POPULATION BIODIVERSITY STUDY OF RUBELLA VIRUS USING STATISTICAL APPROACH

Keerti S. Athalye (1803)
Mrudula G. Belgal (1805)
Anjali R. Nehul (1838)

OBJECTIVES

- To study the extent of diversification in the population of Rubella virus.
- To identify the optimum number of clusters present in the virus population.
- To explore the possibility of emergence of new sub- clusters.
- To check the presence of admixture, if any.

About RUBELLA VIRUS

- Realm: *Riboviria*
- Kingdom: *Orthornavirae*
- Phylum: *Kitrinoviticota*
- Class: *Alsuviricetes*
- Order: *Hepelivirales*
- Family: ***Matonaviridae***
- Genus: ***Rubivirus***
- Species: ***Rubella virus***

Some KEY TERMINOLOGIES

- **Genetic diversity** : It refers to the differences in the **genetic make-up of a distinct species and to the genetic variations within a single species.**
- **Genome** : The complete set of genes or genetic material present in a cell or organism.
- **Genotype** : It is the collection of genes responsible for the various genetic traits of a given organism. It simply means what alleles are carried in a particular organism's DNA. e.g. letter Bb (B-dominant genotype and b-recessive genotype).
- **Phylogenetic Tree** : A branching diagram showing the inferred evolutionary relationships among various biological species or other entities based upon similarities and differences in their genotypic or phenotypic characteristics.

- **Allele** : It is one of the alternative forms of the same gene or same genetic locus (dominant and recessive)
- **Allele Frequency** : It is the relative frequency of an allele at a particular locus in a population, expressed as a fraction or percentage. They are used to describe the amount of variation at a particular locus or across multiple loci.
- **Linkage** : It is the tendency of genes or other DNA sequences at specific loci to be inherited together as a consequence of their inherited proximity on a single chromosome.
- **Linkage equilibrium** : This means Statistical independence of alleles at all loci.
- **Hardy-Weinberg principle** : The principle states that in a large breeding population, allele frequencies will remain the same from generation to generation assuming that there is no mutation, gene migration, selection or genetic drift and the population is said to be at Hardy–Weinberg equilibrium.

MOTIVATION

- Rubella virus causes German measles or three-day measles. This disease is **often mild, self-limiting and has symptom based treatments**, yet, it has **proven to be deadly for pregnant women and newborn babies**.
- Based on the 2018 Global Vaccine Action Plan (GVAP) Assessment Report by the WHO Strategic Advisory Group of Experts (SAGE) on Immunization, **two regions (African and Eastern Mediterranean) have not yet set rubella elimination or control target**.
- According to the Centers for Disease Control and Prevention, **effectiveness of one dose of the available vaccine is about 97%, which means that 3% of the population is still at risk**.
- **We hope that this study turns out to be useful for the people and communities working for the same.**

LITERATURE REVIEW

- **Analysis of genetic structure and relationship among nine indigenous Chinese chicken populations by the Structure program, Chen et al (2009).** The paper explains how the Structure was used to infer the genetic structure of nine indigenous Chinese chicken populations based on 16 microsatellite markers.
- **Inference of Population Structure Using Multilocus Genotype Data, Pritchard et al (2000).** A model based clustering method is described for multilocus genotype data to infer population structure and assign individuals to populations that are characterized by a set of allele frequencies at each locus.
- **Introduction To Adegenet 2.0.0, Thibaut Jombart.** The paper provides various illustrations for the users of Adegent. It consists of instructions to install the package in R and also includes R algorithms to carry out import/export of data, data conversion, object creation, Multivariate analysis (PCA), spatial analysis and simulation with examples.

DATA

- **Secondary data**.
- **Complete genomes of 52 wild Rubella viruses** downloaded from NCBI Nucleotide database, curated by the Department of Bioinformatics, SPPU.
- **Maximum Length** of a sequence was **9777**.
- **Two lineages 1 and 2**.
- **Eleven sub-lineages** with strains are **1A:5; 1B:3; 1C:3; 1D:3; 1E:11; 1F:2; 1G:2; 1I:1, 1J:2; 2B:18; 2C:2**.
- Out of these, 1I was not considered for the analysis.
- Remaining groups named as 1, 2, 3, 4, 5, 6, 7, 8, 10, 11.

SOFTWARE and APPLICATIONS

- **Mega (X)**: Mega is a software suite developed to analyse DNA and protein sequences from species and populations.
- **LIAN (3.7)**: The hypothesis of Linkage equilibrium is tested using Monte Carlo method in LIAN.
- **R software (3.6.2)**: It is used to carry out wide range of computations with the help of different packages.
- **STRUCTURE (2.3.4)**: It implements a model based clustering method which is used to infer the population structure using genotype data.
- **Structure Harvester**: It is used to detect optimum number of clusters (K) of individuals using Evanno method.
- **iTOL**: This application provides means to explore Phylogenetic trees.

MSA and ‘PI SITES’ EXTRACTION

- **MULTIPLE SEQUENCE ALIGNMENT (MSA)** is the alignment of three or more biological sequences (protein or nucleic acid) of similar length.
  We used MEGA for MSA.
  After **MSA and deletion of gaps in the sequences, length of each sequences was 9716**.

- **PARSIMONY INFORMATIVE SITES** are the sites that contain at least two types of nucleotide bases and at least two of them occur with minimum frequency of two.
  From MSA, these sites were extracted using MEGA and used as an input for STRUCTURE software.
  Number of PI sites was 1932.

1. AF435865.	t	t	a	t	g	a	a	a	a	g	c	g	c	g	a
2. AF435866.	t	t	a	t	g	a	a	a	a	g	c	g	c	g	a
3. JN635281.	t	t	a	t	g	a	a	a	a	g	c	g	c	g	a
4. M15240.2	t	t	a	t	g	a	a	a	g	g	c	g	c	g	a
5. DQ085339	t	c	a	t	g	a	a	a	g	g	c	g	c	g	a
6. JN635282.	t	c	a	t	g	a	a	a	g	g	c	g	c	g	a
7. DQ388281	t	t	a	t	g	a	a	a	g	g	c	g	c	c	a

Loci highlighted by yellow are PI sites

LIAN (LINKAGE ANALYSIS)

- This tests the statistical hypothesis of linkage equilibrium formulated as
  $H_0 : V_D = V_E$ (Null Hypothesis of Linkage Equilibrium)
  where, V_D is Variance obtained from re-sampled data
  V_E is Variance expected to establish linkage equilibrium

- The extent of linkage can be inferred from the following parameter:

$$I^S{}_A = \frac{1}{l-1}\left(\frac{V_D}{V_E} - 1\right)$$

  where, l : the total number of loci.

