method. Since k=6 means 4096 possible words to be compared for each pair of genomes, I have also studied if a smaller subset of words can be used to do the classification. This then would help us to devise effective methods to classify any unknown CRF and predict its parental subtypes. I have also developed a simple software tool for analysis of genome of any length using the CGR method.

Chapter-2

Material and Methods

2.1 Overview –

HIV sequences were downloaded from the HIV database (http://www.hiv.lanl.gov). These sequences were then analysed using CGR method and cladograms were created. In order to generate CGR, programs were written in C++ and CGR co-ordinates were plotted using GNUPLOT. A set of HIV-1 whole genome sequences were given input to program as a multiple FASTA file. C++ program then processes this input file as per given word-length 'k' and pair-wise distance matrix between set of input genome is generated as output. Using Neighbor Joining (NJ) Algorithm of PHYLIP 3.69 clustering was performed and cladograms were generated. Lastly, cladograms were edited, rooted and labeled using MEGA 5.

2.2 Material –

All 55 HIV-1 CRFs available on Los Alamos database were downloaded. Filtering of search results on database was done using sequence tool. Filtering of sequences was done by using various subtypes CRFs as filtering parameter.

For clustering of the HIV whole genome sequences, 37 sequences of different subtypes were used (Table 2.1) from HIV-1 reference dataset (Leitner *et al.* 2005). Lengths of all downloaded HIV whole genome sequences were in range of 9000bp-9700bp.

Table 2.1: List of HIV-1 reference dataset sequences (Leitner *et al.* 2005)

Subtype	Sequence	Acc. No.	Year of sampling	Country of sampling (origin)
A1	92UG037.1	U51190	1992	Uganda
A1	Q23	AF004885	1994	Kenya
A1	SE7253	AF069670	1994	Sweden (Somalia)
A1	UG57136	AF484509	1998	Uganda
A2	CDKTB48	AF286238	1997	DRC
A2	CY017	AF286237	1994	Cyprus
B	HXB2	K03455	1983	France
B	BK132	AY173951	1990	Thailand
B	671	AY423387	2000	Netherlands
B	1058	AY331295	1998	USA
C	ETH2220	U46016	1986	Ethopia
C	92BR025.8	U52953	1992	Brazil
C	IN21068	AF067155	1995	India
C	SK164B1	AY772699	2004	South Africa
D	ELI	K03454	1983	DRC
D	94UG114.1	U88824	1994	Uganda
D	4412HAL	AY371157	2001	Cameroon
D	A280	AY253311	2001	Tanzania
F1	93BR020-1	AF005494	1993	Brazil
F1	VI850	AF077336	1993	Belgium (DRC)
F1	FIN9363	AF075703	1993	Finland
F1	MP411	AJ249238	1996	France
F2	MP255	AJ249236	1995	Cameroon
F2	MP257	AJ249237	1995	Cameroon
F2	0016BBY	AY371158	2002	Cameroon
F2	CM53657	AF377956	1997	Cameroon
G	SE6165	AF061642	1993	Sweden (DRC)
G	HH8793.1.1	AF061640	1993	Finland (Kenya)
G	DRCBL	AF084936	1996	Belgium (DRC)
G	NG083	U88826	1992	Nigeria
H	056.1	AF005496	1990	Cent. Afr. Rep.
H	VI991	AF190127	1994	Belgium (?DRC)
H	VI997	AF190128	1993	Belgium (?DRC)
J	SE9280.9	AF082394	1994	Sweden (DRC)
J	SE9173.3	AF082395	1993	Sweden (DRC)
K	EQTB11C	AJ249235	1997	DRC
K	MP535	AJ249239	1996	Cameroon

Apart from this, 4 SIV(cpz) reference sequences were also downloaded to be used as an outgroup in cladogram. (Table 2.2)

Table 2.2 : SIV(cpz) reference dataset sequences used as an outgroup for clustering (Leitner *et al.* 2005)

Sequence		Acc. No.	Year of sampling	Country of sampling (origin)
CPZ	GAB	X52154	NA	Gabon
CPZ	ANT	U42720	NA	DRC
CPZ	TAN1	AF447763	2000	Tanzania
CPZ	CAM5	AJ271369	1998	Cameroon

GeneCutter –

It is an online sequence alignment and extraction tool available on Los Alamos HIV database (http://www.hiv.lanl.gov/content/sequence/GENE_CUTTER/cutter.html). It inputs nucleotide sequences, aligns them against a reference sequence (HXB2 in case of HIV-1) and extracts required genes from whole genome.

Geography tool –

Apart from using HIV Los Alamos database for downloading HIV whole genome sequences and GeneCutter, I also used Geography tool available on database to get information about current HIV CRFs and subtypes scenarios worldwide.

2.3 About Software and Algorithms used –

(i) Dev C++

All C++ programs were written, compiled and executed using Dev C++ version 4.9.9.2. It is freely available under GNU General Public Licence and can be downloaded from http://www.bloodshed.net/devcpp.html.

(ii) PHYLIP 3.69

PHYLIP (PHYLogeny Ineferring Package) is a package developed by Joseph Felsenstein. It contains various programs which include parisomy , distance matrix and likelihood based methods to infer phylogenies. Using PHYLIP 3.69, Neighbor Joining (NJ) Algorithm was used which is a distance matrix based method and generates output tree in Newick format. (Felsentein ; 1985)

(iii) Neighbor Joining Algorithm

It was created by Saitou and Nei in 1987 as a clustering method. It takes distance matrix between various pairs as inputs and calculates Q-matrix as per equation given below.

$$Q(i,j) = (r-2)d(i,j) - \sum_{k=1}^{r} d(i,k) - \sum_{k=1}^{r} d(j,k)$$

where r denotes total number of genomes in distance matrix and d(i,j) denotes distance matrix.

Once Q-matrix is constructed, pair having lowest Q-matrix is clustered together (i.e. closest neighbor are joined together). This process of calculating Q-matrix and clustering closest neighbor is iterated until all pairs are clustered and thus an output tree is generated.
(Saitou , Nei ; 1987)

(iv) MEGA 5

MEGA is an integrated tool for sequence analyses. It was developed by Tamura *et al*. It consists of various algorithms, models but I purposely used it for viewing Newick trees (cladograms) generated from PHYLIP (See section (ii) and (iii) above). Rooting , labeling and other editing of trees was also done using MEGA 5. (Tamura *et al.* 2011)

Chapter – 3

Analysis of CRFs

3.1 Introduction –

As discussed in chapter 1, identification and classification of CRFs is a tedious task and is currently dependent on phylogenetic methods based on individual gene-based phylogenies.CRFs continuously emerges over time and circulate into host population. These CRFs can be simple bi-recombinants or complex recombinants of multiple subtypes. (Fig. 3.1 and 3.2)

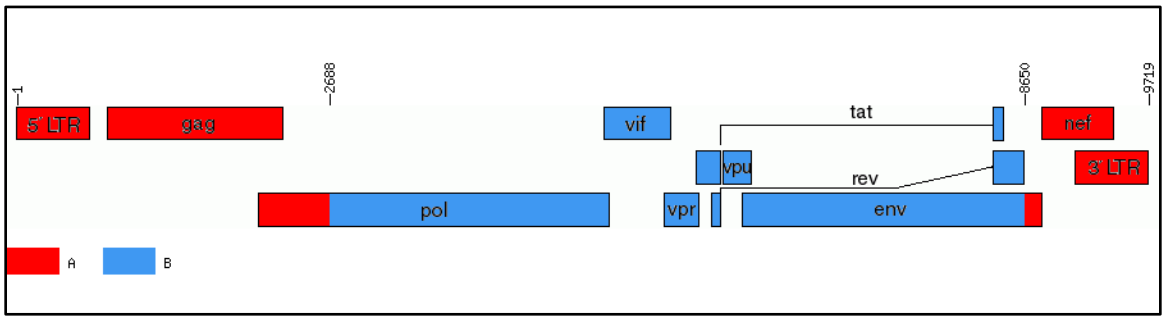


Fig 3.1 : **CRF03** : recombinant of subtype A and Subtype B (A=38% ; B=62%) Total genome length is 9719bp.

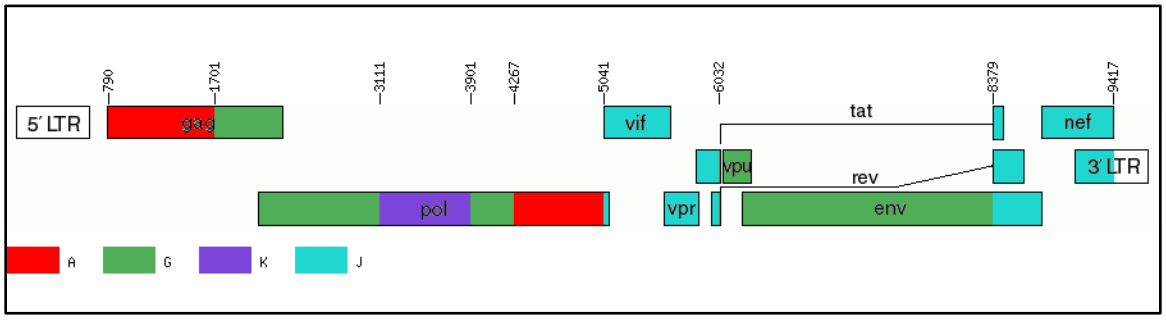


Fig 3.2 : **CRF06** : recombinant of 4 subtypes A,G,J,K (A=20%, G=46% , K=10% and J=24%). Total genome length is 8627bp.

Apart from this it is also reported that CRFs can recombine with other subtypes to generate new CRFs. (Fig.3.3)

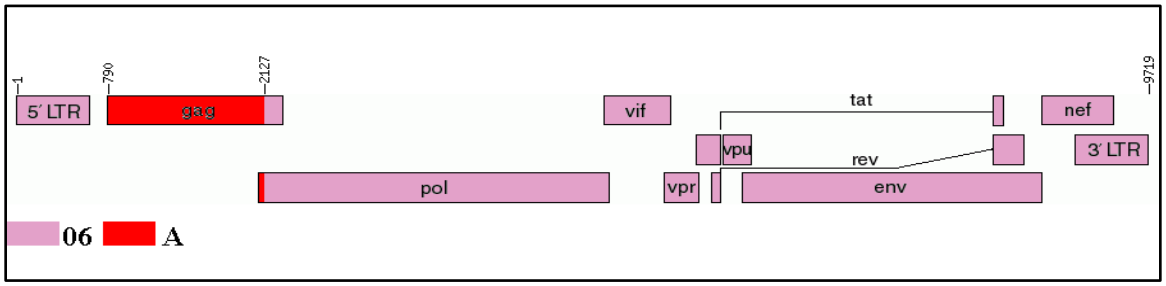


Fig 3.3 : **CRF32** : Recombinant between CRF06 and subtype A. (CRF06 = 86% and A= 14%) Total genome length is 9719bp.

Recombination is an ongoing process and as these newly emerging CRFs evolves, it is essential to identify them and track them to see how particular subtype or CRF is emerging and spreading globally.

3.2 Distribution of CRFs worldwide –

Presently there are 55 different CRF genomes identified and present in HIV Los Alamos database. Each CRF is named by giving a number which is in order of timeline when particular CRF was identified. Hence, CRF01 was first CRF identified by Carr *et al.* in 1996. Note, there is sharp increase in number of cases of CRFs as reported after 2003 (Lau *et al.* 2013) (Fig 3.4).

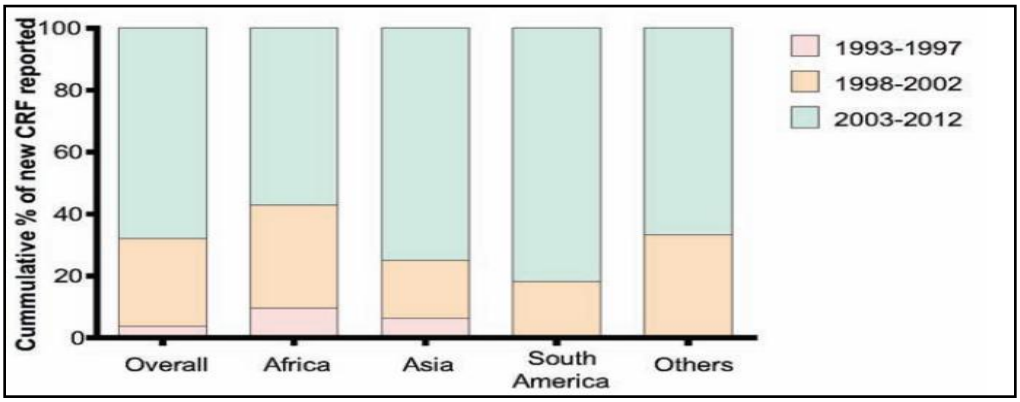


Fig. 3.4: Percentages of new CRFs reported during specific year-intervals.

It is clearly evident that earlier CRFs were mostly found in Africa and Asia regions but as time progresses new CRFs were identified in other parts of world as well.

Table 3.1 : Various CRFs in HIV Los Alamos database

Total CRFs available on HIV Los Alamos database	55	
Number of Bi-recombinants	42	
Number of Complex recombinants	13	
Subtype with maximum recombinants	B	(27 CRFs out of 55)
Number of CRFs between B and F1	11	
Number of CRFs between B and G	4	
Number of CRFs between B and C	3	
Number of CRFs between B and CRF01	9	

Table 3.1 clearly shows that subtype B is highly recombinant as it is found in nearly 50% of CRFs reported.

Using Geography tool present in HIV Los Alamos database, I looked into the current scenario of subtypes and recombinants worldwide and in India.

It is seen that subtype B is most prevalent globally (Fig. 3.5) (57.2% of all reported cases). While among all recombinants CRF01_AE has maximum occurrence (48.3% of all reported recombinant cases). (Fig. 3.6)

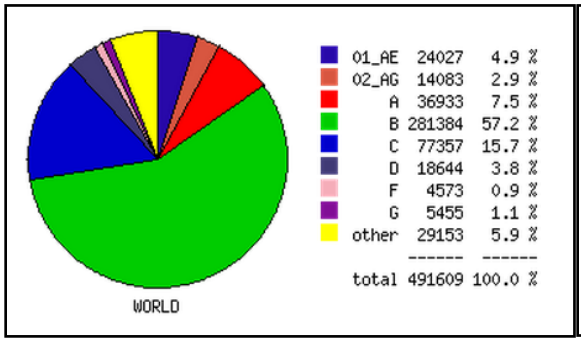


Fig. 3.5 : Distribution of subtypes worldwide

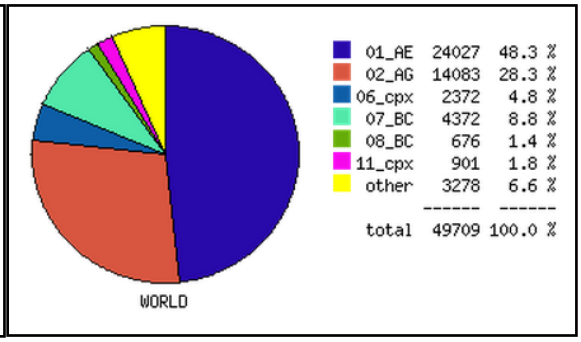


Fig. 3.6 : Distribution of CRFs worldwide

When similar analysis using geography tool is performed for India , it is seen that most HIV cases in India is due to subtype C (95.8% of all reported cases).

From past studies it is well established that subtype C is highly prevalent in South Asia mostly in India and China. (Lau *et al.* 2013)

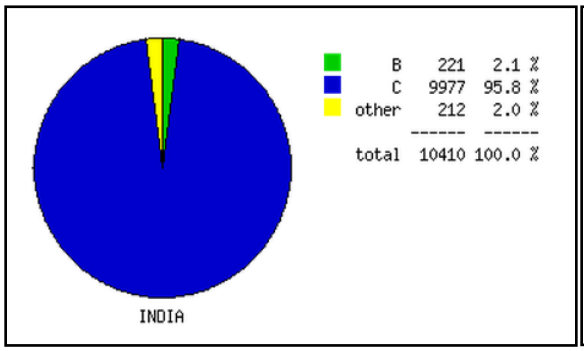


Fig. 3.7 : Distribution of subtypes in India

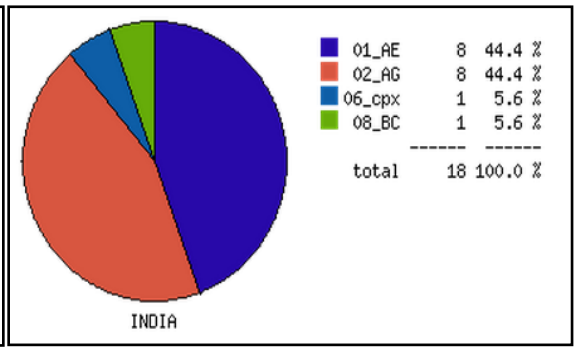


Fig. 3.8 : Distribution of CRFs in India

3.3 CGR analysis of CRFs –

CGR were generated for CRFs as well as parental subtypes. It was observed that visually all of them look practically the same. Hence, just by visual analysis we cannot extract any useful information from CGR of these CRFs and parental subtypes.(Fig. 3.9)

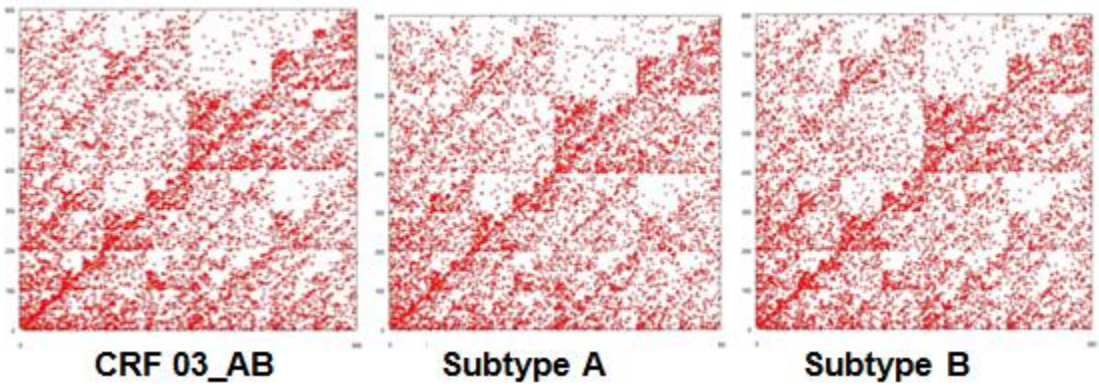


Fig. 3.9 : CGR of CRF03_AB and parental subtypes A and B

CGR had been used to classify HIV subtypes (Pandit, Sinha ; 2010) successfully at minimum word-length k=6. I wanted to study if the same method could be applied for CRFs as well. For this, first, I examined if the CRFs can be classified at lower word-lengths. Then I studied the clustering of CRFs in a cladogram, along with other subtypes, at k=6.

