



SELINUS UNIVERSITY
OF SCIENCES AND LITERATURE

**UNSUPERVISED CLUSTERING APPROACH
TO CHARACTERIZE THE CPG ISLAND
DISTRIBUTION OF THE ANDES
HANTAVIRUS**

by Emilio Mastriani

Supervised Coordinator
By Prof. Salvatore Fava Ph.D.
Dr. Prof. Mauro Berta

Scientific Coordinators
Dr. Prof. Shu-Lin Liu Ph.D.

A DISSERTATION

Presented to the Department of
Computational Biology
program at Selinus University

Faculty of Life & Earth Science
in fulfilment of the requirements
for the degree of
Doctor of Philosophy
In Computational Biology

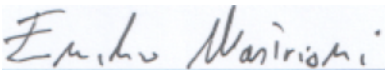
2020

ACKNOWLEDGEMENTS

Writing down my dissertation, I want to thank Prof. Shu-Lin Liu for his availability and professional support. My gratitude goes also to Dr. Alexey Rakov for his constant help and friendly relation. Special thanks to all the collaborators, brothers and sisters that encouraged me to cultivate the passion for the science and the research. A good science can drive to the Truth.

DECLARATION

I hereby attest that I am the first author of this research project titled “UNSUPERVISED CLUSTERING APPROACH TO CHARACTERIZE THE CPG ISLAND DISTRIBUTION OF THE ANDES HANTAVIRUS” and that its contents are only the result of the readings and research I have done. Prof. Shu-Lin Liu is the main contributor to this dissertation and all the information in this research was obtained and presented in accordance with academic rules and ethical conduct. I fully cited by the references all materials and results that not original to this work.

Signature: 

DEDICATION

This thesis is dedicated to God Almighty for the gift of life given to me to complete this work, to my parents Bruno and Cristina, to my parents in law Bartolomeo and Sara, to my wife Serena and my son Andrea for their caring help and infinite patience.

ABSTRACT

Hantaviruses belong to family of *Bunyaviridae* and small mammals host them. Humans are infected either by inhaling virus-containing aerosols or through contact with the animal droppings. Even if rodents host the pathogenic species and humans are considered dead-end hosts, they get accidentally infected and the Andes orthohantaviruses (ANDV) seems to be the unique species for which person-to-person transmission has been documented. Hemorrhagic fever with renal syndrome (HFRS) and hantavirus cardiopulmonary syndrome (HCPS) are two important syndromes associated with hantavirus infections, with a mortality rate close to 40%. CpG repression in RNA viruses has been known for decades and both the estimation of the CpG odds ratio and the correlation with their genome polarity were dominant factors to determine the CpG bias. In this study we conducted the differential analysis of the CpG odds ratio for all the OrthoHantaViruses on the full segmented genomes (L, M, S). The results suggested the statistical significance of the three groups and indicated the “*Small*” genomes as the more informative from the CpG odd ratio point of view. Therefore, focusing the attention to the *small* genomic segments as the more significant with respect to the CpG variation, we calculated the CpG odds ratio for all the OrthoHantaViruses within these segments and estimated the correlation coefficient with the relative coding sequences. Preliminary results confirmed both the CpG odds ratio as the lowest among all the nucleotides and highlighted the Andes virus as that whose CpG odds ratio within CDS is highest. The use of these two measures as features for the three mains unsupervised clustering algorithms has brought to the identification of four different sub-groups inside of the *Orthohantaviridae* family and corroborated the evidence that the Andes Hantavirus (similar, in some way, to Tula H.) exhibits a peculiar CpG odds ratio distribution, perhaps linked to its unique prerogative to pass from human-to-human.

Keywords: Viruses, OrthoHantaViruses, Andes OrthoHantaVirus, Segmented genomes, CpG islands, CpG odd ratio, ANOVA analysis, Unsupervised clustering