- For our data, $I_A{}^S$ is **0.0949** which is extremely less deviated from the threshold (0.05), hence, it indicates **weak linkage equilibrium**.

Summary Statistics

V_D	63278.7971
V_e	343.3530
$I_A{}^S$	0.0949

Testing Null Hypothesis (H_0: $V_D = V_e$)

	Monte Carlo (1000 resamplings)
Var(V_D)	1017.1932
P	< 1.00 x 10^{-03}
L	398.1252

Genetic Diversity

Mean genetic diversity (H): 0.2975 +/- 0.0040

ANALYSIS in R

- Converted the PI sites data in **Genind** object, this object stores allelic frequencies along with other information.

- Carried out Multivariate Analysis using following tools

  1. **Principle Component Analysis** : To reduce dimensions of data and retain the genetic diversity as much as possible.

  2. **Multidimensional Scaling** : To reduce the data dimensions on the basis of distance matrix calculated using '**percentage distance**'.

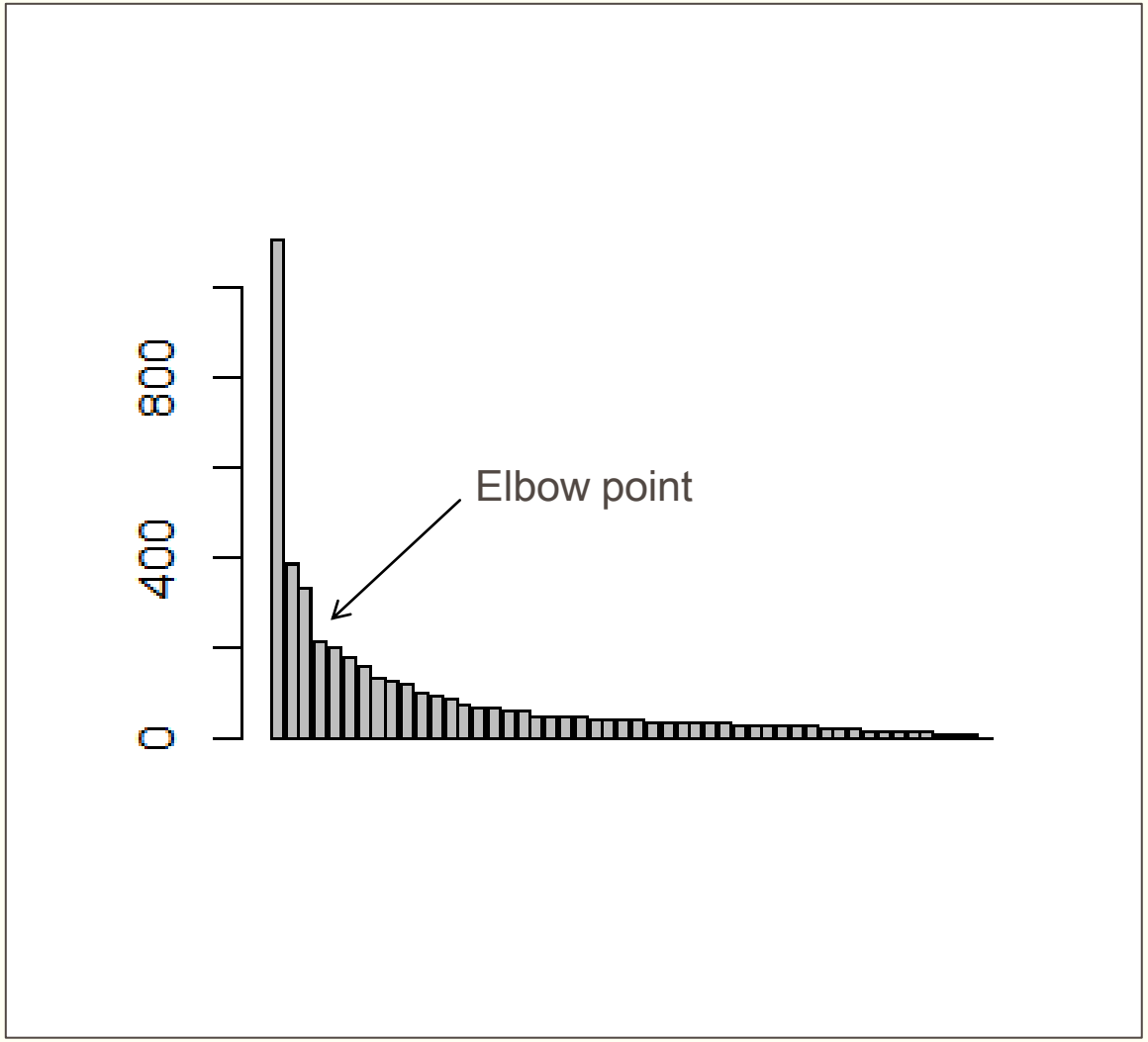
  here, percentage distance = $\frac{\text{no. of loci for which individual differ}}{\text{total no. of loci}}$

  3. **K-Means clustering** : To obtain clusters based on PCA with criterion Within Sum of Squares.