3.3.1 Classification of CRFs at lower word-length –

Once it is established that CRFs clusters with parental subtype having maximum recombination length, then effect of word-length on classification was checked. As we know from previous studies (Sinha,Pandit; 2010) that minimum word-length k=6 is required for classification of HIV subtypes using CGR method. So, word-length was reduced below k=6 to see if CRFs can be clustered at lower word-lengths.

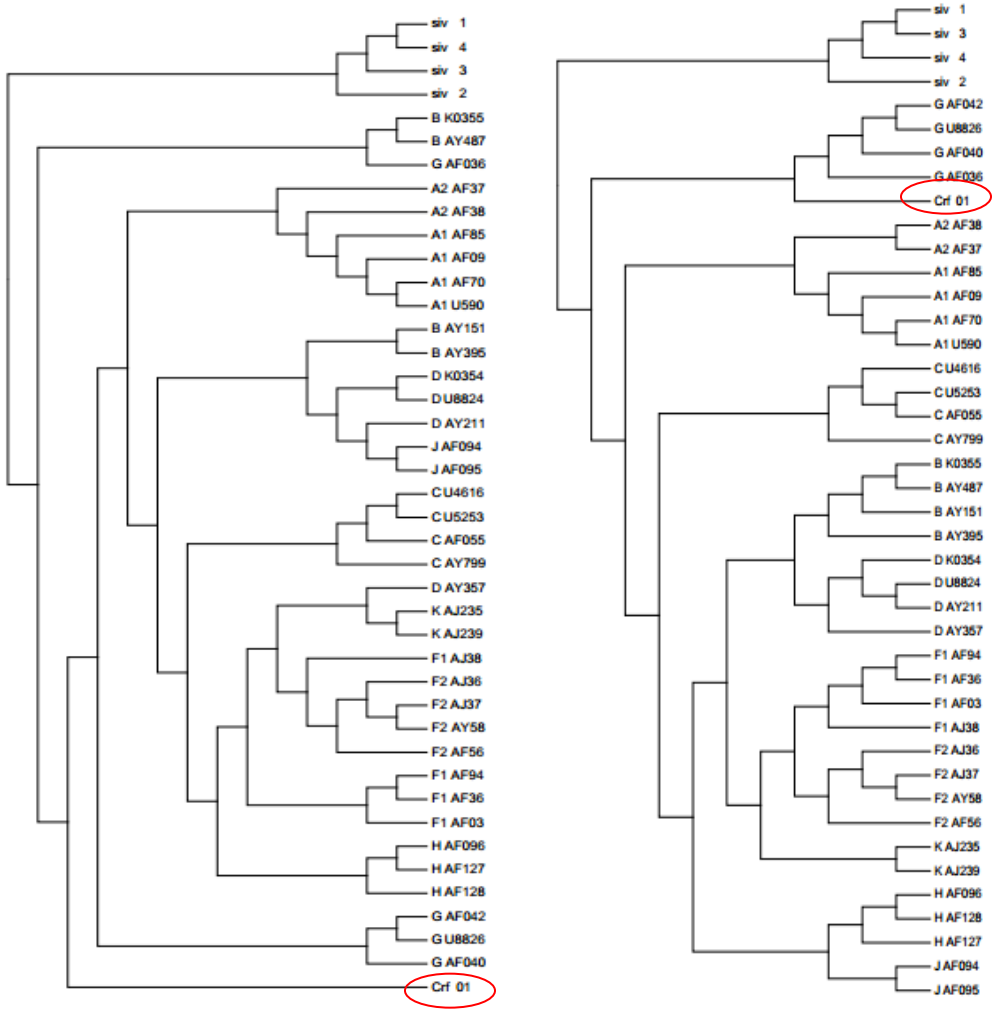


Fig 3.10 : **CRF01** : (A) k=4; (B) k=5 - CRF01 has A=69% and E/U =31%

Note: E was earlier defined as a subtype but later it was found that it always exists as a recombinant only, hence its status as a subtype was removed, so CRF01 is made up of subtype A and U (Unclassified) region.

For k=4, CRF01 should cluster with subtype A but it did not. Also subtypes B, G, D were not classified correctly (Fig. 3.10(A)). But when word-length was increased to k=5, then subtypes were clustered properly but CRF clustered with the wrong subtype (Fig.3.10(B))

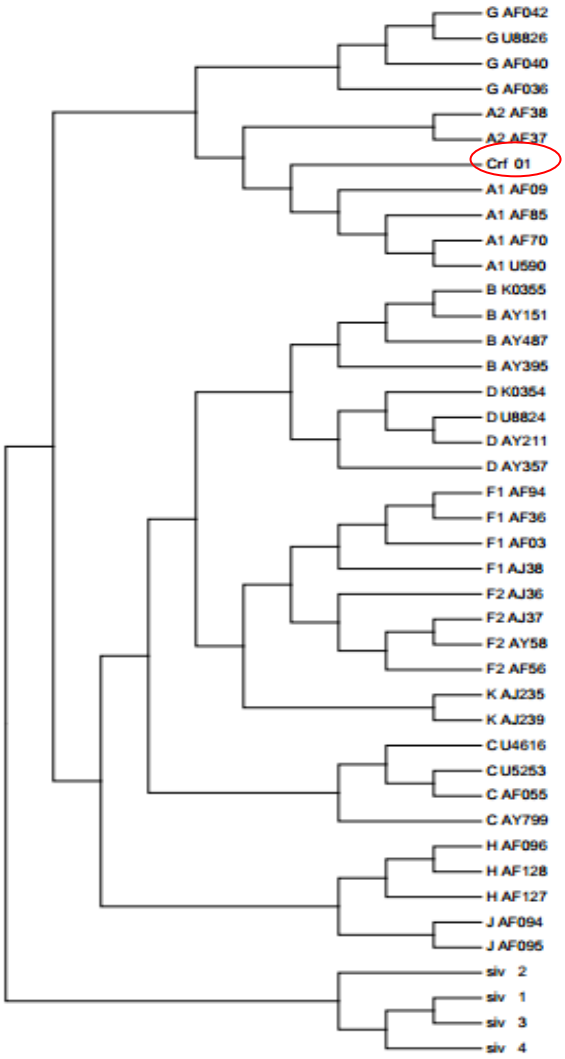


Fig. 3.11: **CRF01** at k=6 (All subtypes and CRF are properly clustered)

Thus, it is seen that for CRF classification as well, minimum word-length k=6 is required. (Fig.3.11)

3.3.2 Positioning of CRFs in cladograms at k=6 –

CRFs along with other HIV-1 subtypes and SIV (outgroup) were analysed and clustered as per methods mentioned in Chapter-2.

Clustering was performed using four SIV_{cpz} reference dataset sequences as an out-group. (Table 2.2)

Figures Fig.3.12 to Fig. 3.64 show the cladograms for all the CRFs present in HIV Los Alamos database. (Listed in Table 3.2)

It is seen that consistently the bi-recombinant CRFs cluster with that parental subtype, which has comparatively larger length in the genome.

For CRFs that were complex recombinants, it was seen that all, except five complex CRFs, cluster with parental subtypes having larger length in genome. These 5 CRFs are discussed later after Table 3.2.

First letter in each legend denotes subtype and followed by accession number of the given sequence.

(Cladograms on next page Fig. 3.12 to Fig. 3.64)