Table of Contents

<u>ACKNOWLEDGEMENTS</u>	<u>2</u>
<u>DECLARATION.....</u>	<u>3</u>
<u>DEDICATION.....</u>	<u>4</u>
<u>ABSTRACT.....</u>	<u>5</u>
<u>LIST OF FIGURES AND TABLES.....</u>	<u>8</u>
<u>INTRODUCTION.....</u>	<u>10</u>
HANTAVIRUSES.....	10
ANDES HANTAVIRUS	10
CPG DINUCLEOTIDES IN RNA VIRUSES	12
UNSUPERVISED CLUSTERING AND K-MEANS ALGORITHM.....	13
OBJECTIVE OF THE RESEARCH	14
<u>CHAPTER ONE: SEGMENTED GENOME AND STATISTICAL DIFFERENCE OF CPG ODDS RATIO</u>	<u>16</u>
STATISTICAL SIGNIFICANCE	16
DECISION ABOUT THE NULL HYPOTHESIS.....	18
KRUSKAL-WALLIS TEST INTERPRETATION AND CONCLUSIONS	18
<u>CHAPTER TWO: THE SMALL GENOMIC SEGMENTS CLIQUE AS THE MORE INFORMATIVE GROUP.....</u>	<u>20</u>
DUNN TEST FOR MULTIPLE COMPARISONS OF GROUPS	20
STATISTICAL CLUES TO IDENTIFY THE MOST SIGNIFICANT GROUP WITH RESPECT TO CPG FREQUENCY.....	21
VARIANCE OF THE DINUCLEOTIDE ODD RATIO	21
AVERAGE AND MEDIAN OF VARIANCES FOR THE DINUCLEOTIDE ODD RATIO	22
STATISTICAL ANALYSIS OF CPG ODDS RATIO AND CONCLUSION	23
<u>CHAPTER THREE: INFLUENCE OF THE CPG ODDS RATIO FROM NON-CODING REGIONS.....</u>	<u>26</u>
ODDS RATIO INSIDE CDS REGIONS.....	26
ANDES HANTAVIRUS AND CPG FREQUENCY FROM CDS REGIONS.....	26
CPG ODDS RATIO IN CDS REGIONS AND CONCLUSIONS.....	28
<u>CHAPTER FOUR: UNSUPERVISED LEARNING TO CLUSTERIZE HANTAVIRIDAE FAMILY</u>	<u>30</u>
INTRODUCTION	30

WHAT'S CLUSTERING?	30
OVERVIEW OF CLUSTERING TECHNIQUES	30
CHOOSE THE APPROPRIATE NUMBER OF CLUSTERS	33
UNSUPERVISED CLUSTERING AND HANTAVIRUSES	35
OPTIMAL NUMBER OF CLUSTERS FOR HANTAVIRUSES	35
K-MEANS, DBSCAN AND HCA VS HANTAVIRUS	37
<u>METHODS AND MATERIALS</u>	<u>40</u>
<u>DISCUSSION AND CONCLUSIONS</u>	<u>42</u>
<u>APPENDIX.....</u>	<u>44</u>
LIST OF GENOMIC SEQUENCES	44
LIST OF R SCRIPTS	50
<u>REFERENCES.....</u>	<u>52</u>