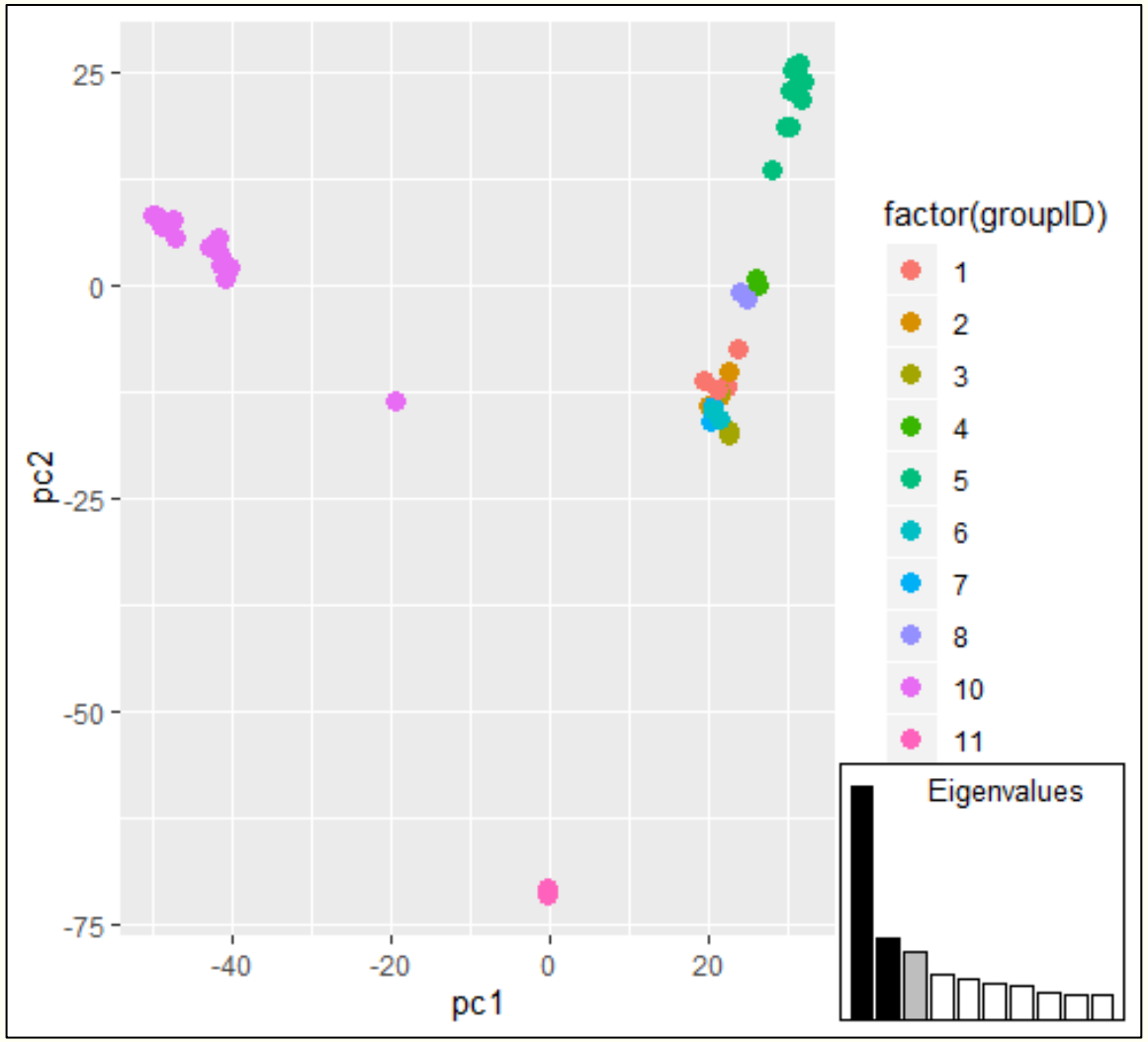
  4. **Discriminant Analysis** : To create functions which would differentiate each strain from the others.

I. PCA

- Considering PI sites as variables and gene sequences as observations, we ran PCA for the encoded data.
- The command used was ‘dudi.pca()’.
- dudi.pca() shows the bar plat of eigen values and asks user the number of PCs to be retained.
- 4 axes were retained in PCA.