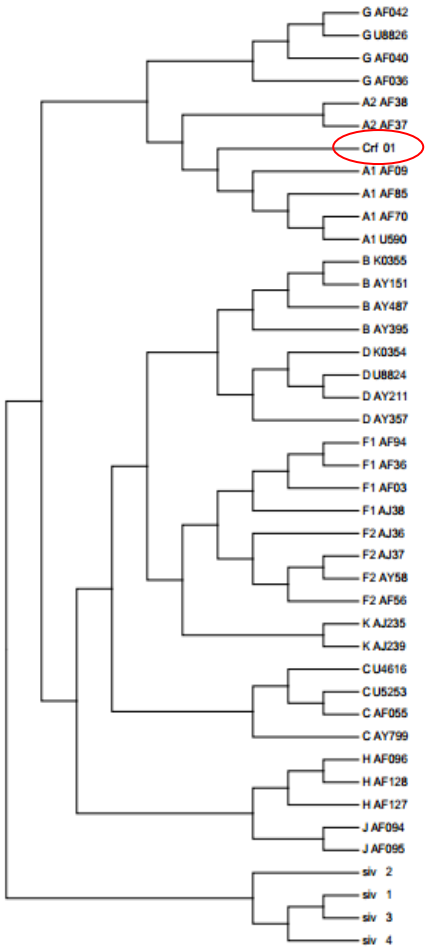


Fig. 3.12 : CRF01_AE

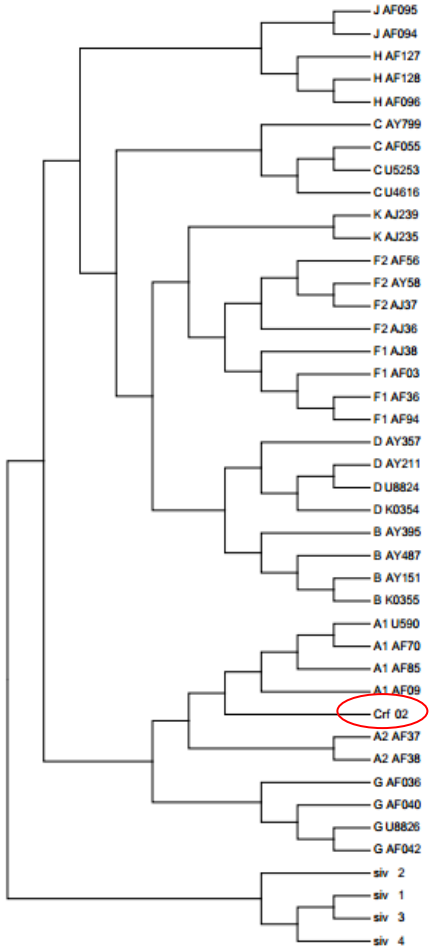


Fig. 3.13 : CRF02_AG

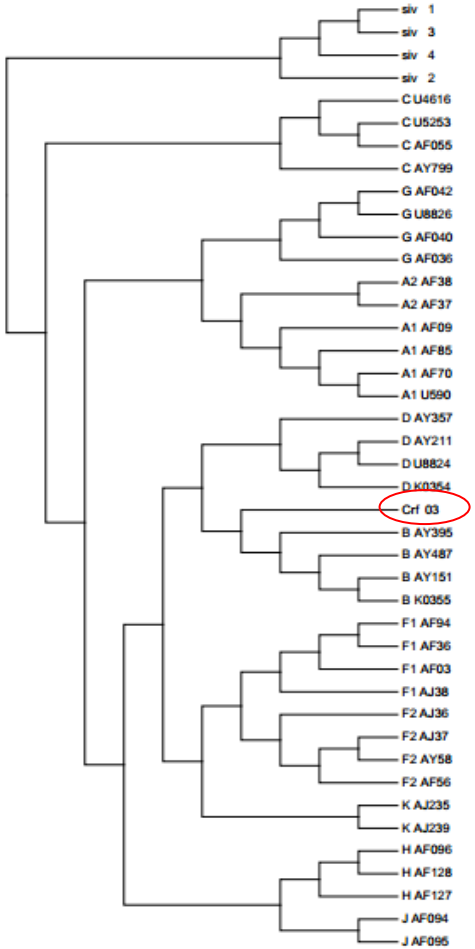


Fig. 3.14 : CRF03_AB

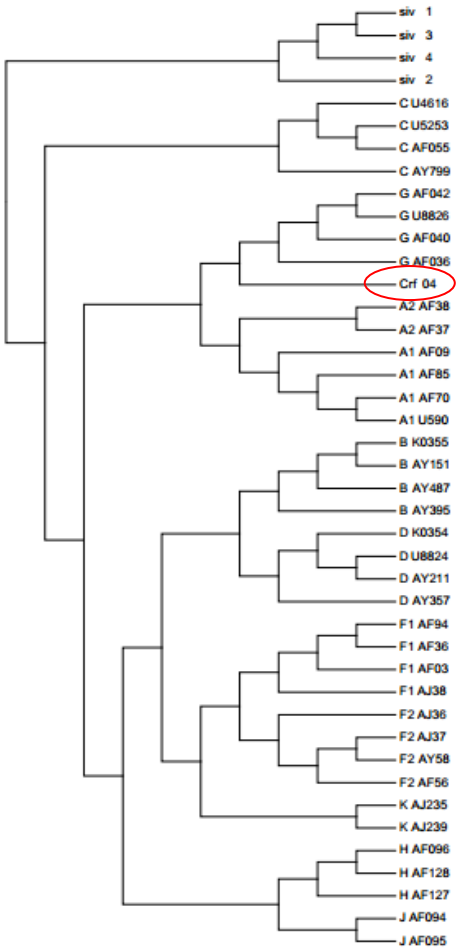


Fig. 3.15 : CRF04_cpx

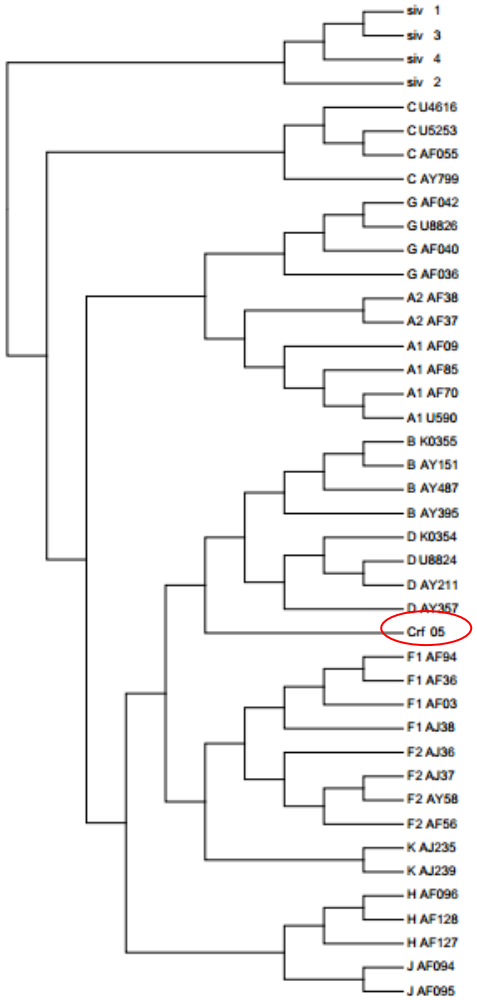


Fig. 3.16 : CRF05_DF

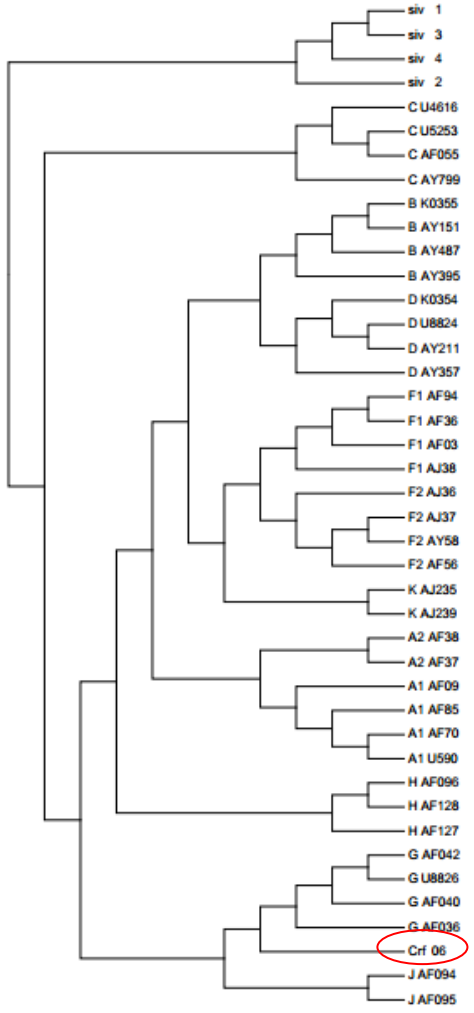


Fig. 3.17 : CRF06_cpx

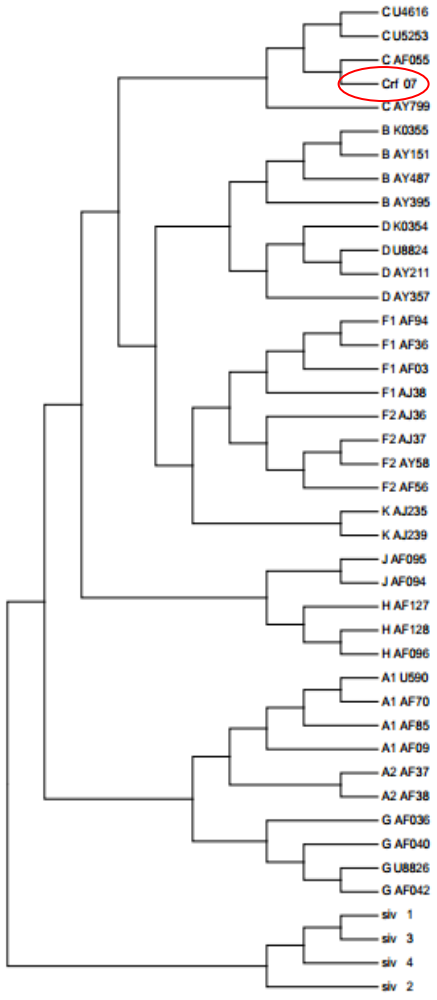


Fig. 3.18 : CRF07_BC

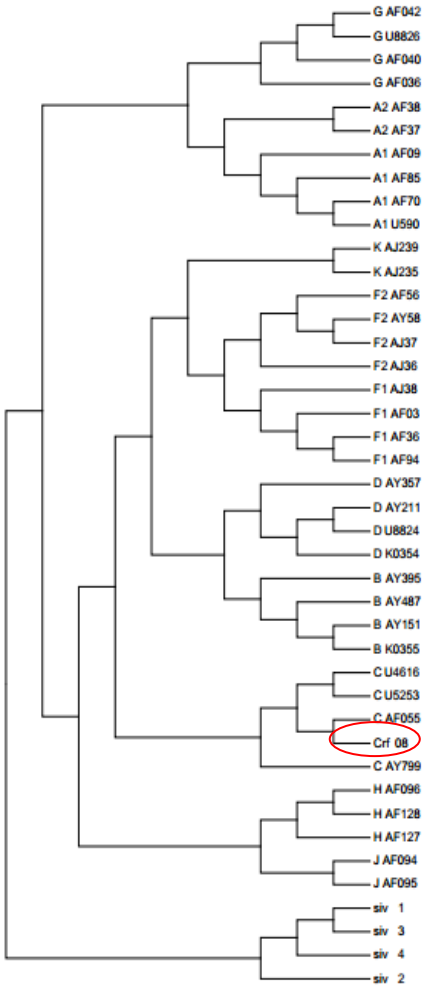


Fig. 3.19 : CRF08_BC

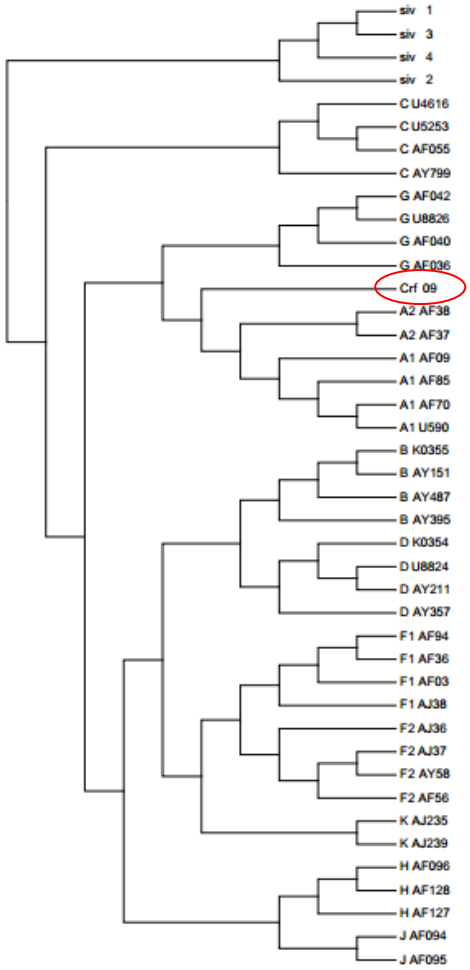


Fig. 3.20 : CRF09_cpx

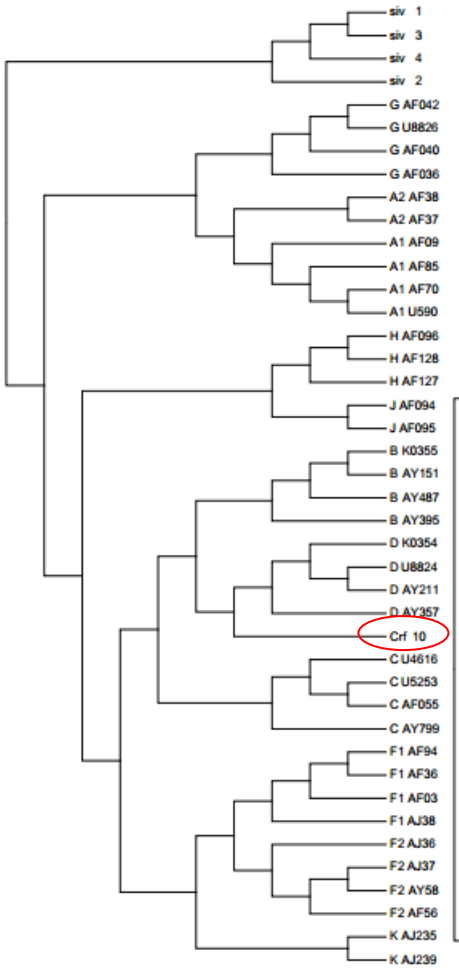


Fig. 3.21 : CRF10_CD

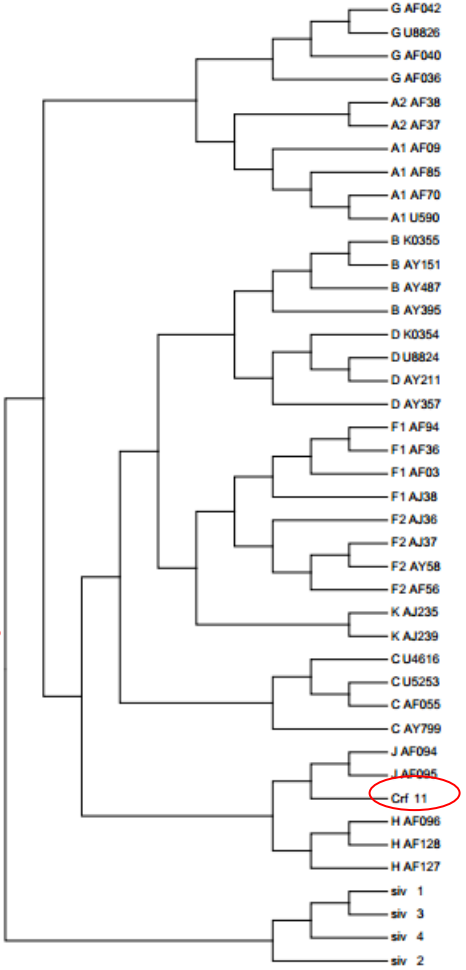


Fig. 3.22 : CRF11_cpx

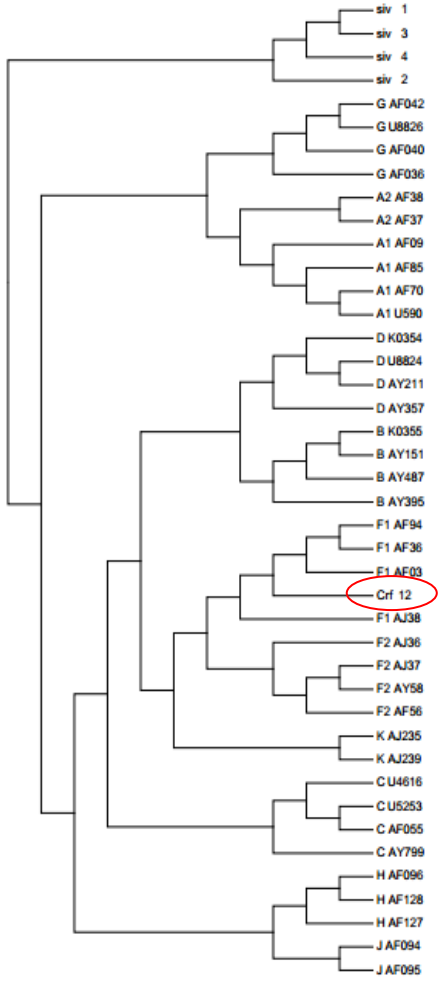


Fig. 3.23 : CRF12_BF

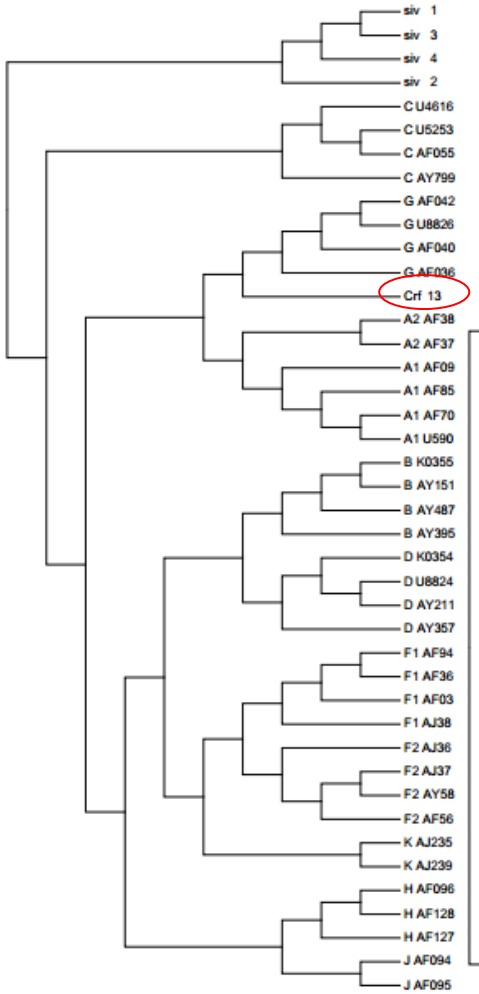


Fig. 3.24 : CRF13_cpx

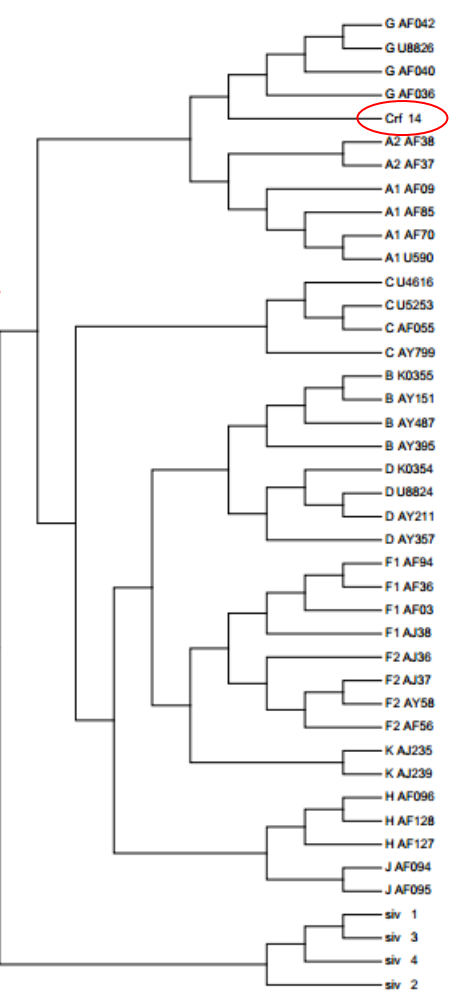


Fig. 3.25 : CRF14_BG

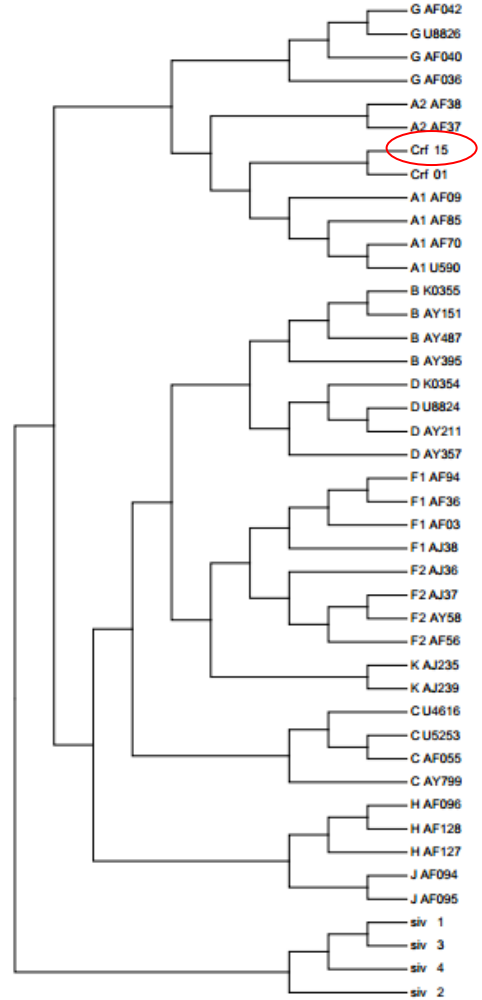


Fig. 3.26 : CRF15_01B

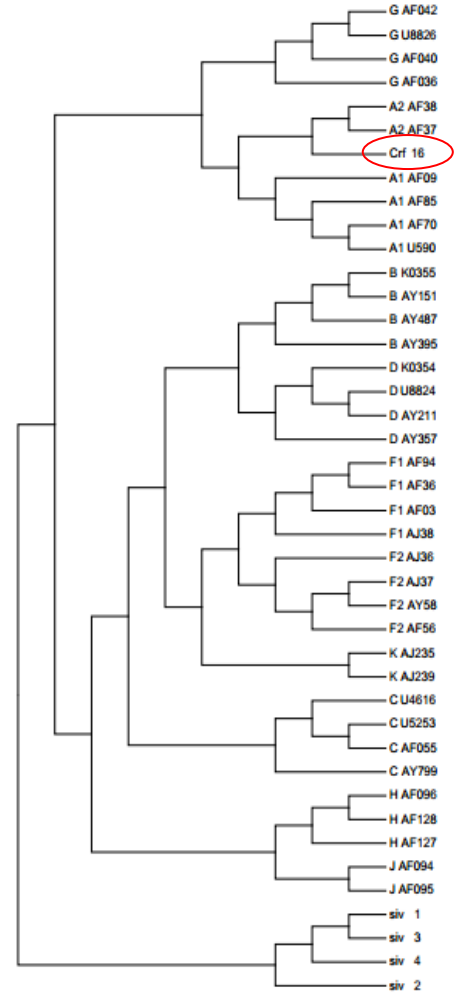


Fig. 3.27 : CRF16_A2D

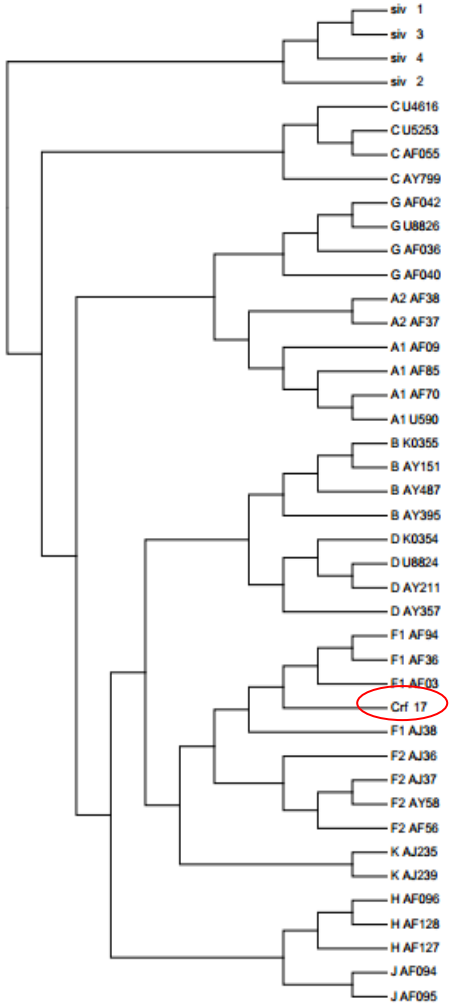


Fig. 3.28 ： CRF17_BF

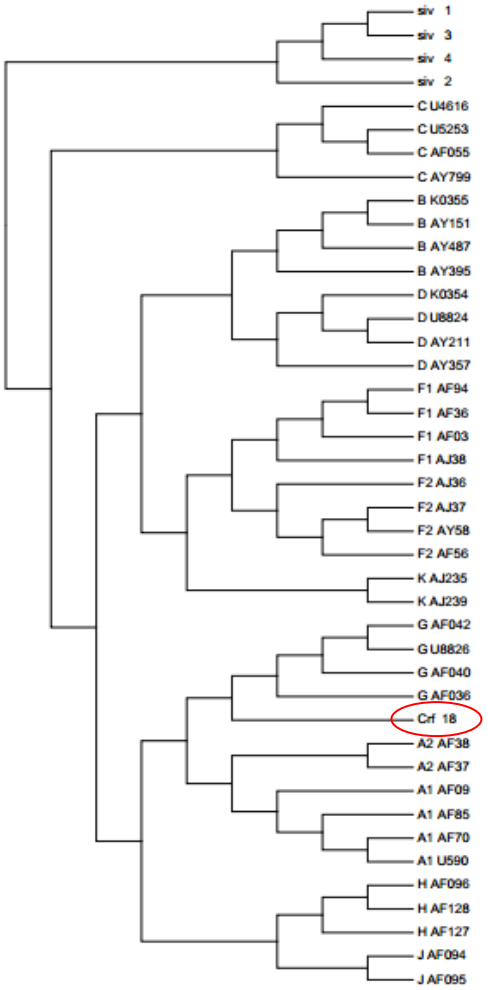


Fig. 3.29 ： CRF18_cpx

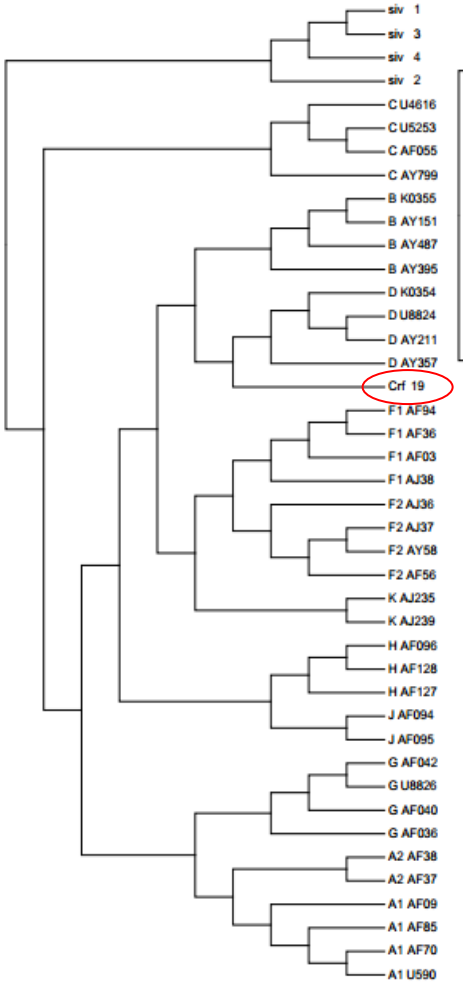


Fig. 3.30 ：CRF19_cpx

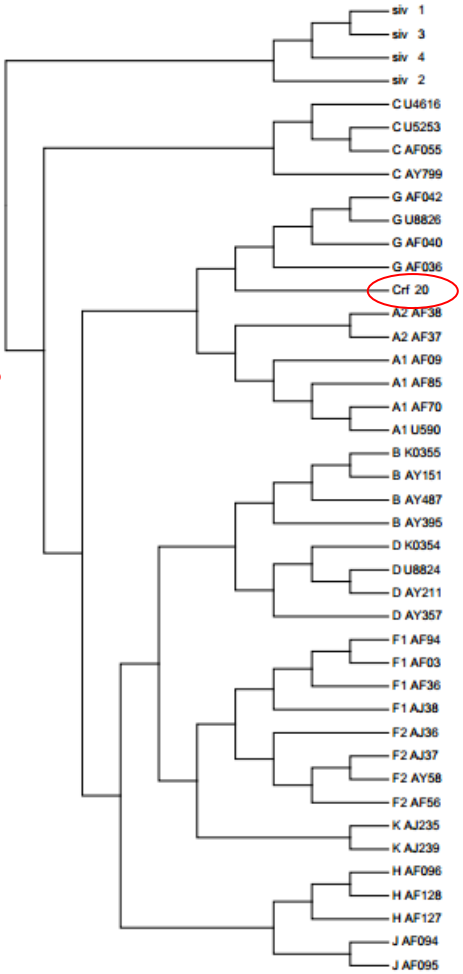


Fig. 3.31 ： CRF20_BG

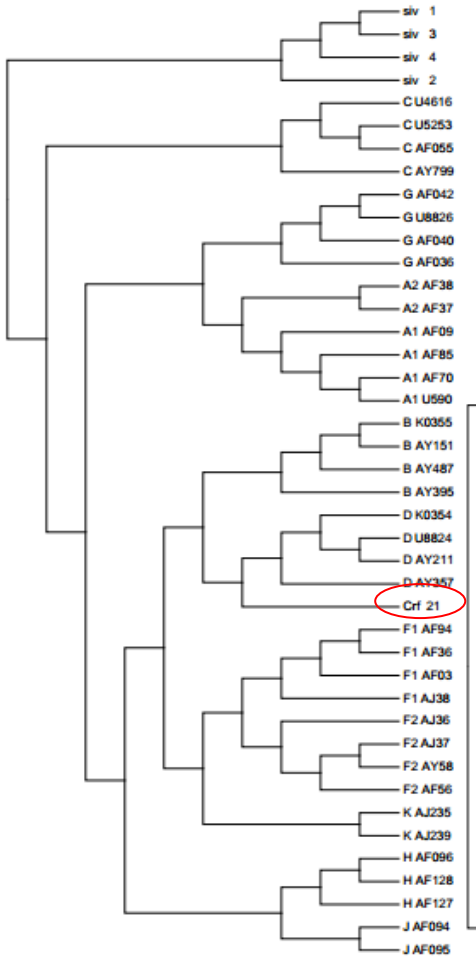


Fig. 3.32 : CRF21_A2D

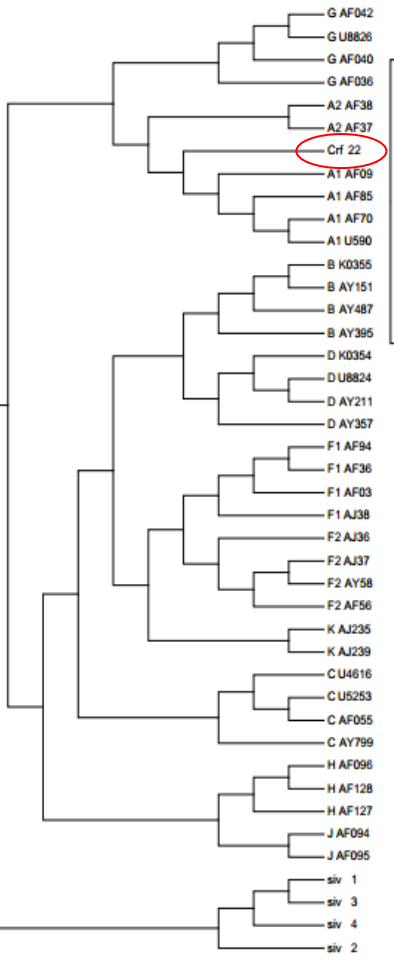


Fig. 3.33 : CRF22_01A1

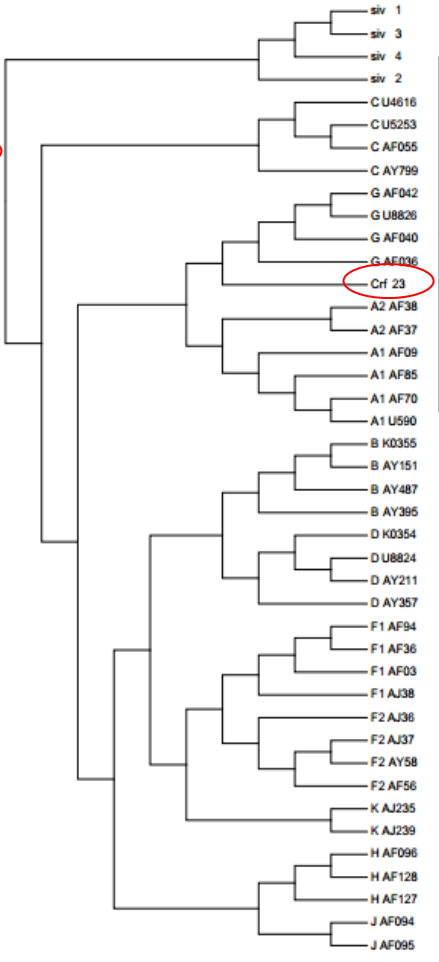


Fig. 3.34 : CRF23_BG

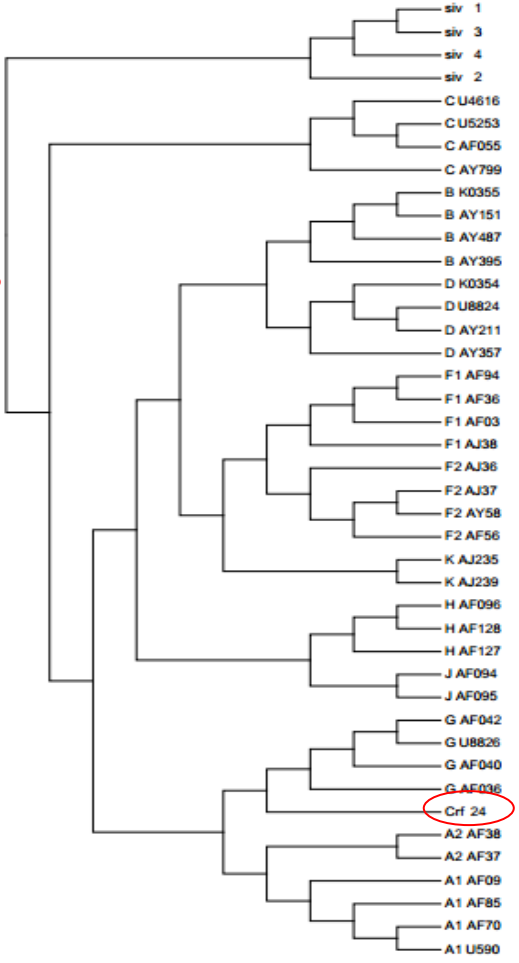


Fig. 3.35 : CRF24_BG

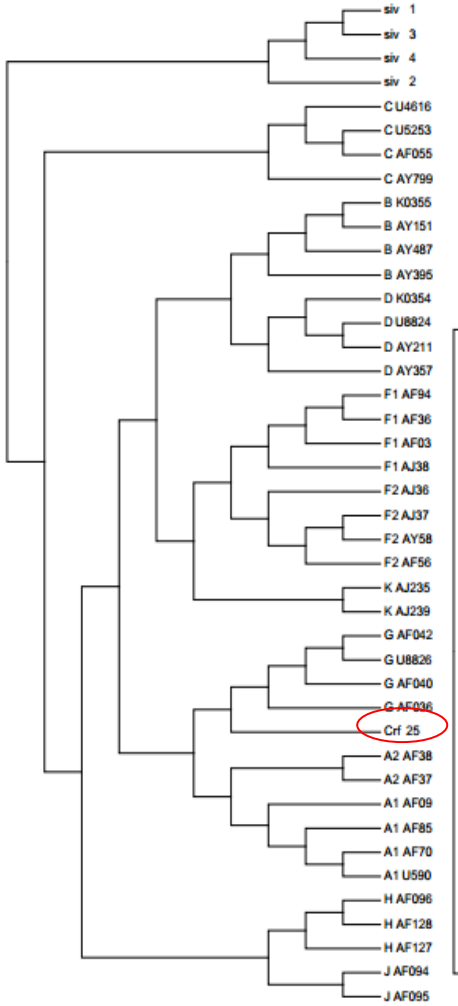


Fig 3.36 : CRF25_cpx

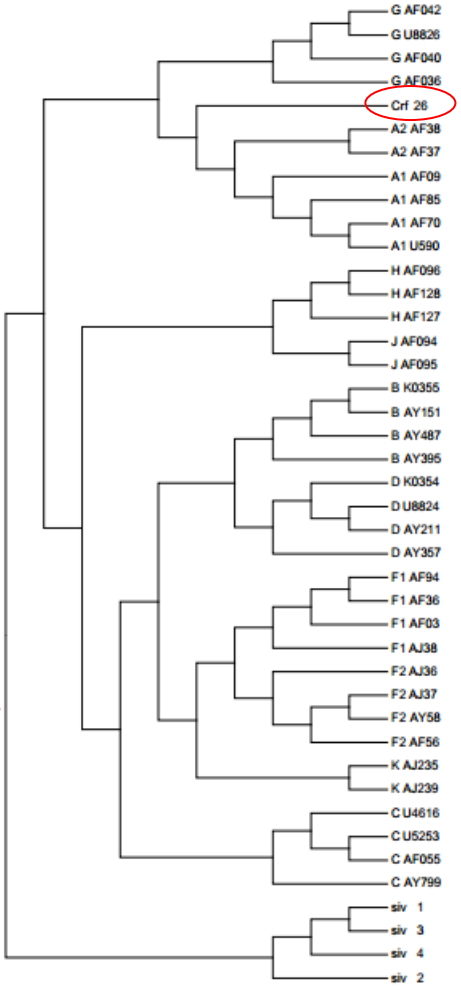