List of Figures and Tables

FIGURE 1 REGION OF PERU INDICATING THE HANTAVIRUS TOWNS DESCRIBED IN 1996	11
FIGURE 2 TRANSMISSION TREE FOR HPS CASES IN SOUTHERN ARGENTINA, SEPTEMBER 1996.	12
FIGURE 3 CPG DINUCLEOTIDES. THE 5'—C—PHOSPHATE—G—3' " SEQUENCE OF NUCLEOTIDES, IS INDICATED ON ONE DNA STRAND (UPPER SIDE). ON THE REVERSE DNA STRAND (DOWN SIDE), THE COMPLEMENTARY 5'—CPG—3' SITE IS SHOWN.....	13
FIGURE 4 METHYLATION AND DEAMINATION OF CPG DINUCLEOTIDE. HOW METHYLATION OF CPG FOLLOWED BY SPONTANEOUS DEAMINATION LEADS TO A LACK OF CPG SITES IN METHYLATED DNA.....	13
TABLE 1 NORMALITY TEST PERFORMED USING SHAPIRO-WILK APPROACH	16
FIGURE 5 NORMALITY QQ PLOTS, 1 STAY FOR GROUP L, 2 FOR MEDIUM AND 3 FOR SMALL RESPECTIVELY	17
TABLE 2 TEST OF HOMOGENEITY OF VARIANCE.....	17
FIGURE 6 BOXPLOTS TO VISUALLY CHECK FOR OUTLIERS. 1 STAY FOR GROUP L, 2 FOR MEDIUM AND 3 FOR SMALL RESPECTIVELY	18
TABLE 3 KRUSKAL-WALLIS RANK SUM TEST. COMPARISON OF X BY GROUP.....	20
FIGURE 7 BOXPLOTS REPRESENTATION OF THE DUNN'S TEST.....	21
EQUATION 1 DEFINITION OF VARIANCE	22
EQUATION 2 AVERAGE OF VARIANCES.....	22
EQUATION 3 MEDIAN OF VARIANCES.....	22
EQUATION 4 DISTANCE BETWEEN THE AVERAGE VARIANCE OF A GENERAL DINUCLEOTIDE AND THE CPG VARIANCE.....	23
EQUATION 5 DISTANCE BETWEEN THE MEDIAN VARIANCE OF A GENERAL DINUCLEOTIDE AND THE CPG VARIANCE.....	23
FIGURE 8 VARIANCE OF THE DINUCLEOTIDE FREQUENCY FOR THE THREE GENOMIC GROUPS (L, M AND S).....	24
FIGURE 9 COMPARISON BETWEEN THE ODDS RATIO VARIANCE OF CPG DINUCLEOTIDE AND THE AVERAGE AND MEDIAN VARIANCE FOR GENERIC DINUCLEOTIDE GROUPED BY GENOMIC SEGMENTS (L, M AND S).....	25
FIGURE 10 COMPARISON OF THE DISTANCES BETWEEN THE AVERAGE OF THE VARIANCE FOR ALL THE DINUCLEOTIDES (AVERAGE, BLUE DIAMOND), THE DISTANCE OF CPG ODDS RATIO VARIANCE FROM THE AVERAGE MEASURE (CG_DELTA_AVG, RED SQUARE) AND THE DISTANCE OF CPG ODDS RATIO VARIANCE FROM MEDIAN OF THE VARIANCE FOR ALL THE DINUCLEOTIDES (CG_DELTA_MED, GREEN TRIANGLE). THE VALES ARE GROUPED BY GENOMIC SEGMENT TYPE (L, M AND S).....	25
FIGURE 11 DINUCLEOTIDE ODDS RATIO INTO CDS REGIONS FOR THE 10 VIRUSES. THE CDS REGIONS BELONG TO THE GROUP OF SMALL GENOMIC SEGMENTS.....	26

TABLE 4 CPG ODDS RATIO FROM CDS REGIONS AND FROM FULL GENOME INTO THE GROUP OF SMALL GENOMIC SEGMENTS.....	27
FIGURE 12 ODDS RATIO OF CPG INTO CDS REGIONS.....	27
TABLE 5 COMPARISON (Δ) OF THE CPG ODDS RATIO IN CDS, CPG ODDS RATIO FROM FULL GENOME AND MEDIAN VALUES FOR THE VIRUSES WITH THE TOP FREQUENCIES	28
FIGURE 13 THE ANDES HANTAVIRIDAE SHOWS THE HIGHEST VALUES IN ALL THE THREE CASES (CPG ODDS RATIO INTO CDS, CPG ODDS RATIO FROM FULL GENOME AND MEDIAN VALUES)	28
FIGURE 14 EXAMPLE OF CLUSTERING BASED ON THE SHAPE FEATURE	30
ALGORITHM 1 K-MEANS ALGORITHM	31
ALGORITHM 2 HAC ALGORITHM.....	32
ALGORITHM 3 DBSCAN ALGORITHM	33
FIGURE 15 ELBOW CURVE METHOD	34
ALGORITHM 4 ELBOW CURVE METHOD.....	34
ALGORITHM 5 SILHOUETTE SCORE METHOD	35
FIGURE 16 SILHOUETTE SCORE OPTIMAL K POINT	35
FIGURE 17 OPTIMAL NUMBER OF CLUSTERS ACCORDING TO ELBOW, SILHOUETTE AND GAP METHODS	36
FIGURE 18 CLUSTER TREE REPRESENTATION	37
FIGURE 19 K-MEANS WITH K=4	38
FIGURE 20 DBSCAN AND FOUR GROUPS OF VIRUSES	39
FIGURE 21 HCA DIVISIVE (AGNES)	39
FIGURE 22 HCA CLUSTERING.....	40
FIGURE 23 FLOWCHART OF EXECUTED STEPS TO CALCULATE THE CPG ODDS RATIO	41
TABLE 6 LIST OF LARGE RNA SEQUENCES	44
TABLE 7 LIST OF MEDIUM RNA SEQUENCES	44
TABLE 8 LIST OF SMALL RNA SEQUENCES	45
FIGURE 24 SCRIPT TO CONDUCT ANOVA ANALYSIS IN R.....	50
FIGURE 25 SCRIPT TO CONDUCT THE UNSUPERVISED CLUSTERING IN R.....	51

INTRODUCTION

The current section is intended to introduce the reader to the main arguments argued into the project providing the basic knowledge.