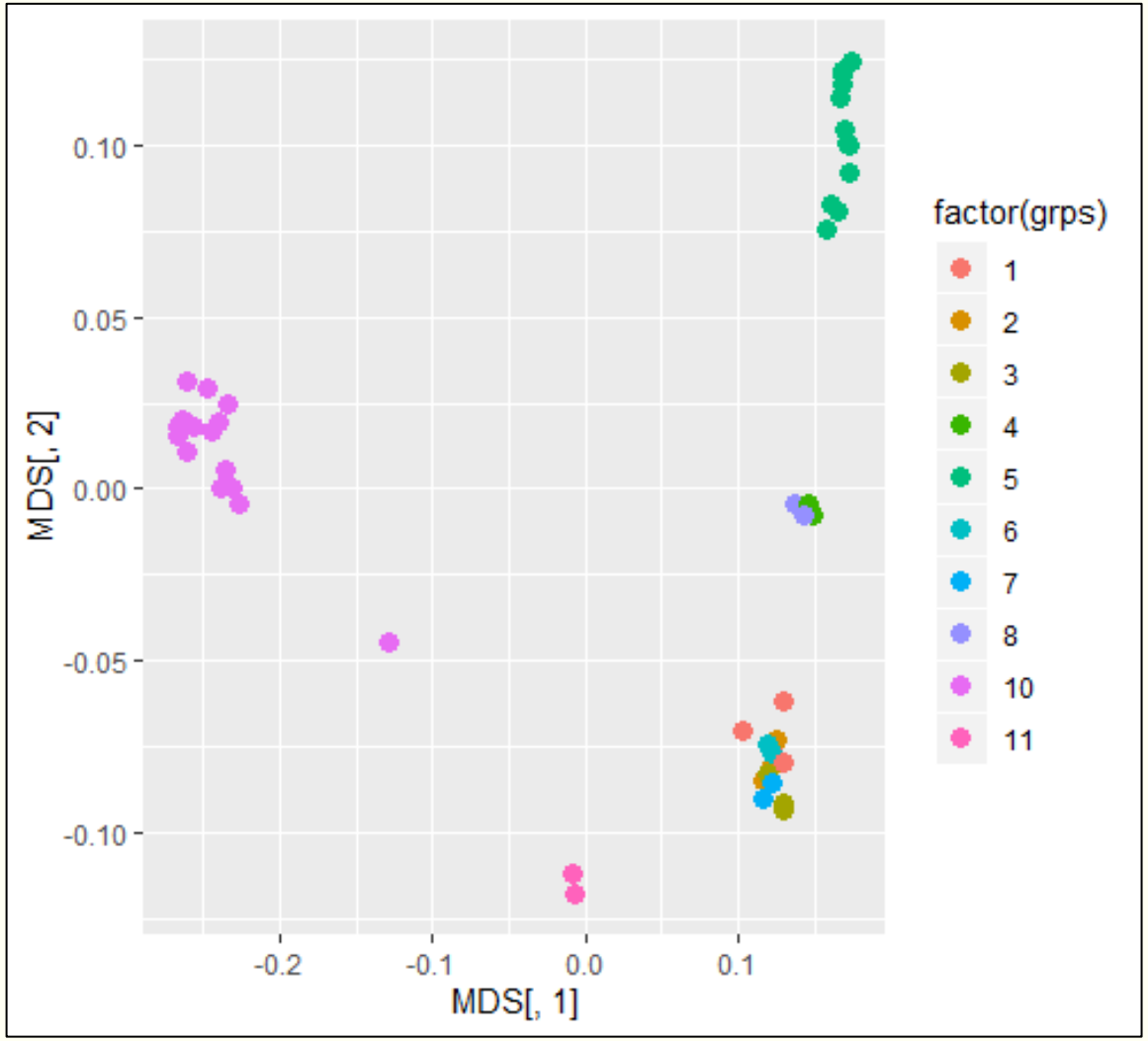


- This graph is PC2 against PC1.
- 5th and 11th group belong to their own separate clusters.
  10th group is divided into 2 clusters. One cluster involves 17 strains out of 18. Second one has single old strain.
  The next cluster is formed by 4th and 8th group.
  Remaining groups numbered 1, 2, 3, 6 and 7 form the last cluster



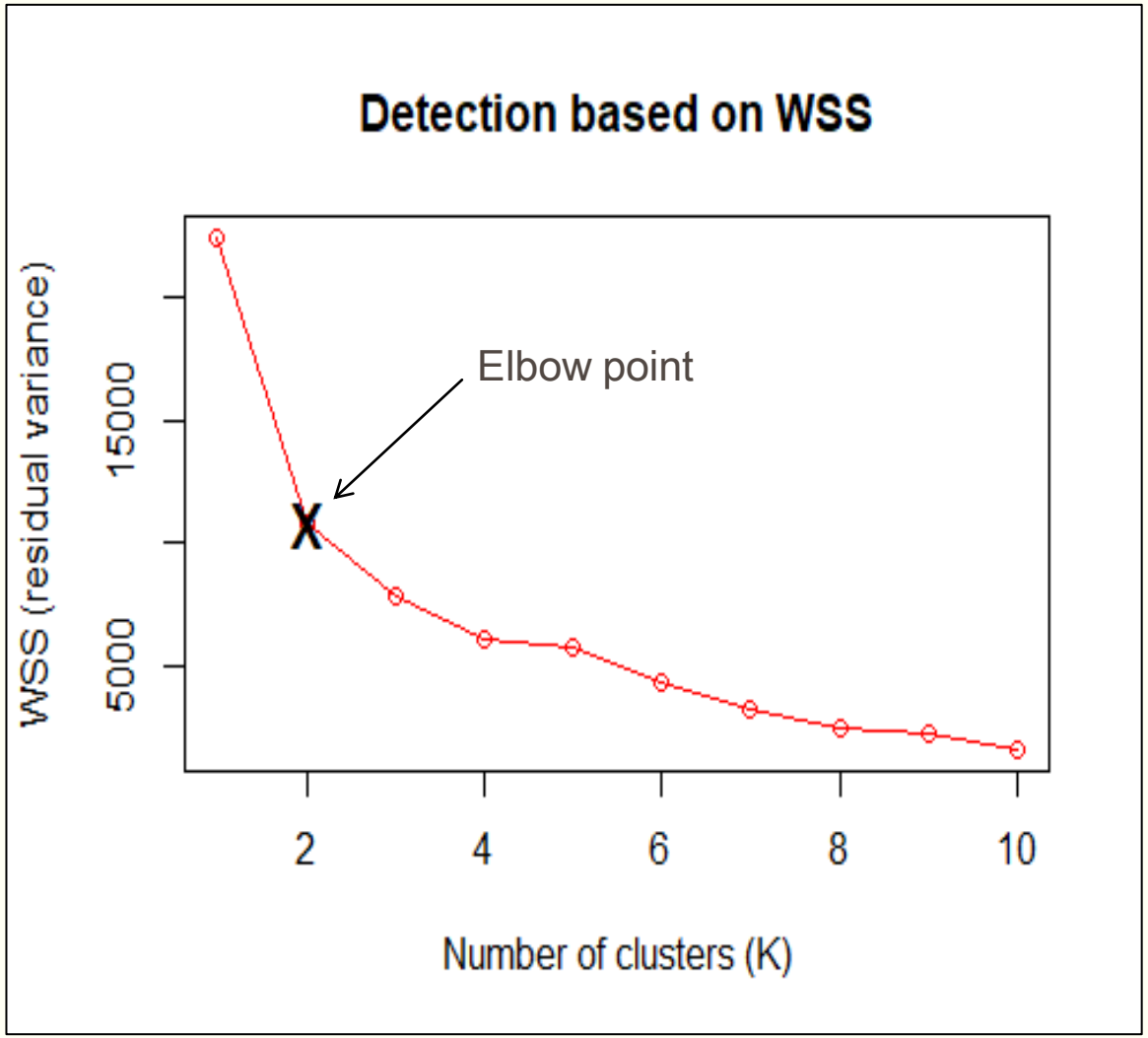
II. MDS

- Genetic distance was calculated using ‘dist.gene’ command using ‘percentage’ method.
- The command used was ‘cmdscale()’.
- 2 axes were retained in MDS.
- This graph is MDS2 against MDS1.
- The graph showed similar results as that of PCA.



III. K-Means Clustering

- ADEgenet provides a function to perform Kmeans clustering on genind object.
- The command used was ‘find.clusters()’. Criterion was Within Sum of Squares (WSS).
- Elbow rule gives opt K to be 2.
- Built-in WSS based criterion gives opt K to be 6.
- Seed was set to 101.