Fig. 3.37 : CRF26_AU

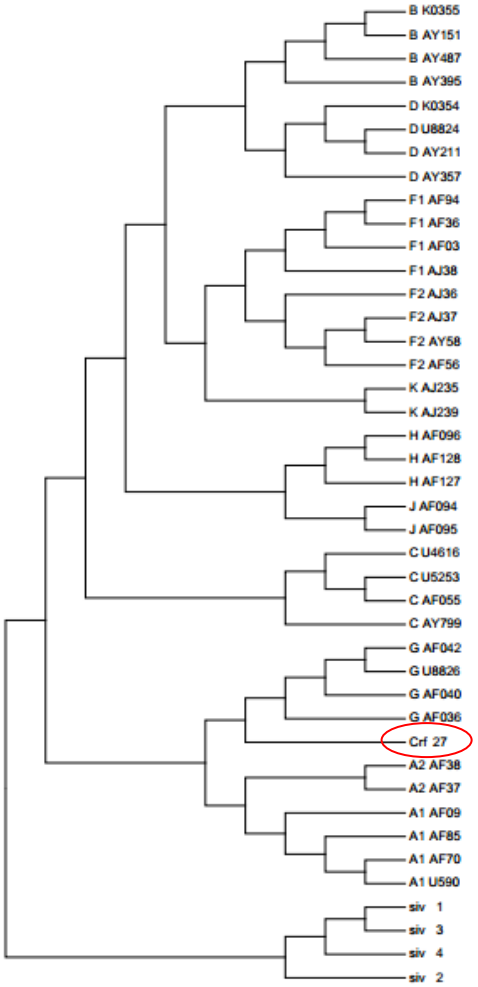


Fig. 3.38 : CRF27_cpx

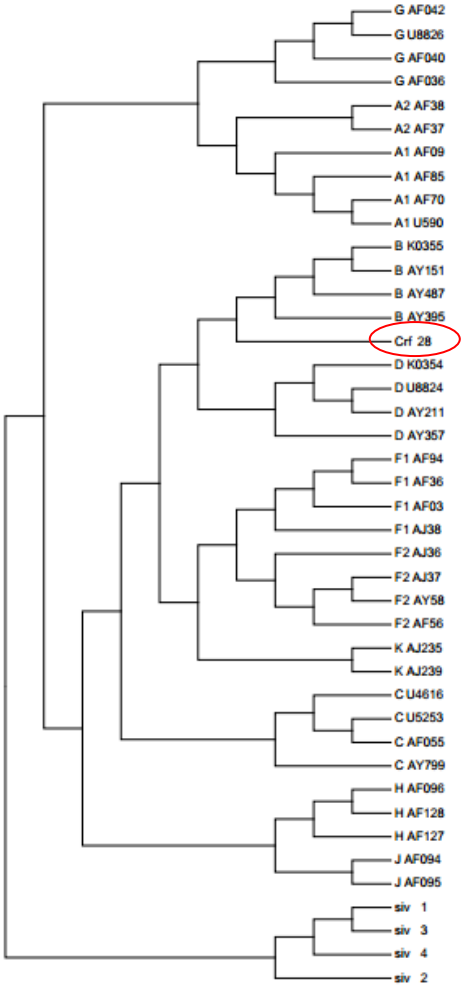


Fig. 3.39 : CRF28_BF

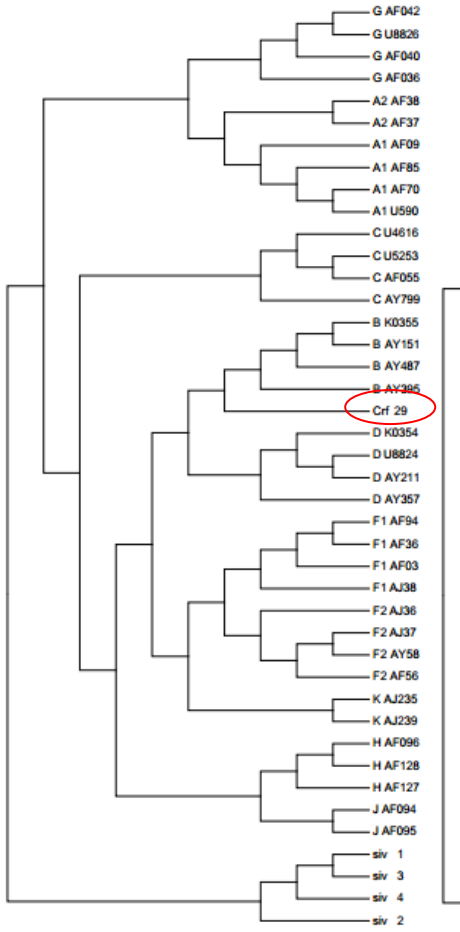


Fig. 3.40 : CRF29_BF

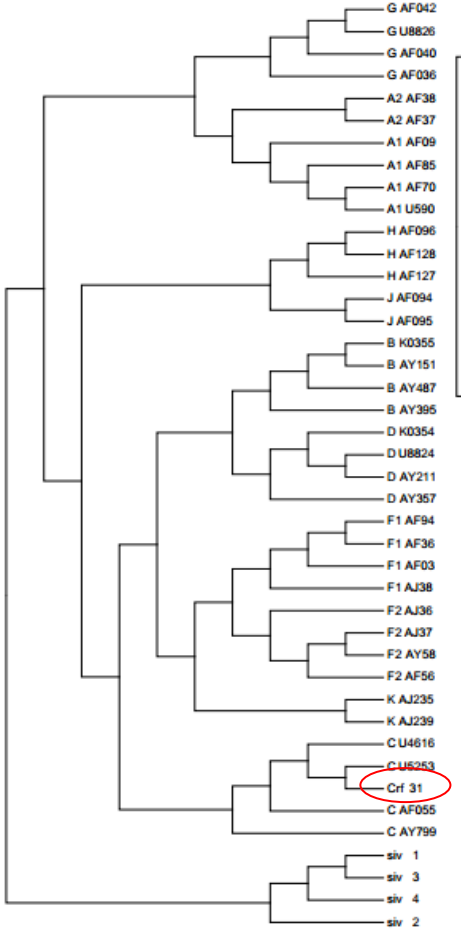


Fig. 3.41 : CRF31_BC

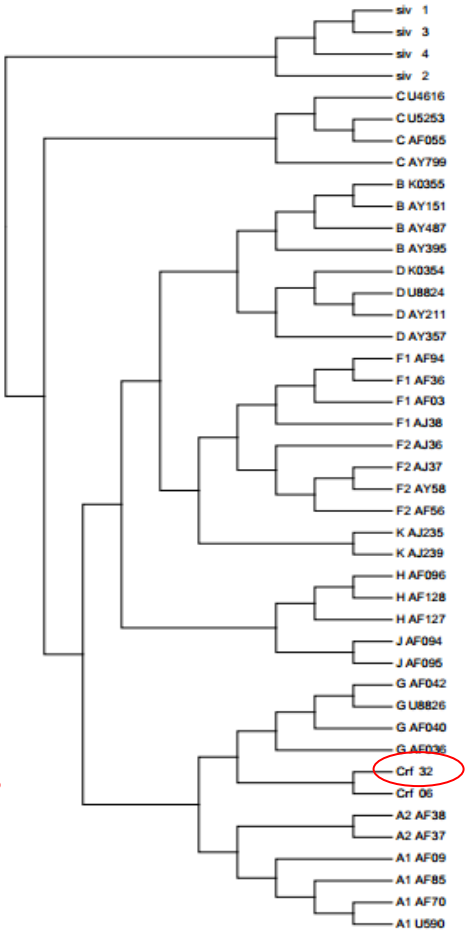


Fig. 3.42 : CRF32_06A

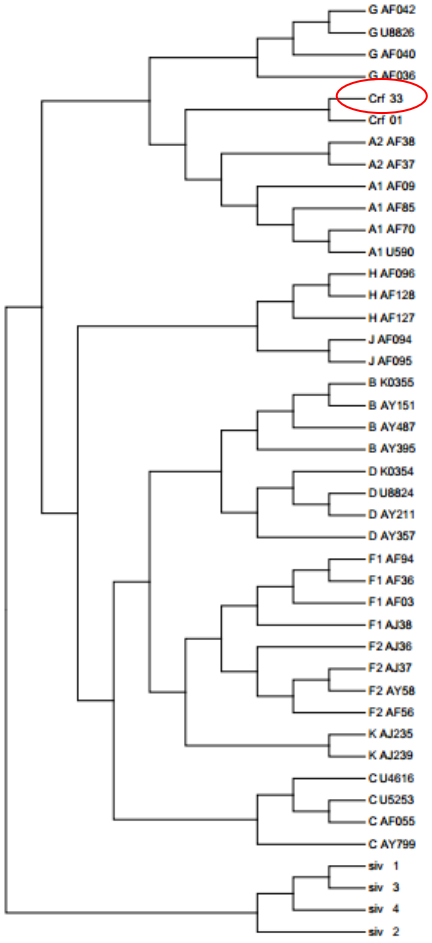


Fig. 3.43 : CRF33_01B

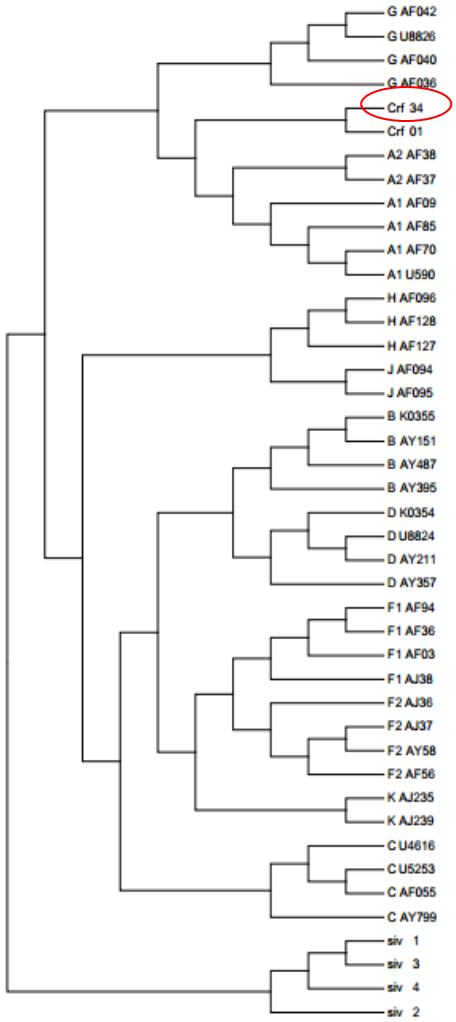


Fig. 3.44 : CRF34_01B

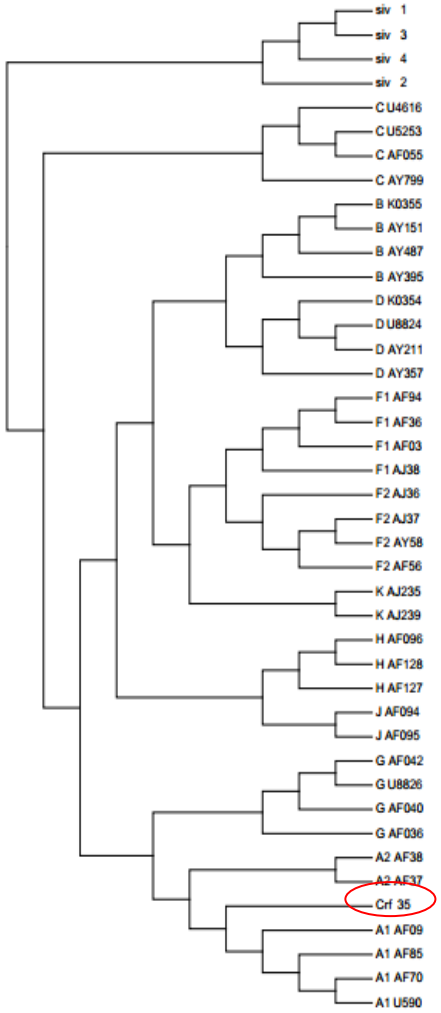


Fig. 3.45 : CRF35_AD

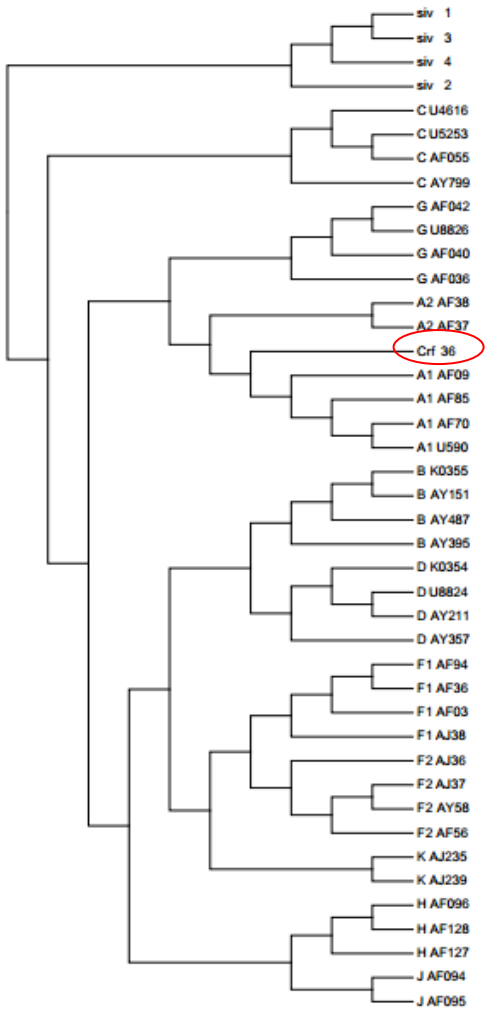


Fig. 3.46 : CRF36_cpx

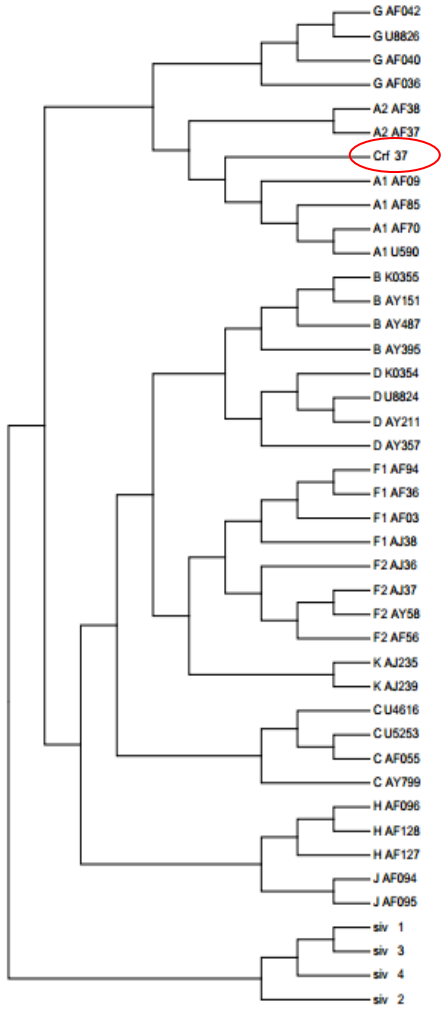


Fig. 3.47 : CRF37_cpx

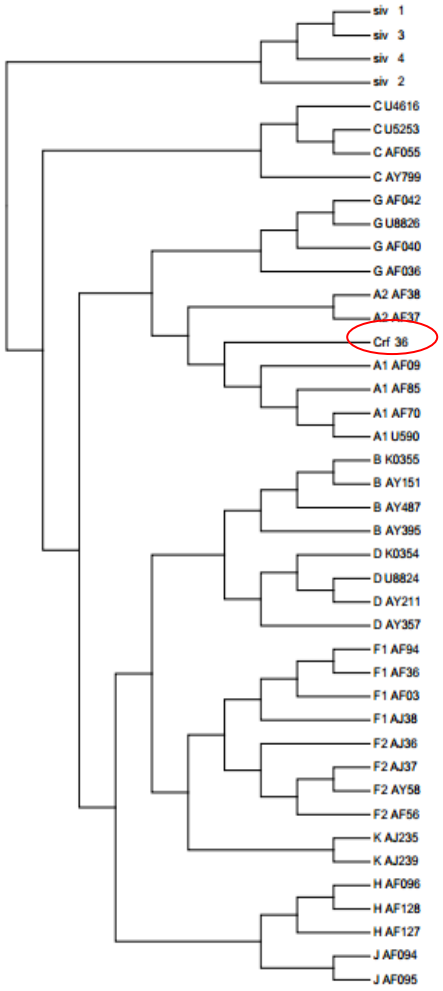


Fig. 3.48 : CRF36_cpx

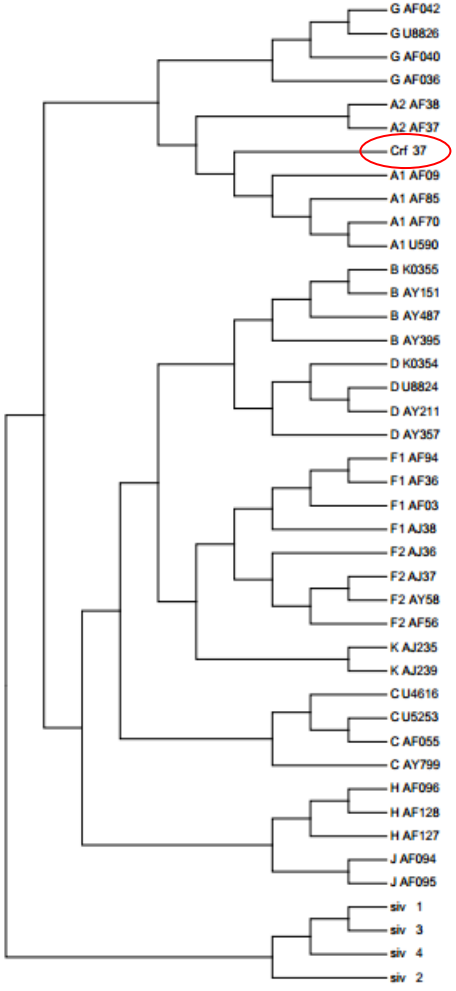


Fig. 3.49 : CRF37_cpx

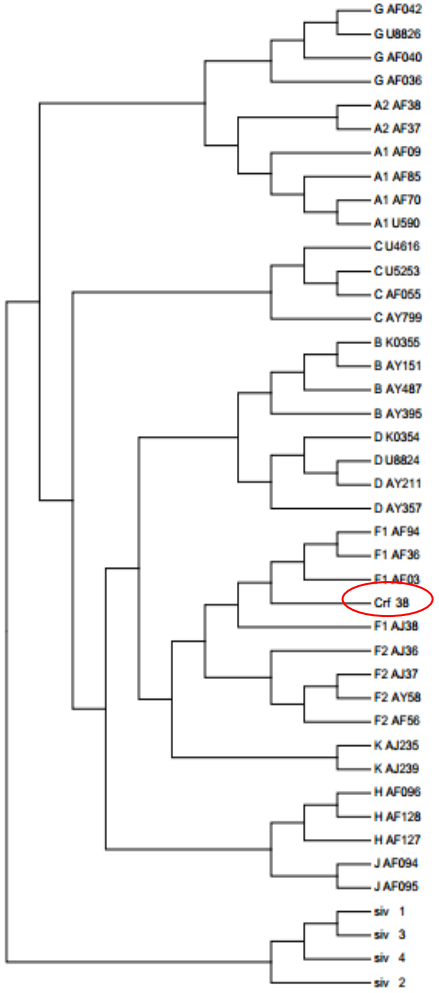


Fig. 3.50 : CRF38_BF

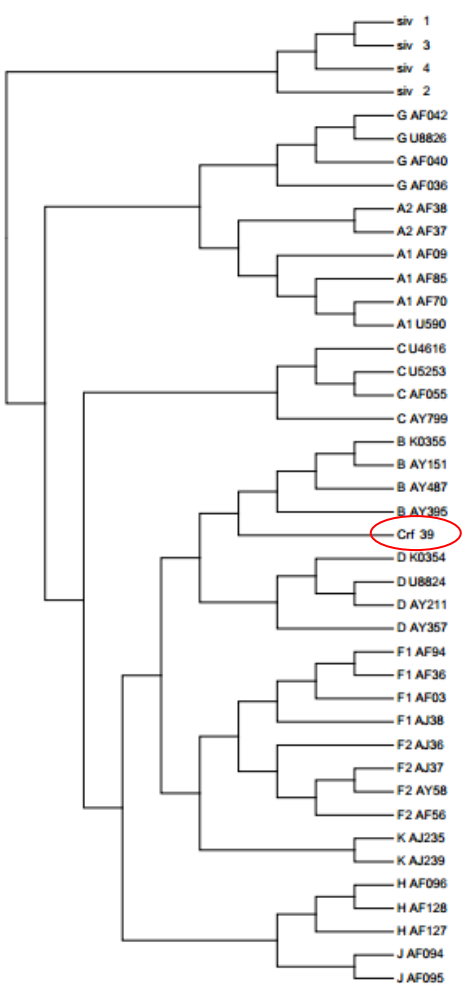


Fig 3.51 : CRF39_BF

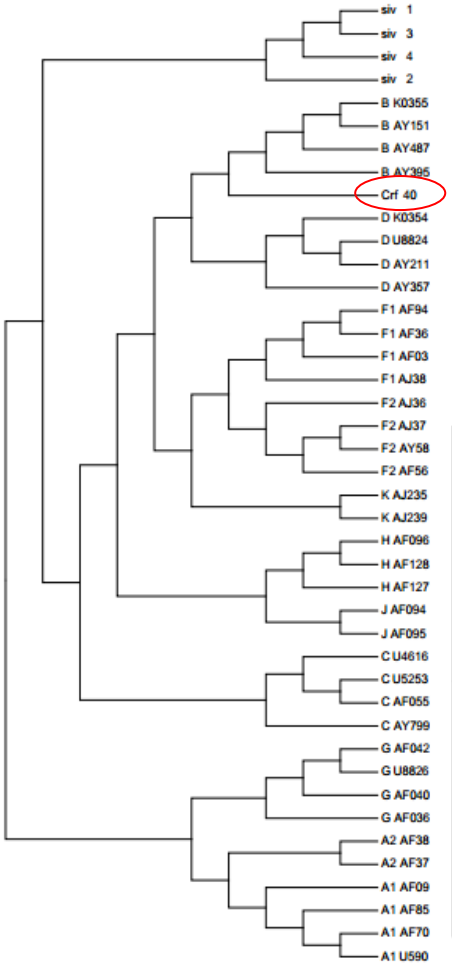


Fig. 3.52 : CRF40_BF

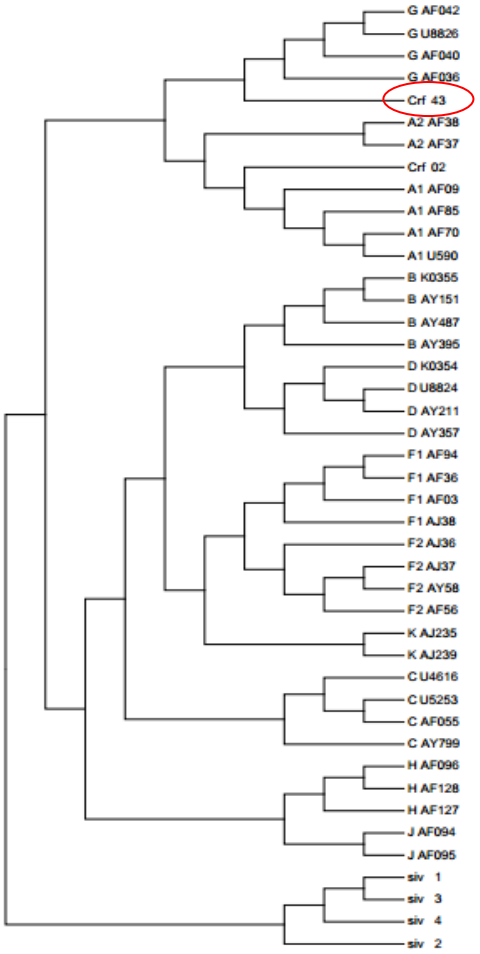


Fig. 3.53 : CRF43_02G

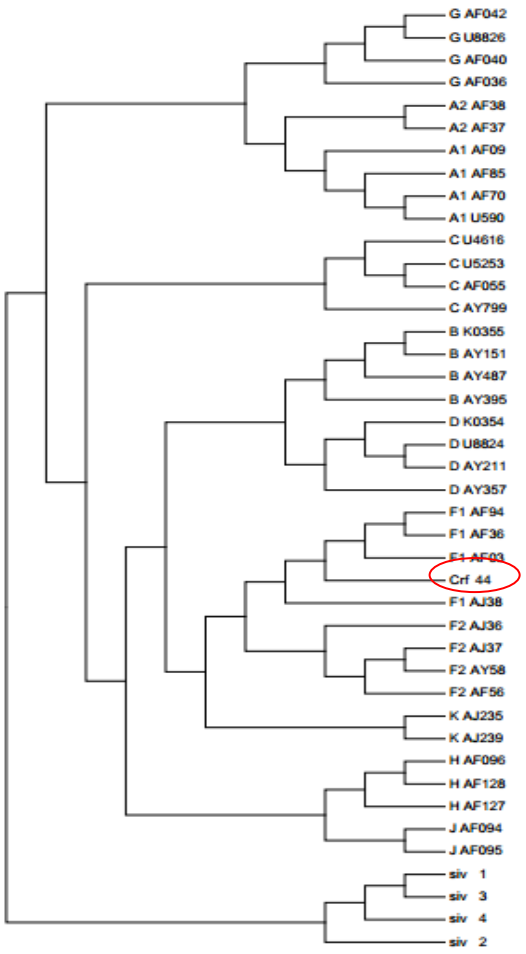


Fig. 3.54 : CRF44_BF

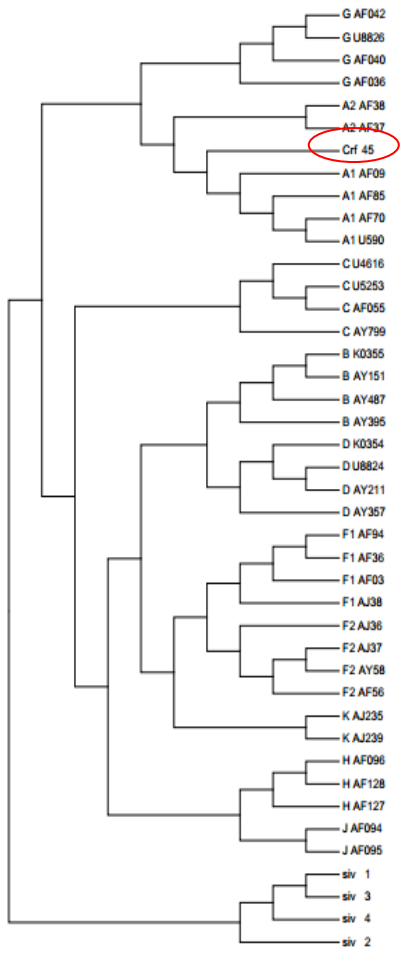


Fig. 3.55 : CRF45_cpx

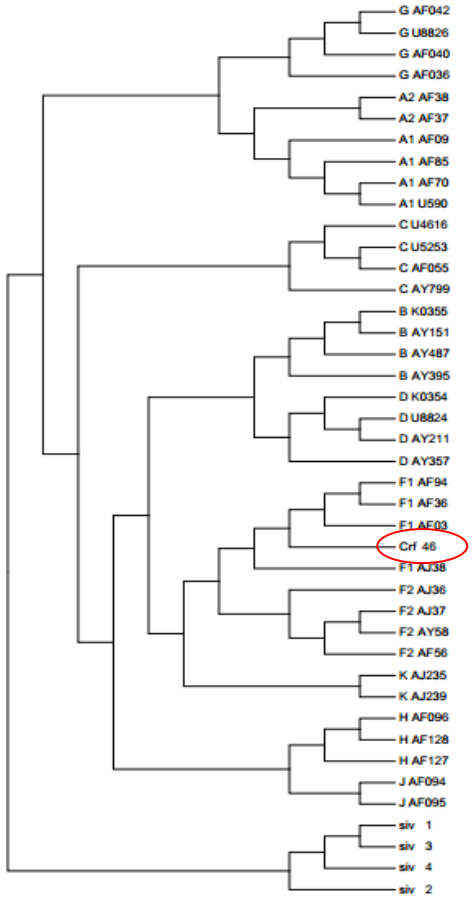


Fig. 3.56 : CRF46_BF

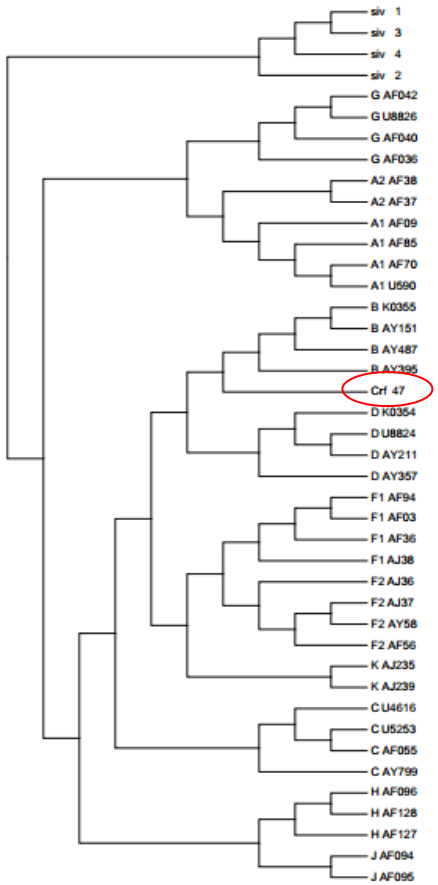


Fig. 3.57 : CRF47_BF

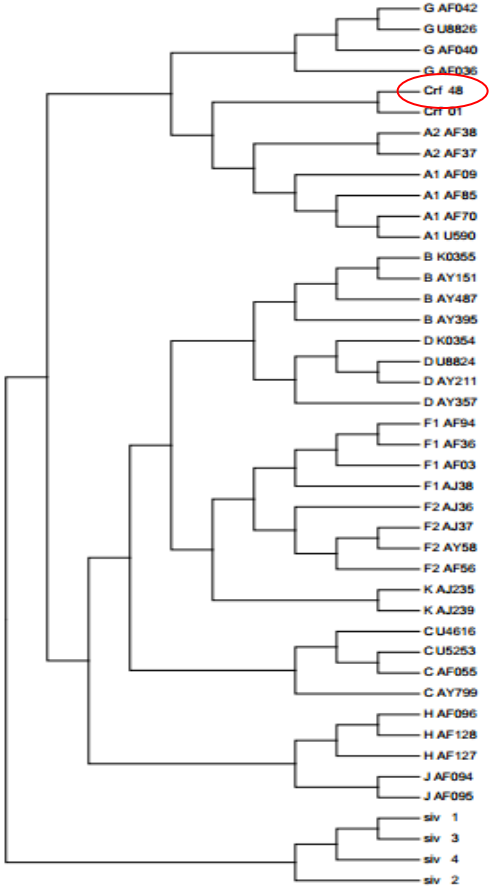


Fig. 3.58 : CRF48_01B

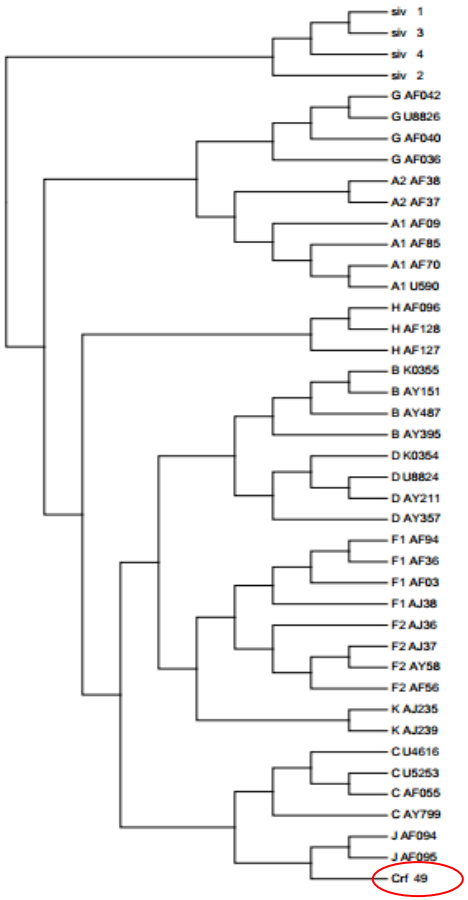


Fig 3.59 : CRF49_cpx

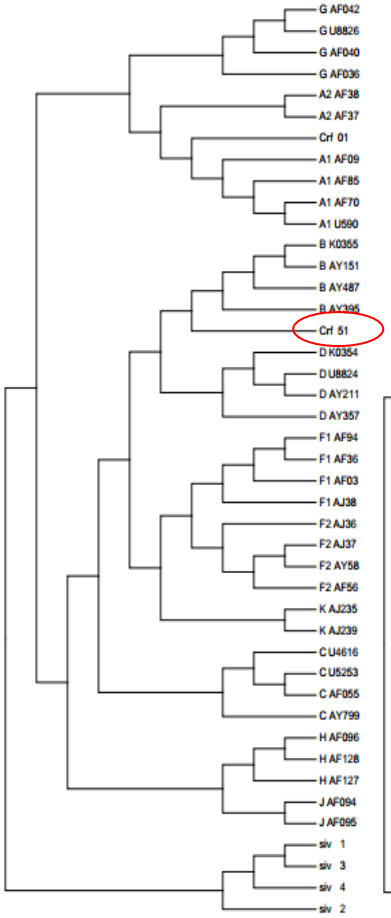


Fig. 3.60 : CRF51_01B

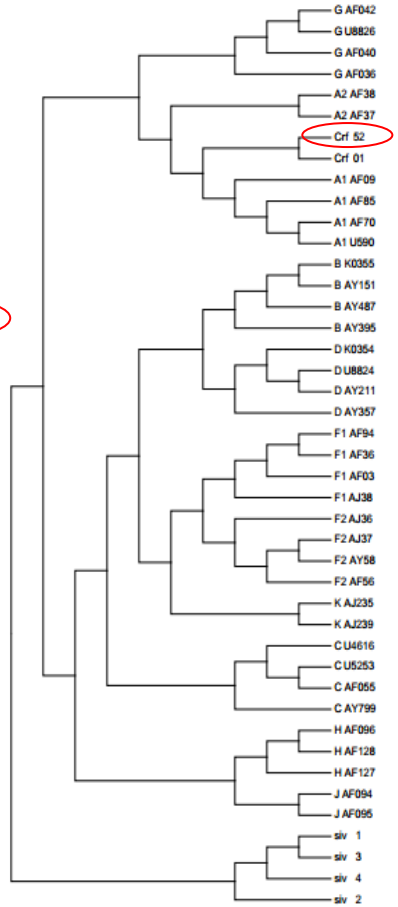


Fig. 3.61 : CRF52_01B

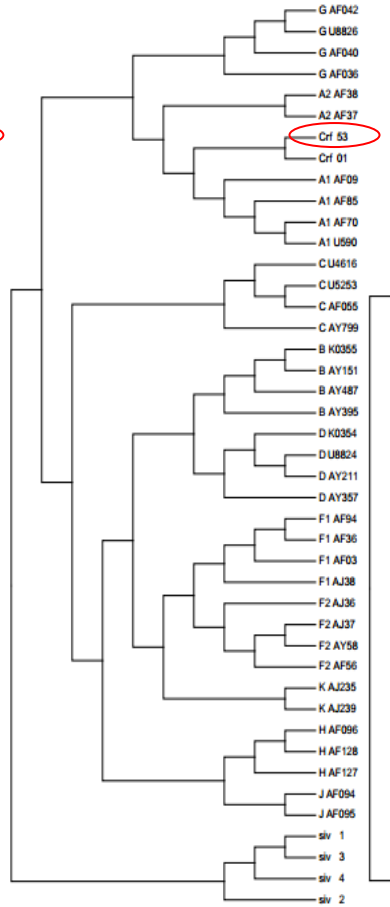