Hantaviruses

Hantaviruses are enveloped RNA viruses with negative-sense, tri-segmented genome. The large (L), the medium (M) and the small (S) code for viral transcriptase or polymerase, glycoprotein precursors (GPC) and the N protein that makes up the nucleocapsid, respectively. [1]. Hantaviruses are transmitted to humans by infected rodents without causing any significant illness in them. There are four rodents in the United States that have been shown to carry the New World hantaviruses: the deer mouse (*Peromyscus maniculatus*), the white-footed mouse (*Peromyscus leucopus*), the rice rat (*Oryzomys palustris*) and the cotton rat (*Sigmodon hispidus*). *Oligoryzomys* spp. rodents appear to be the principal reservoir for most Andes viruses, including the CASV variant [2, 3]. The broad geographic distribution of *Sigmodontinae* and *Oligoryzomidae* rodents suggests that human cases of HCPS will eventually be identified from all countries in the Americas.

Hantavirus cardiopulmonary syndrome (HCPS) is an acute, severe, and sometimes fatal respiratory disease caused by an infection from Andes orthohantavirus. Initial symptoms are linked to the respiratory apparatus (shortness of breath, progressive cough, and tachycardia), muscle aches, fatigue, and fever, making it difficult to distinguish from a simple flu. HCPS symptoms can quickly evolve and, in extreme cases, infected individuals may be incubated and receive oxygen therapy [4]. Complications of cardiogenic shock, lactic acidosis and hemoconcentration can cause death within hours of hospitalization. In South America, Andes hantavirus (ANDV) is the primary etiologic agent. In Chile, over 600 cases of ANDV-related hantavirus have been reported between 2001-2009 with fatality rate of 36%.

Andes hantavirus

Andes OrthoHantaVirus (ANDV) is a major causative agent of hantavirus cardiopulmonary syndrome [5], severe respiratory disease with a fatality rate of 35–40% [6]. Andes orthohantavirus, is the only hantavirus that can spread by human to human by bodily fluids or long-term contact [7-9]. The Andes virus causes the HPS into human hosts and has been identified for the first time in 1995 in samples from patients in southern Argentina [10], even if sporadic cases of HPS have been retrospectively identified [11] in the same country from as early as 1987. In 1995 has been identified for the first time in the lungs of a patient from El Bolson and the outbreak studied in a past dispatch began in September 22, 1996. The Figure 1

reports the towns involved in the 1996 HPS outbreak in southern Argentina, while the Figure 2 shows the transmission tree for the HPS cases in the same outbreak, indicating dates of onset of symptoms, survivor status and hypothetical line of transmission. *Oligoryzomys* spp. rodents appear to be the principal reservoirs for most Andes viruses [3]. A previous study [12] presented the *N. spinosus* mice as a reservoir for the Andes virus variant found in Madre de Dios and Puno. If these mice will be confirmed as reservoir for this virus, the human population at risk for hantavirus infection by transmission from *N. spinosus* mice could be large.