- K = 2

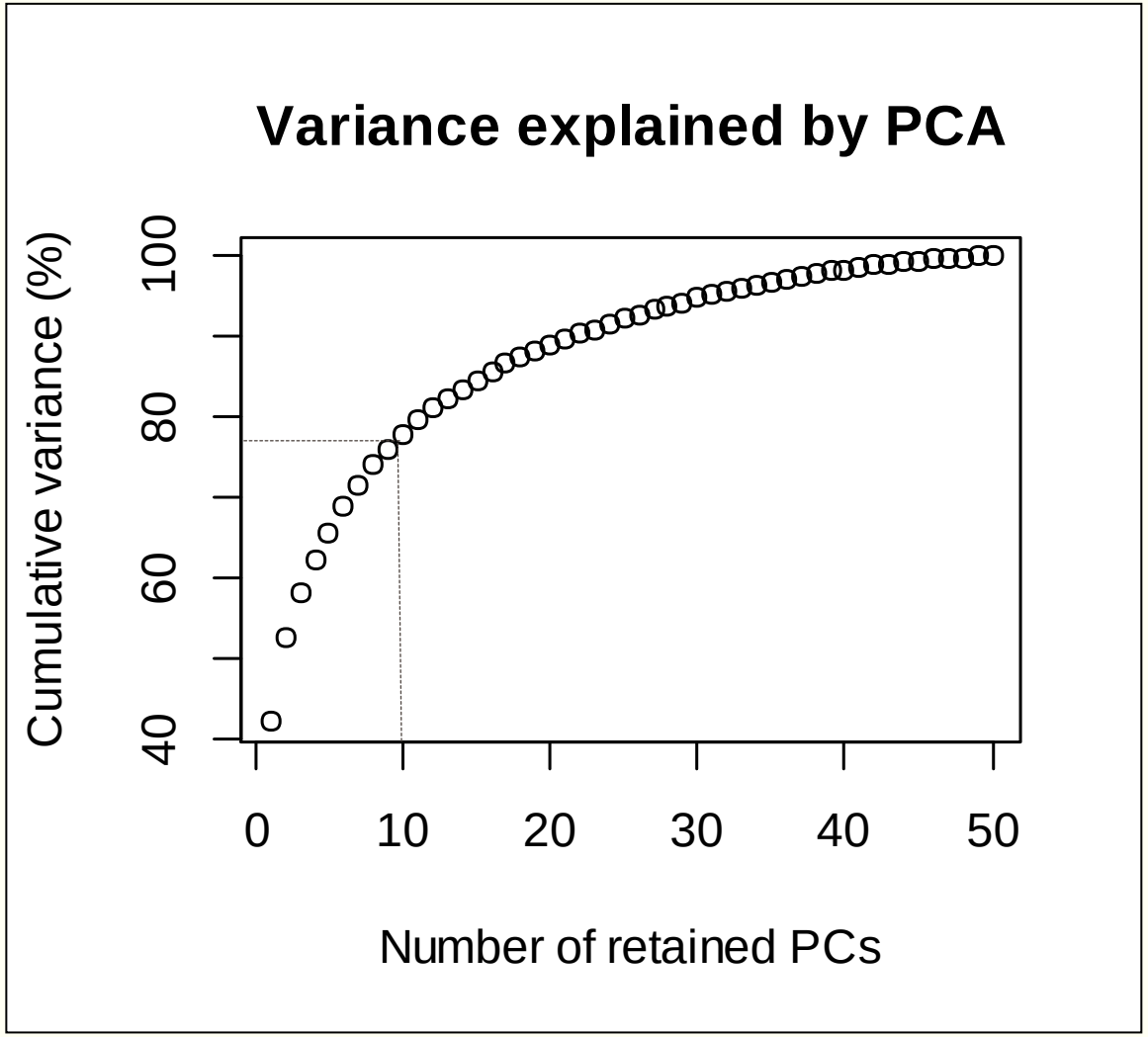
New Group	Original group									
	1	2	3	4	5	6	7	8	10	11
1	5	3	3	2	12	2	2	2	0	2
2	0	0	0	0	0	0	0	0	18	0

- K = 6

New Group	Original Group									
	1	2	3	4	5	6	7	8	10	11
1	5	2	3	0	0	2	0	0	0	0
2	0	0	0	0	0	0	0	0	9	0
3	0	0	0	0	0	0	0	0	9	0
4	0	1	0	0	0	0	2	0	0	0
5	0	0	0	2	12	0	0	2	0	0
6	0	0	0	0	0	0	0	0	0	2

IV. DAPC

- We performed discriminant analysis by considering the original sub-lineages as the groups.
- The command used was ‘dapc()’.
- 10 PCs were selected for DAPC and 4 discriminant functions were retained.
- Output of dapc() includes membership probabilities for each observation.



- This table shows the groups predicted by the functions.
- Since, the strains from a group are getting predicted to be in the same group, this may be the indication that their genetic structure behaves in same way and there might be no admixture. We proceed to the STRUCTURE for reliable results.

Original Group	1	2	3	4	5	6	7	8	10	11
1	**5**	0	0	0	0	0	0	0	0	0
2	0	**3**	0	0	0	0	0	0	0	0
3	0	0	**3**	0	0	0	0	0	0	0
4	0	0	0	**2**	0	0	0	0	0	0
5	0	0	0	0	**12**	0	0	0	0	0
6	0	0	0	0	0	**2**	0	0	0	0
7	0	0	0	0	0	0	**2**	0	0	0
8	0	0	0	0	0	0	0	**2**	0	0
10	0	0	0	0	0	0	0	0	**18**	0
11	0	0	0	0	0	0	0	0	0	**2**

Predicted Group

WHY GO TO STRUCURE ?

- We cannot completely rely on analysis in R, because **the methods implemented are distance based** and they have some **disadvantages** as follows

  1. The clusters identified may be **heavily dependent on distance metric used and visualizations techniques used**.

  2. It is **difficult to measure the confidence** that we should put in the formation of meaningful clusters using these methods.

- Thus, **distance-based methods are more suitable for initial exploratory analysis**.