Fig. 3.62 : CRF53_01B

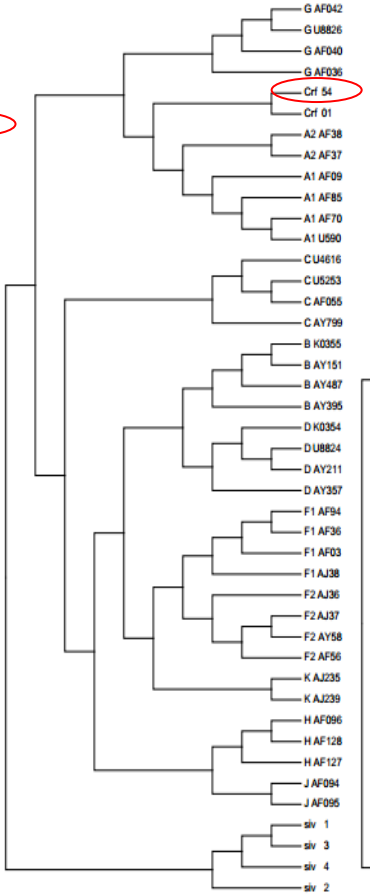


Fig. 3.63 : CRF54_01B

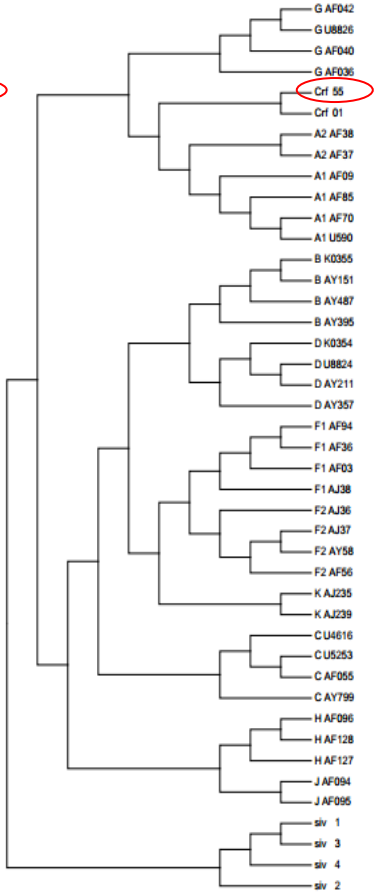


Fig. 3.64 : CRF55_01B

Table 3.2 : Information regarding CRFs and their clustering in cladograms

Name of CRF	Parental subtypes with % of genome length							Parental subtype clustered with CRF	% of clustered parental
CRF46_BF	B (1%)	F1 (99%)						F	99%
CRF31_BC	B (2%)	C (98%)						C	98%
CRF53_01B	CRF01 (97%)	B (3%)						CRF01	97%
CRF33_01B	CRF01 (93%)	B (7%)						CRF01	93%
CRF16_A2D	A2 (91%)	D (9%)						A	91%
CRF26_AU	A (89%)	U (11%)						A	89%
CRF35_AD	A (89%)	D (11%)						A	89%
CRF55_01B	CRF01 (89%)	B (11%)						CRF01	89%
CRF34_01B	CRF01 (88%)	B (12%)						CRF01	88%
CRF44_BF	B (12%)	F1 (88%)						F	88%
CRF48_01B	CRF01 (88%)	B (12%)						CRF01	88%
CRF32_06A1	CRF06 (87%)	A1 (13%)						CRF06	87%
CRF20_BG	B (14%)	G (86%)						G	86%
CRF38_BF	B (14%)	F1 (86%)						F	86%
CRF08_BC	B (15%)	C (85%)						C	85%
CRF28_BF	B (85%)	F1 (15%)						B	85%
CRF24_BG	B (17%)	G (83%)						G	83%
CRF23_BG	B (20%)	G (80%)						G	80%
CRF07_BC	B (22%)	C (78%)						C	78%
CRF14_BG	B (22%)	G (78%)						G	78%
CRF15_01B	CRF01 (77%)	B (23%)						CRF01	77%

CRF17_BF	B (23%)	F1 (77%)						F	77%
CRF12_BF	B (18%)	F1 (72%)						F	72%
CRF01_AE	A (69%)	E/U (31%)						A	69%
CRF21_A2D	A2 (32%)	D (68%)						D	68%
CRF52_01B	CRF01 (68%)	B (32%)						CRF01	68%
CRF10_CD	C (26%)	D (66%)	U (8%)					D	66%
CRF47_BF	B (66%)	F1 (34%)						B	66%
CRF51_01B	CRF01 (34%)	B (66%)						B	66%
CRF03_AB	A (38%)	B (62%)						B	62%
CRF02_AG	A (56%)	G (44%)						A	56%
CRF45_cpx	A (56%)	K (12%)	U (32%)					A	56%
CRF05_DF	D (54%)	F (46%)						D	54%
CRF40_BF	B (54%)	F1 (46%)						B	54%
CRF22_01A1	CRF01 (47%)	A1 (53%)						A	53%
CRF29_BF	B (53%)	F1 (47%)						B	53%
CRF39_BF	B (53%)	F1 (47%)						B	53%
CRF37_cpx	CRF01 (3%)	CRF02 (7%)	A (52%)	G (37%)	U (1%)			A	52%
CRF54_01B	CRF01 (52%)	B (48%)						CRF01	52%
CRF25_cpx	A (19%)	G (50%)	U (31%)					G	50%
CRF49_cpx	A1 (22%)	C (19%)	J (48%)	K (5%)	U (6%)			J	48%
CRF06_cpx	A (20%)	G (46%)	J (24%)	K (10%)				G	46%
CRF19_cpx	A1 (34%)	D (46%)	G (20%)					D	46%
CRF13_cpx	E (14%)	A (22%)	G (39%)	J (22%)	U (3%)			G	39%

CRF43_02G	CRF02 (62%)	G (38%)						G	**38%**
CRF11_cpx	A (50%)	E (3%)	G (4%)	J (33%)	U (10%)			J	**33%**
CRF27_cpx	A (3%)	CRF01 (23%)	G (31%)	H (10%)	J (27%)	K (5%)	U (1%)	G	31%
CRF36_cpx	CRF01 (22%)	CRF02 (38%)	A (26%)	G (14%)				A	**26%**
CRF18_cpx	A1 (37%)	F (4%)	G (21%)	H (17%)	K (4%)	U (17%)		G	**21%**
CRF04_cpx	A (31%)	G (18%)	H (6%)	K (16%)	U (29%)			G	**18%**
CRF09_cpx	A	G	U	(Break-points not available on database)				A	NA
CRF30_0206	CRF02	CRF06	(Break-points not available on database)						NA
CRF41_CD	C (18%)	D (72%)	(Whole genome sequence not available)						NA
CRF42_BF	B	F1	(Break-points not available on database)						NA
CRF50_A1D	A1	D	(Break-points not available on database)						NA

Note : CRFs % in grey (CRF_4, CRF_11,CRF_18,CRF_36 and CRF_43) are CRFs in which CRF did not cluster with parental subtype having maximum recombinant length. All these CRFs are complex CRFs. All bi-recombinant CRFs clustered with parental subtype having maximum recombination length as seen in Table 3.2. Even for genome length as low as 31% , clustering was correct. It is because 31% is relatively higher genome length than rest of subtypes in that CRF.