- **We use STRUCTURE to get clusters based on model-based approach** to eliminate the above disadvantages.

About STRUCTURE

- Based on **Bayesian approach**.
- Considers that the individuals originally come from K populations whose characteristics are '**a set of allele frequency at each locus**'.
- The program **estimates these allele frequencies** and simultaneously assigns individuals to these K populations.
- Considering that an **individual comes from only one of the K original populations** gives rise to '**without admixture**' model in Structure.
- The Structure also provides '**with admixture**' model, it means, **an individual may have come from more than one of K populations**.

MATHEMATICS behind STRUCTURE

Let X : genotype of individual, known
Z : population of origin of individual, unknown
P : allele frequency in all populations, unknown

Priors are assumed for Z and P and we get

$$Pr[Z, P \mid X] \propto Pr[Z] * Pr[P] * Pr[X \mid Z, P]$$

here, $Pr[X \mid Z, P]$ is known.

Since, computation of exact distribution is not possible, approximate sample is obtained from $Pr[Z, P \mid X]$ using MCMC technique and Gibbs sampling.

Based on this, clustering is done.

Priors change according to the model with or without admixture.

STRUCTURE HARVESTER

- Once we get multiple replications for each K value, we **must find an optimum number of clusters, which is K_{opt}. Structure Harvester makes the job easy**.
- Harvester finds K_{opt} by **Evanno method**.
- Evanno et al. presented a paper describing that **the method provided in Structure, which involved likelihoods, did not always pointed to the true value of K** and **hence, they started using an ad-hoc test of ΔK which was a huge success**.
- **ΔK is an ad-hoc quantity based on the rate of change of likelihood function with respect to K**.
- In summary, Harvester uses likelihoods obtained in Bayesian samples to compute ΔK , relates it with each K and **the ΔK, when plotted against Ks, gives peaks at true value of K.**

PIPELINE

Step 1

- MSA in **MEGA**
- PI sites extraction
- Save in required formats
- Check Annotation

Step 2

- **LIAN**
- Conclude about linkage equilibrium
- Add lineages to data for next step

Step 3

- Start **Structure** minimum at 50000-50000
- Increase Burn-in burn-length by 25000 at least per job

Step 4

- Analyze Structure results in Harvester for each run
- Proceed to next step if get 3 constant optimum K values

Step 5

- Analyze membership scores
- Check phylogenetic tress and geography
- Conclude

PREVIOUS ATTEMPTS

- FIRST ATTEMPT

  Included **all the 52 Rubella wild strains.**

  Results were not consistent even after 6 jobs with high burn-in and burn-lengths.

  Had **single strain from a sub-lineage, its assignment to clusters was changing per job**, it was not reliable to comment upon its membership.

  Decided to remove the sub-lineages with number of strains 1 or 2 from the list.

- **SECOND ATTEMPT**

Started working with **43 strains** and had a consistent **optimum K = 5** for the following burn-in burn-length combinations

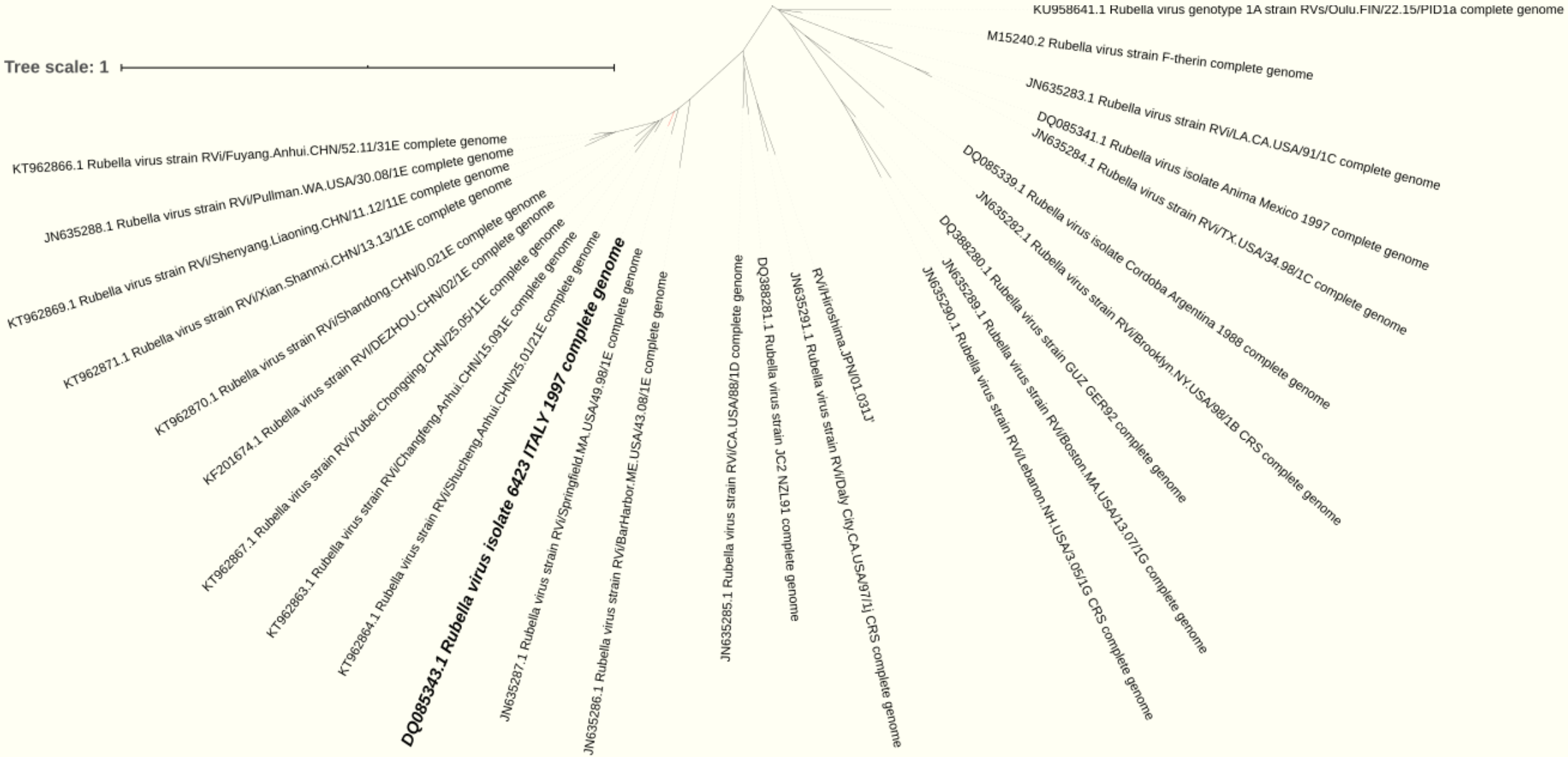
Job number	4	5	6
Burn-in	100000	125000	150000
Burn-length	100000	125000	150000