3.4 Discussion –

It is clear from section 3.3.1 that the minimum word-length k=6 is required for classifying CRFs. While classifying CRFs it was seen that all bi-recombinants clustered properly. However, five CRFs - CRF04, CRF11, CRF18, CRF36, CRF43 – did not cluster with parental subtype having maximum recombination length.

Note that all 5 of them are complex CRFs created by recombination from earlier CRFs. So, 5 out of 13 complex CRFs were not clustered properly (61.5% of the complex CRF recombinants clustered properly). Thus in case of complex CRFs, efficiency of this method is low. But for all normal bi-recombinants, CRFs always clustered with parental subtype having maximum recombination length (100% clustered properly).

It is clearly evident that this method worked for all bi-recombinants and most other CRFs. But complex CRFs made up of recombination of CRFs require further investigation.

Lastly as this method worked for all bi-recombinants, hence we can predict parental subtype with larger genome length in any unknown bi-recombinant CRF.

Chapter – 4

Selected words analysis

4.1 Introduction –

CRFs were analysed using CGR and cladograms in previous chapter at word-length k=6. Word-length k=6 means that all 6 letter words that can be formed using A,T,G,C. Thus, there will be 4^6 (i.e. 4096) different 6-letter words. Word-frequencies of all these 6 letter words were taken into consideration while calculating Euclidean distances between genomes. To look into these 6-letter words, comparison between words from CRF and its parental subtype were made.

Table 4.1 : Word-comparison between CRF03_AB and parental subtypes A and B

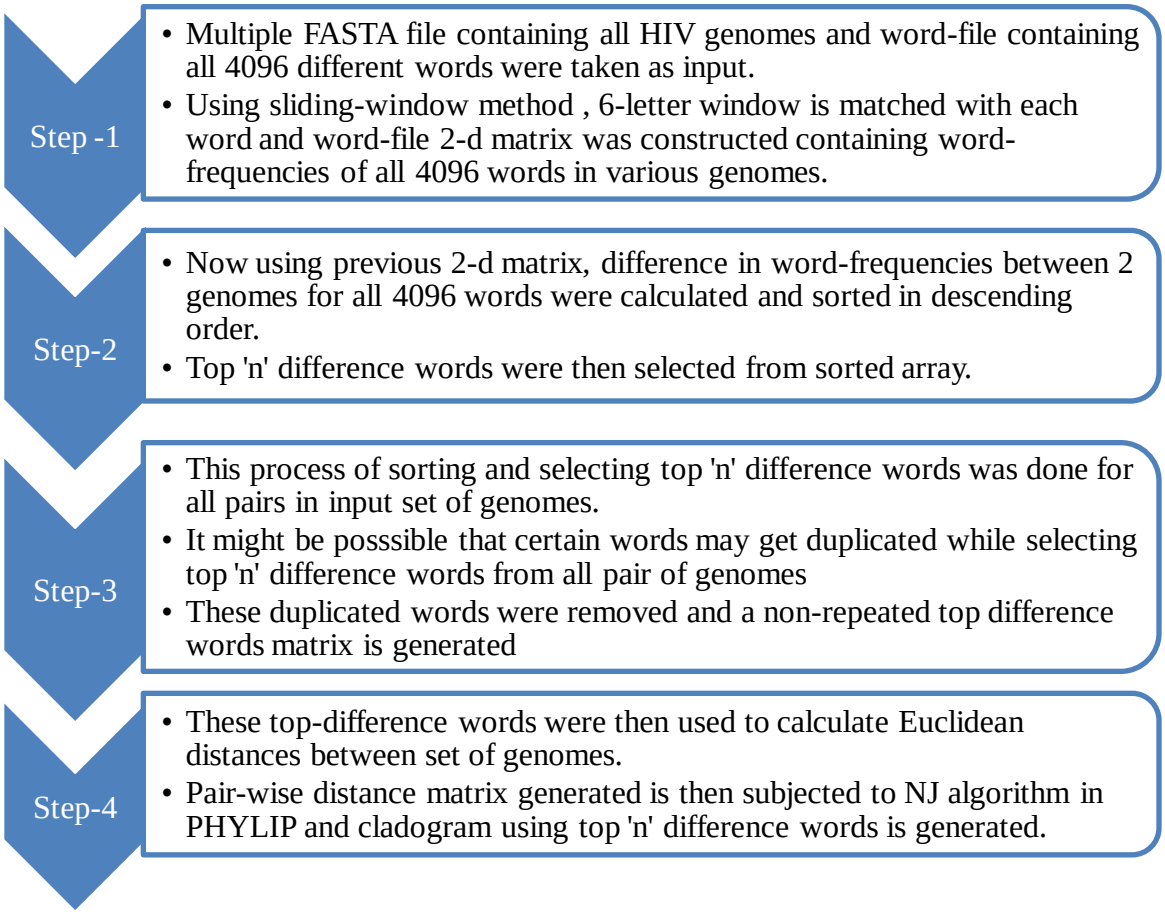
Total 6 letter words	4096 words
Words having same freq. in A and CRF	1624 words
Words having same freq. in B and CRF	1741 words
Words having same freq. in CRF, A and B	972 words
Case of (0,0,0) among same freq. words	718/972 ≈ 74%

So, it is seen that roughly 700 words out of 4096 words were non-existent in HIV-1 genome and did not contribute in generating distances between genomes. Same could also be analysed by looking empty spaces in CGR for CRF03. These empty spaces denote position words having zero frequency in CRF03 (Fig. 3.9)

But clustering in cladogram is based on Euclidean distances between genomes. Thus words which create large frequency difference contributes more towards distance

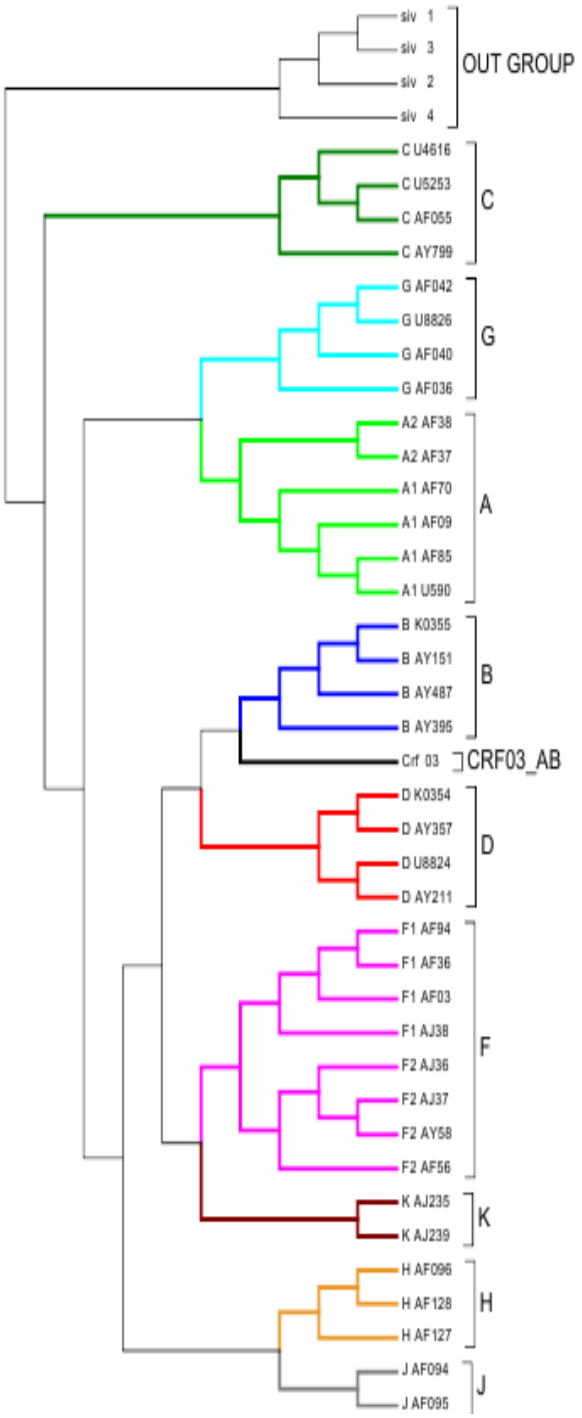
between genomes and so will provide better information required for correct clustering pattern. Hence , It was thought that instead of calculating all 4096 words for cladograms, lesser number of selected words should be used that creates maximum Euclidean distances. These words were termed as ***top difference*** words. Algorithm to select these top difference words is mentioned in next section.

4.2 Algorithm for selected words –



4.3 CRF analysis with reduced word set –

Different top words like top 50, top 10 , top 5 and so on were selected and cladograms were constructed for each case. It is seen that for higher top words like top 50, top 10 CRFs clustered properly but for very less top words like top 2 CRFs didn’t clustered properly and certain subtypes got inter-mixed. (Fig. 4.1, 4.2 , 4.3 , 4.4)



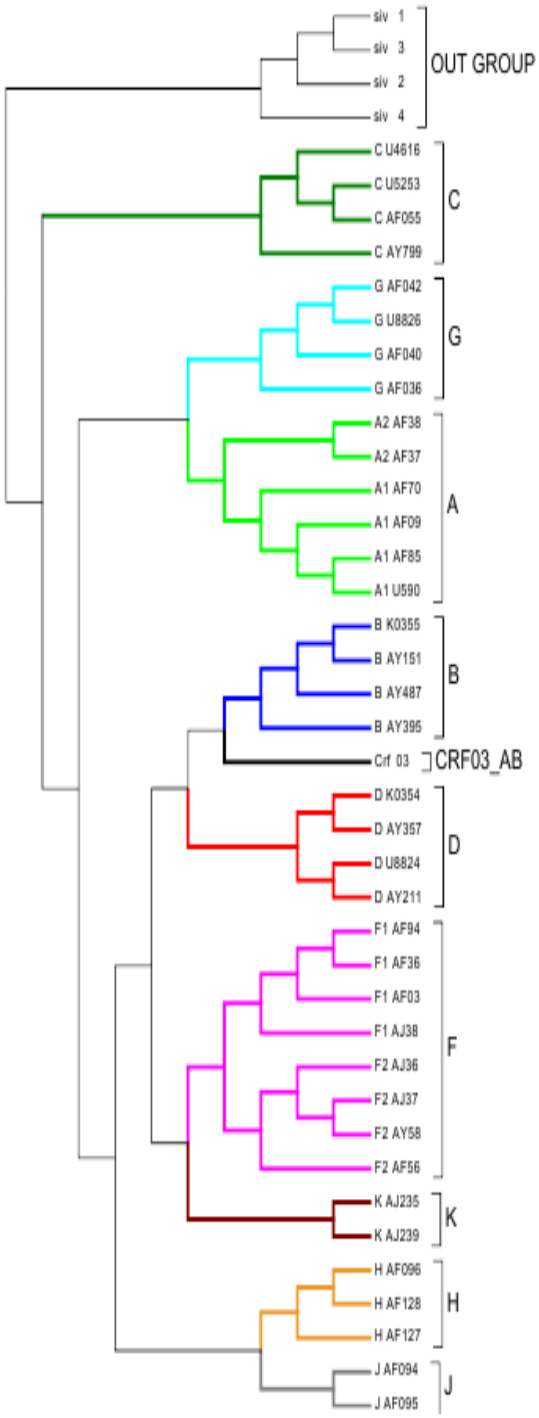


Fig. 4.1 : CRF03_AB using all 4096 words (A = 38% , B =62% of genome length 9719 bp)

Fig. 4.2 : CRF03_AB using top 15 words (725 words)

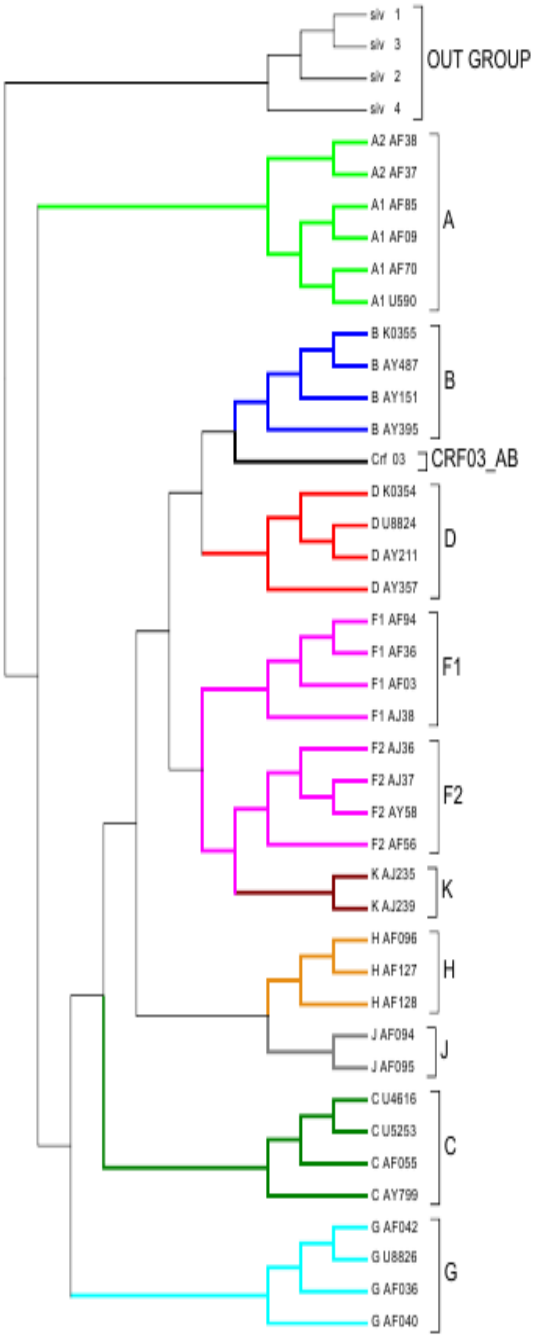


Fig. 4.3 : CRF03_AB using top 5 words (400 words)

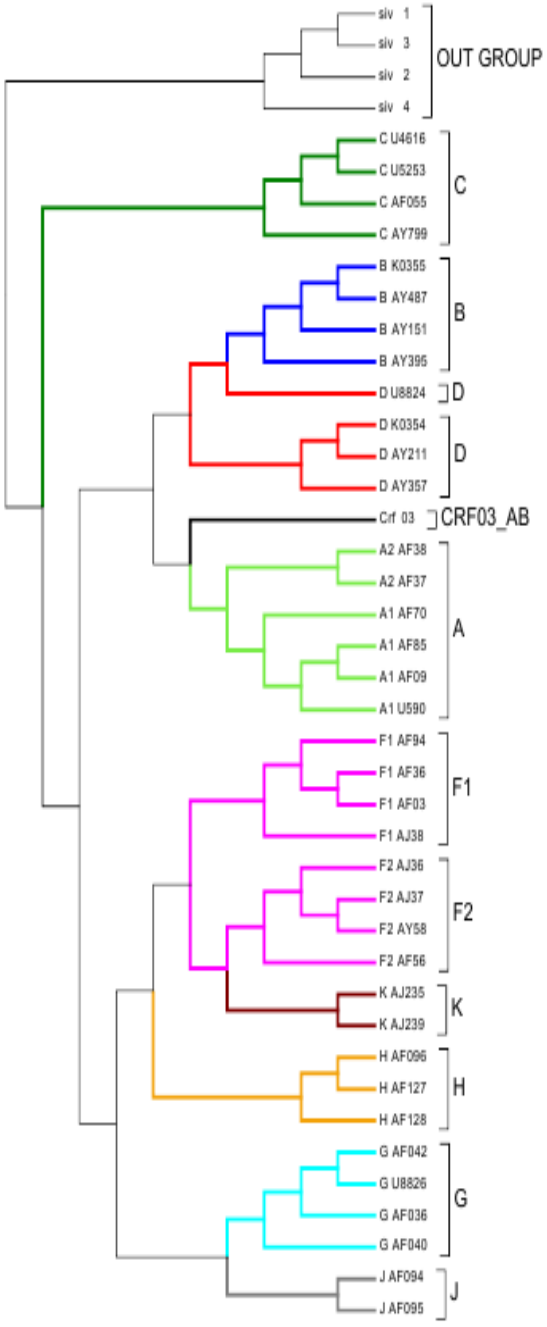


Fig. 4.4 : CRF03_AB using top 2 words (220 words)

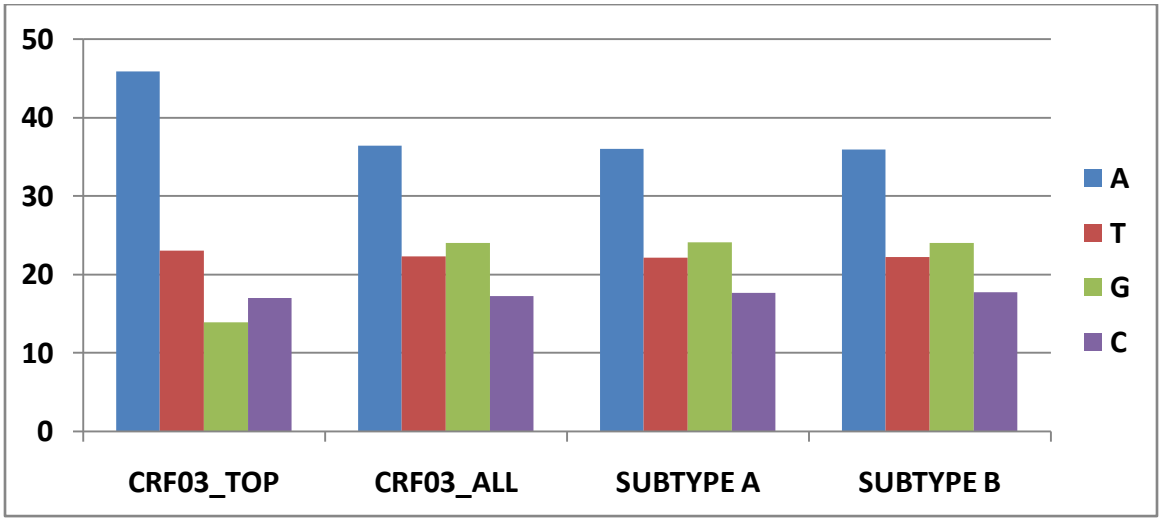
Thus , it is seen that at lower top difference words, total number of words used to construct cladograms were reduced to very low values insufficient for proper clustering ,

but when a sufficient number of words were selected then correct clustering pattern was obtained. So, top 10 difference words were selected as they gave correct clustering pattern for various CRFs. These top 10 difference words amounts to roughly 600 words.