By the membership score, **a strain from 1D 'DQ085343' was constantly getting clustered with other strains from lineage 1E** and the **same situation was in phylogenetic tree**.

When checked in the NCBI repository, it actually belonged to the Population 5. This was **regarded as an annotation error** in the original data file.

This was corrected, but, **the lineage 1D now remained with only two strains**.

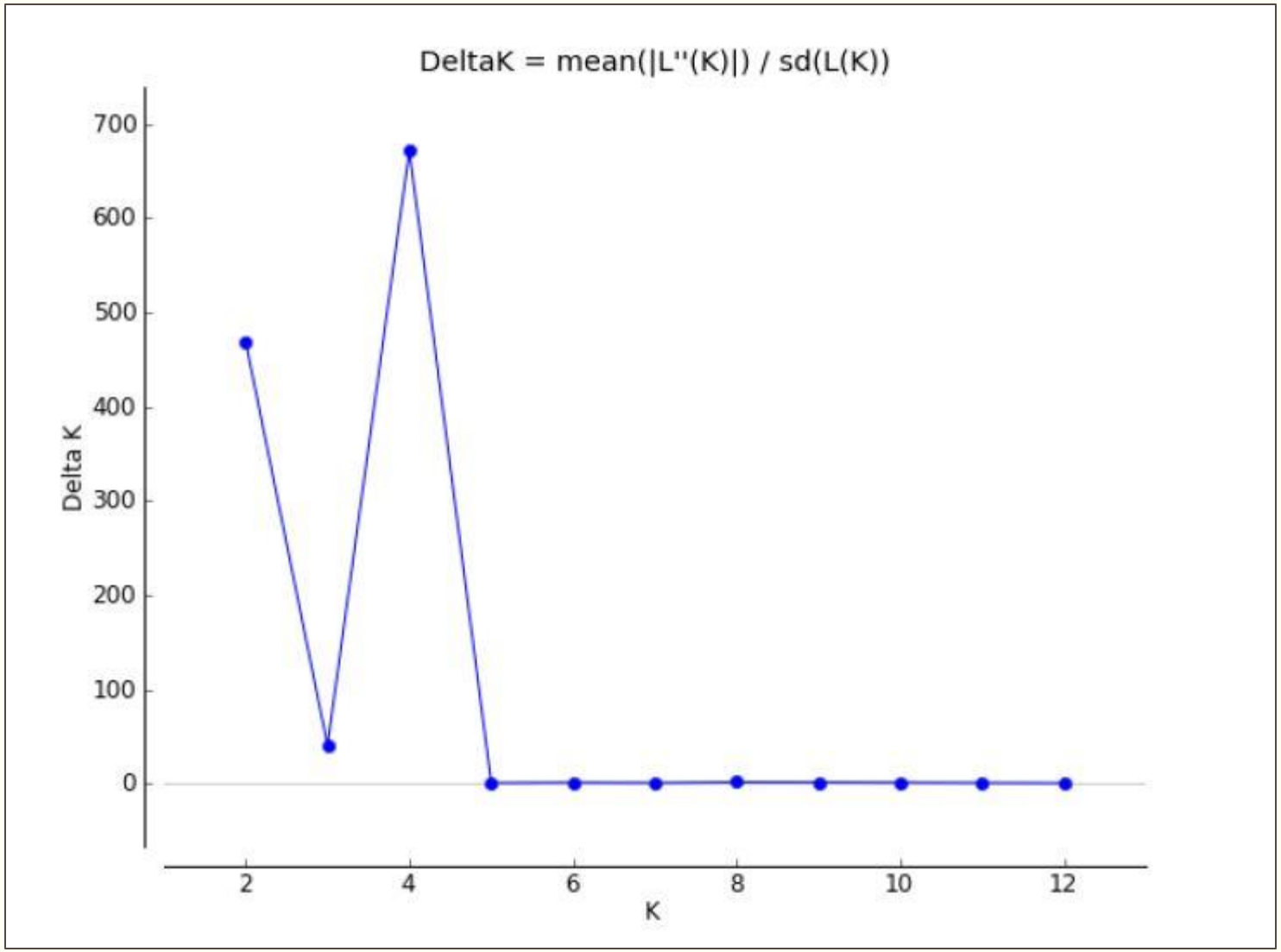
Thus, we started the third attempt just before the lockdown started.