Although lesser words (nearly 600 words) were used instead of 4096 words, but computational time nearly remained same as time was spent in algorithm for searching top difference words. However, using this method we selected out words that were contributing maximum to distances between genome. These words were later analysed by base composition analysis to see if these words are some specific words like rich or deficient in particular nucleotides.

4.4 Base composition analysis of selected words –

Base composition analysis was performed for CRF (using top 10 words) , CRF (using whole genome) and parental subtypes of CRFs. It was seen that in case of top words , these words were more A – rich than whole genome and parental subtypes. (Fig. 4.5)



(Fig. 4.5 : Base composition analysis of top 10 difference words in CRF03_AB
Y-axis shows percentage of various subtypes)

Table 4.2 : Percentages of base composition analysis of CRF03_AB (Fig. 4.6)

	CRF03_TOP	CRF03_ALL	SUBTYPE A	SUBTYPE B
A	45.81	36.42	36.03	35.93
T	23.02	22.32	22.18	22.22
G	13.9	23.99	24.13	24.04
C	17	17.23	17.64	17.79

Although HIV genome is itself A-rich but when compared with CRF03_AB whole genome it is seen that percentage of A in top words is 45.81% while that in whole genome is 36.42% . Hence, top difference words are more A–rich and we are not getting this result simply because of HIV genome itself being A-rich genome. It is also worth mentioning that there is a reduction in G in top difference words. (Table 4.2)

Now to see if these A were present at any particular position in 6-letter words, we performed positional base composition analysis of 6-letter top difference words.

Table 4.3 : Positional base composition analysis for various CRFs (Value shown below is % of nucleotide at particular position)

CRF03_AB	1	2	3	4	5	6
A	52.17	45.65	46.32	43.14	43.81	45.48
T	13.88	17.56	16.05	17.89	18.73	20.74
G	19.73	24.08	22.74	24.92	24.08	22.74
C	14.72	13.21	15.38	14.55	13.88	11.54

CRF28_BF	1	2	3	4	5	6
A	51.93	45.56	46.40	43.05	43.55	44.56
T	13.40	17.25	15.91	17.76	18.26	20.44
G	19.77	23.45	22.45	24.96	24.29	23.28
C	14.91	13.74	15.24	14.24	13.90	11.73

CRF14_BG	1	2	3	4	5	6
A	51.91	45.42	46.09	42.93	43.59	45.26
T	13.81	17.47	15.97	17.80	18.64	20.63
G	19.63	23.96	22.63	24.79	23.96	22.63
C	14.64	13.14	15.30	14.47	13.81	11.48

CRF31_BC	1	2	3	4	5	6
A	51.85	45.45	46.30	42.93	43.60	44.78
T	13.47	17.17	15.82	17.17	18.52	20.37
G	20.03	24.07	23.06	25.76	24.24	23.23
C	14.65	13.30	14.81	14.14	13.64	11.62

It is clear from Table 4.3 that 1st position has more than 50% words starting with A, Also lot of words (45%) ends with A, indicating most of these lie in A quadrant.

4.5 Discussion –

Selected word analysis of CRFs tells us that even without looking at all possible 4096 words at k=6, if we extracted lesser number of top difference words from birecombinnats, then distance generated by these words is sufficient enough to cluster CRFs and subtypes properly. Also using these top words , we analysed that these words were more A -rich than words in HIV genome. We also infer that these A rich words are most dominant in 1^{st} position in a 6-letter word. More than 50% probability is that 1^{st} position is A in a 6-letter top difference word. (Table 4.3)

Chapter – 5

Gene-based clustering

5.1 Introduction –

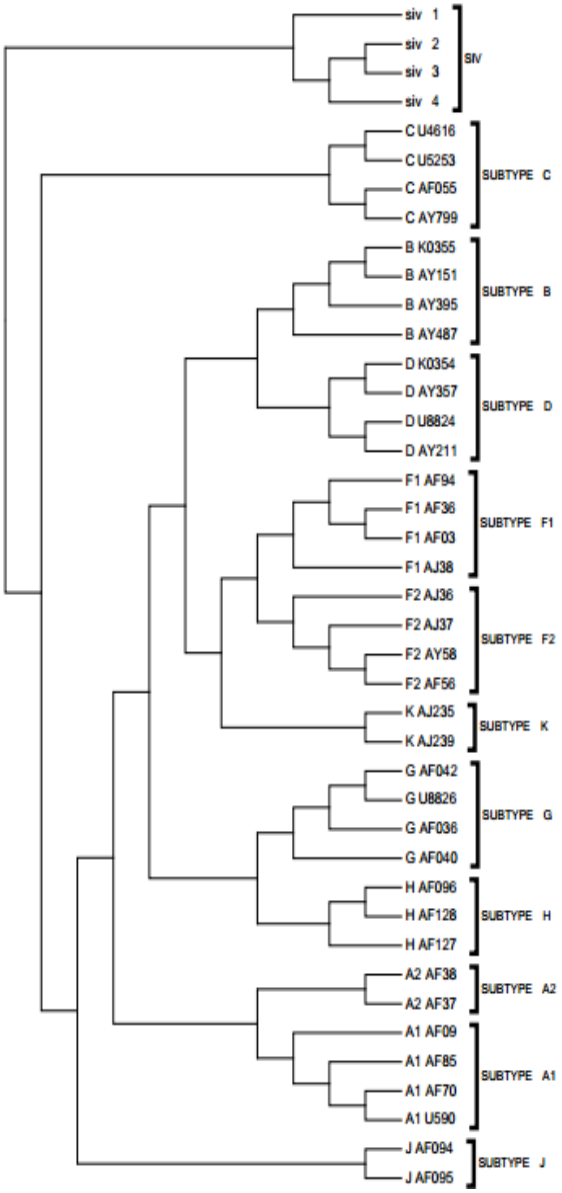
In previous chapters, cladograms were created using whole genome sequences but in this chapter we tried to use particular HIV genes to create cladograms and see if clustering is proper. HIV whole genome is around 9700bp and has 9 genes. They are *gag, pol, env, vif, vpr, tat, nef, vpu, ref* and *tat*. 5 genes out of these 9 namely *gag*, *pol*, *env*, *vif*, *vpr* were used for this study.

Genes were extracted from HIV whole genome using GeneCutter available on HIV Los Alamos database .

5.2 Clustering based on genes –

Using these extracted genes from a set of genomes, cladograms were created. It was observed that cladograms created using larger genes like gag (~1500bp) , pol (~2800bp) and env (~2500bp) showed correct classification (Fig. 5.1) whereas for smaller genes like vif (~500bp) and vpr (~400bp) clustering resulted in inter-mixing of certain subtypes. (Fig. 5.2).

The possible reason for this error is that the small size of DNA yielded fewer number of 6 letter words, which was not enough for arriving at a large sample of points for comparison.



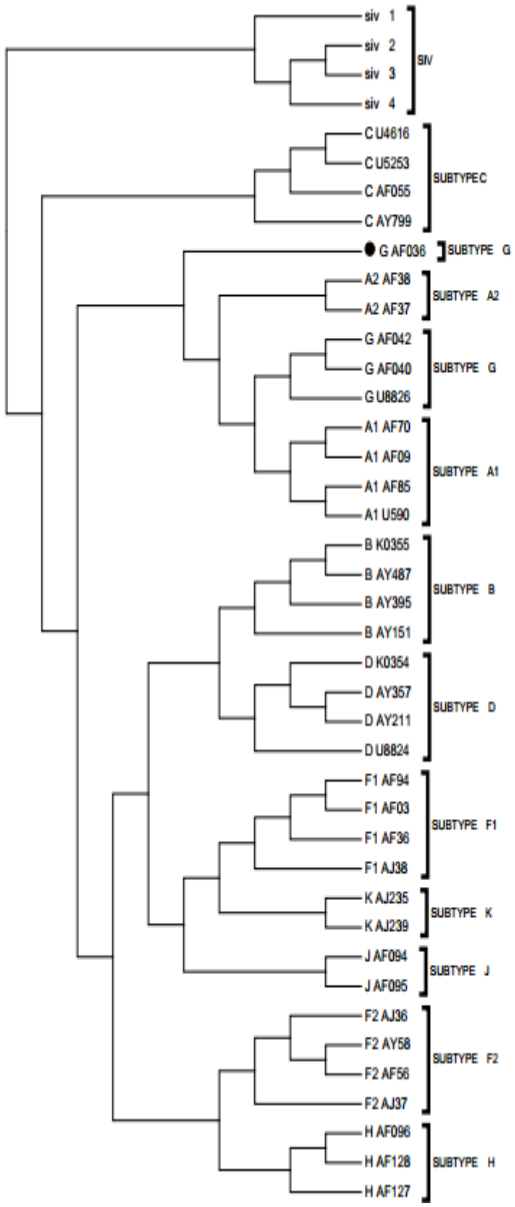


Fig. 5.1 : Cladogram at k=6 for pol genes (subtypes were properly clustered)

Fig. 5.2 : Cladogram at k=6 for vif genes (Subtypes A1,A2,G,F1,F2 got inter-mixed)

Chapter – 6

Software tool developed

6.1 Introduction

A simple software tool which can read HIV sequences, generates CGR and calculate base composition for input sequences was created. This tool was developed using PHP and HTML . PHP is used mainly for coding as it can be easily integrated into HTML codes and is executed along with HTML code when a particular HTML page is opened.

This software tool can be used as stand-alone application on a system or can be used as a web-server. However, right now it is used only as a stand-alone application and development of web-server still require further modification and development as future work. Image below shows snapshot of software (Fig. 6.1)

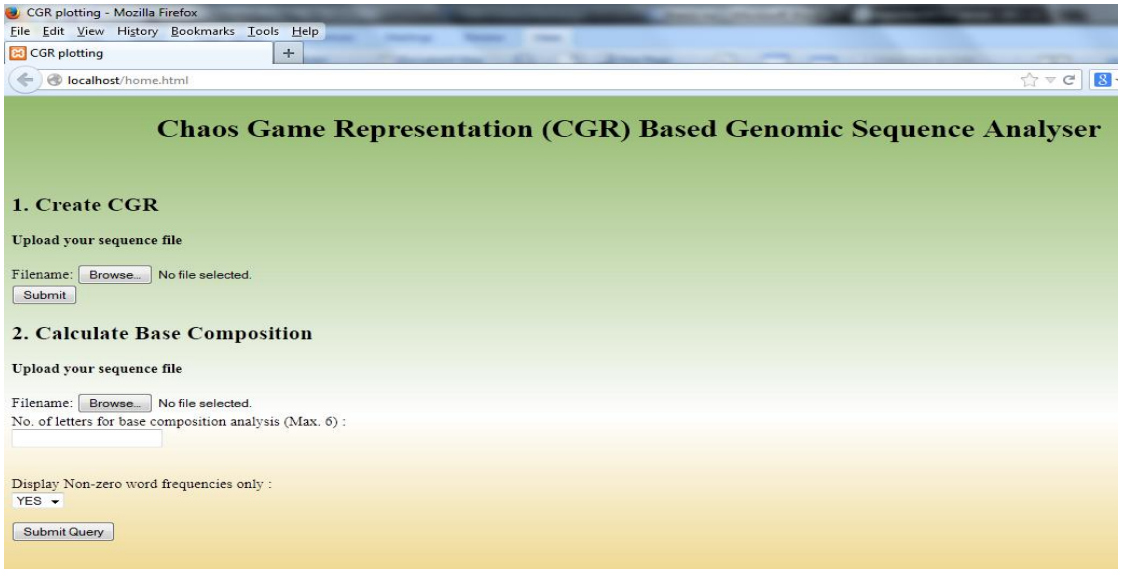
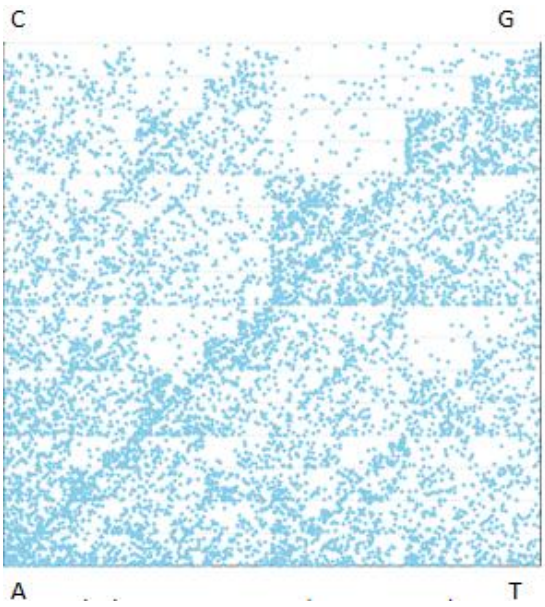


Fig. 6.1 : Front-end of software page developed

First part of software is to create CGR for given input file. Below are few images of CGR generated using software tool.



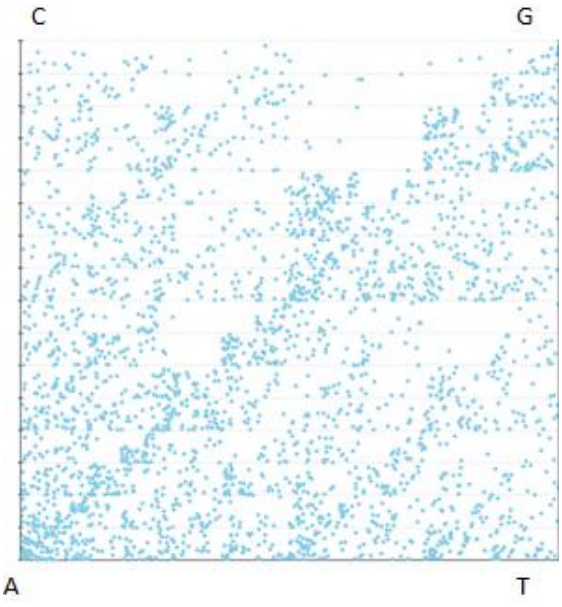


Fig. 6.2 : CGR of sequence U51190 (9719bp) Fig. 6.3 : CGR of U51190 (*pol* gene only)

Second part of software is to calculate base composition for given input file. Below is snapshot of result generated for same using this software tool. (Fig. 6.4)

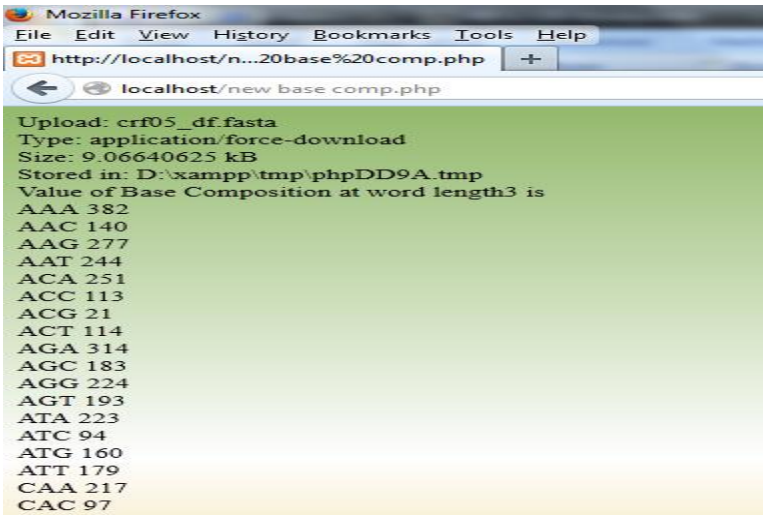


Fig. 6.4 : Snapshot of base composition calculated for CRF05 using software tool

6.2 Future work–

This tool works only as a stand-alone application using Apache localhost. It will be improved in future by adding functionality of creating distance matrix and cladograms. Also if above mentioned functionality is achieved, then it can be made into a web-server.

The question that remains to be addressed is what could be the biological basis for the classification based on higher order words. It will also be useful to see if these top difference words occupy any specific position in the genome and/or contribute to any structural or functional regions in the genome. All these require further study with multiple genomes.

Bibliography

Almeida J, Carriço JA, Maretzek A, Noble PA and Fletcher M : "**Analysis of genomic sequences by chaos game representation**". Bioinformatics (2001) ; 17:429–437

Barnsley M. , "**Fractals Everywhere"** ; Springer-Verlag, New York, (1988)

Carr JK, Foley B, Leitner T, Salminen M, Korber B, McCutchan F : "**Reference sequences representing the principal genetic diversity of HIV-1 in the pandemic**".Human Retroviruses and AIDS (1998) ; p. III-10–19

Felsentein J : "**PHYLIP –Phylogeny Inference Package**. Cladistics (1989) ; 164-166

HIV Los Alamos Database (http://www.hiv.lanl.gov)

Jeffrey HJ: "**Chaos game representation of gene structure**". Nucleic Acids Res ; 1990

Lau K, Wong J : "**Current trends of HIV recombination worldwide**". Infectious Disease Reports (2013) ; 5:s1e4 doi:10.4081/idr.2013.s1.e4

Leitner T, Korber B, Daniels M, Calef C, Foley B : "**HIV-1 subtype and circulating recombinant form (CRF) reference sequences**". HIV sequence compendium ; Los Alamos National Laboratory (2005), 41–48

Saitou N, Nei M : "**The neighbor-joining method: A new method for reconstructing phylogenetic trees**" Molecular Biology and Evolution (1987) 4:406-425.

Pandit A, Sinha S : "**Using genomic signatures for HIV-1 sub-typing**". BMC Bioinformatics (2010) ; 11(Suppl 1):S26 doi: 10.1186/1471-2105-11-S1-S26

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, and Kumar S : "**MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods**". Molecular Biology and Evolution (2011)