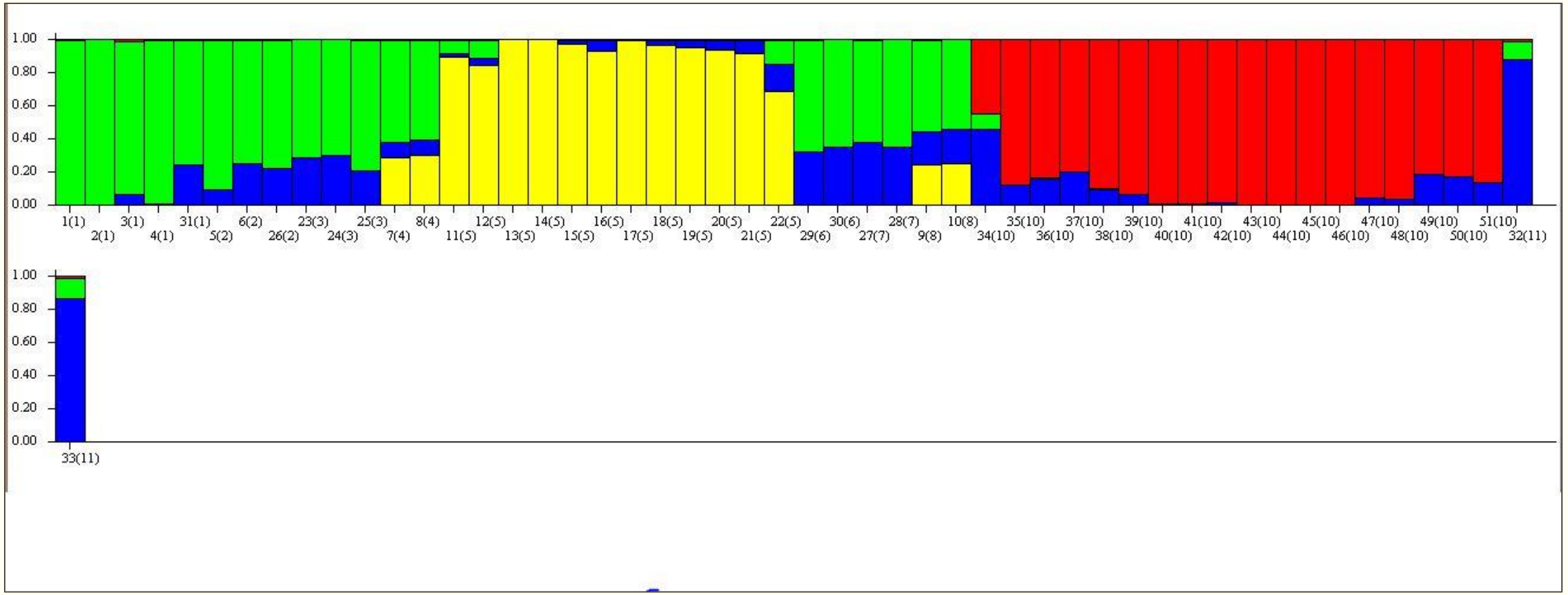
Highlighted strain from 1D is getting placed with strains from 1E

DEMO for STRUCTURE and HARVESTER OUTPUT

- Used data with **51 strains as a part of third attempt**.
- Burn-in and burn-length were 5000 each.
- K ranged from 1 to 13 with 5 replications each.
- The job took 12 hours to complete.
- Harvester plot gave **a clear peak at 4 and 2**.
- The bar plot and membership scores in the Structure were analyzed.
- **The strain 'DQ086338' was found to have large amounts of mixed memberships. 11th group was separate, clearly seen in barplot.**
- Many strains appeared to be admixed, but a definite reason for that would be the extremely low number of simulations.



Plot from Harvester for 5k-5k job



Bar Plot for individuals for K = 4

CONCLUSION

- The data on virus strains was retrieved, processed & modified. The I^S_A value from of LIAN software showed the presence of **weak linkage equilibrium**.
- When analyzing in R, the strain “DQ085338” formed a **cluster of its own**, happens to be **extracted from Israel in 1968**. It was also found to have large amounts of mixed cluster memberships in Structure.
- The separation of 11^{th} group may also be due to the oldness of one of the strains. Strain ‘DQ388279’ was **extracted from Russia in 1967**, another strain “DQ085340” was extracted from Russia according to its a definition.
- The **optimum number of clusters** was found out to be **4 in STRUCTURE** when the **burn-in and burn length values were 5000** each.
- The analysis **using R** packages shows that the **possible number of clusters is 4 to 6** according to the graphs. But, it will also depend on how much discrimination we want to apply.

LIMITATIONS

- For some sub-lineages, there were only single strains available. So it was difficult to cluster those strains. Later, the strains were removed and this **reduced the data**.
- It is a **time consuming process with requirement of high computational power**, hence, could not finish third attempts on local computers.
- Rubella has affected entire world, but, it was noticed that, large amount of strains were extracted from the USA, China and some from Asia-Pacific countries and merely 4-5 strains were obtained from European countries. This imbalance in data will cloud the overall judgment.

SCOPE

- Homogeneity of the sub-populations can be tested using AMOVA (Analysis of Molecular Variance) technique.
- On completion of our third attempt, STRUCTURE results can be compared with R results to understand the difference between distance based and model based clustering techniques.
- It will be interesting to carry out the analysis for diploid organisms.

REFERENCES

- NCBI website, https://www.ncbi.nlm.nih.gov/nucleotide/
- Li H. F., HanW., Zhu Y. F., Shu J. T., Zhang X. Y. and Chen K.W., Analysis of genetic structure and relationship among nine indigenous Chinese chicken populations by the Structure program*, J. Genet.,* 2009, 88, pages 197–203.
- Jonathan K. Pritchard, Matthew Stephens and Peter Donnelly, Inference of Population Structure Using Multilocus Genotype Data, 2000
- G. EVANNO, S. REGNAUT and J . GOUDET, Detecting the number of clusters of individuals using the software STRUCTURE : a simulation study, *Molecular Ecology*, 2005, 14, pages 2611–2620
- Thibaut Jombart*, Genetic Data Analysis Using R : Introduction To Phylogenetics*

- Thibaut Jombart*, An introduction to adegenet 2.0.0*
- Bernhard Houbold, Richard Hudson, Lian 3.0: detecting linkage disequilibrium in multilocus data*, BIOINFORMATICS APPLICATIONS NOTE, 2000, Vol. 16 NO. 9, pages 847-848*
- David A. Lacher And Edward D. O'donnell*,* Comparison Of Multidimensional Scaling And Principal Component Analysis Of Interspecific Variation In Bacteria*, Annals Of Clinical And Laboratory Science,* Vol. 18., No. 6
- Anil Raj, Matthew Stephens, Jonathan K. Pritchard*,* Variational Inference of Population Structure in Large SNP Datasets, 2013
- Jombart et al., Discriminant analysis of principal components: a new method for the analysis of genetically structured populations, *BMC Genetics,* 2010, 11:94, http://www.biomedcentral.com/1471-2156/11/94
- iTOL : Interactive Tree Of Life, https://itol.embl.de/

THANK YOU !