Figure 1 Region of Peru indicating the Hantavirus towns described in 1996

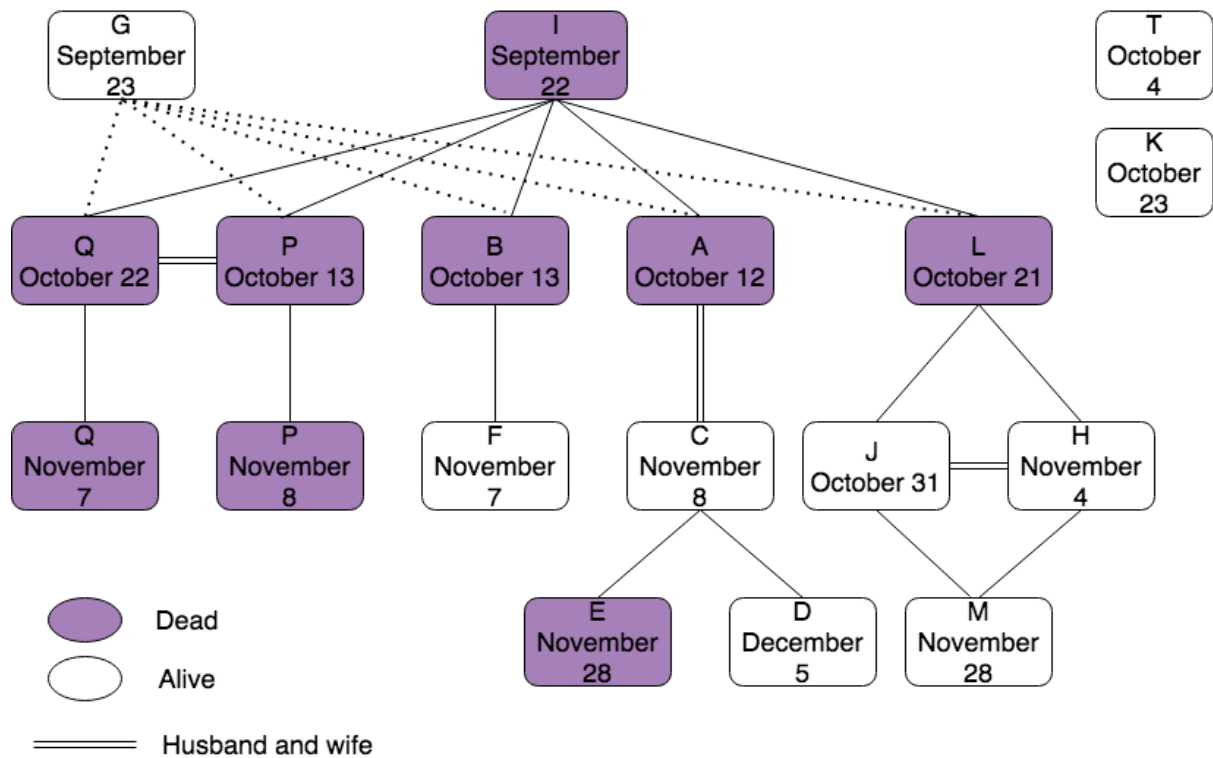


Figure 2 Transmission tree for HPS cases in southern Argentina, September 1996.

CpG dinucleotides in RNA viruses

The CpG sites are regions of DNA or RNA where a cytosine nucleotide is followed by a guanine nucleotide in the linear sequence of bases along the 5' – 3' direction, as illustrated by Figure 3. CpG sites occur with high frequency in genomic regions called CpG islands. CpG dinucleotides have long been observed to occur with a much lower frequency in the sequence of vertebrate genomes than would be expected due to random chance. For example, the frequency of CpG dinucleotides in human genomes is less than one-fifth of expected frequency. This underrepresentation is a consequence of the high mutation rate of methylated CpG sites: the spontaneously occurring deamination of methylated cytosine results in thymine, and the resulting G:T mismatched bases are often improperly resolved to A:T; whereas the deamination of cytosine results in uracil, which as a foreign base is quickly replaced by a cytosine (base excision repair mechanism). Figure 4 depicts the process just mentioned. The transition rate at methylated CpG sites is ~10 fold higher than at unmethylated sites. Thus, the overrepresentation of CpA and TpG is considered to be a consequence of the underrepresentation of CpG. CpG has also been observed to be predominantly under-represented in RNA viruses [13, 14] and the mechanism that contributes to the deficiency in case of riboviruses (RNA nucleic acid) is largely unknown. Because riboviruses do not form DNA intermediates during genome replication, the methylation-deamination model is unlikely to apply, while the *host innate immunity model evasion* seems to be more appropriate. In fact the CpG odds ratio values

of mammals-infecting riboviruses are lower than the riboviruses infecting other taxa and the CpG motif in an AU-rich oligonucleotide can significantly stimulate the immune response of plasmacytoid dendritic cells [15]. Previous research also pointed out the huge variations of CpG bias in RNA viruses and brought out the observed under-representation of CpG in RNA viruses as not caused by the biased CpG usage in the non-coding regions but determined mainly by the coding regions [16].

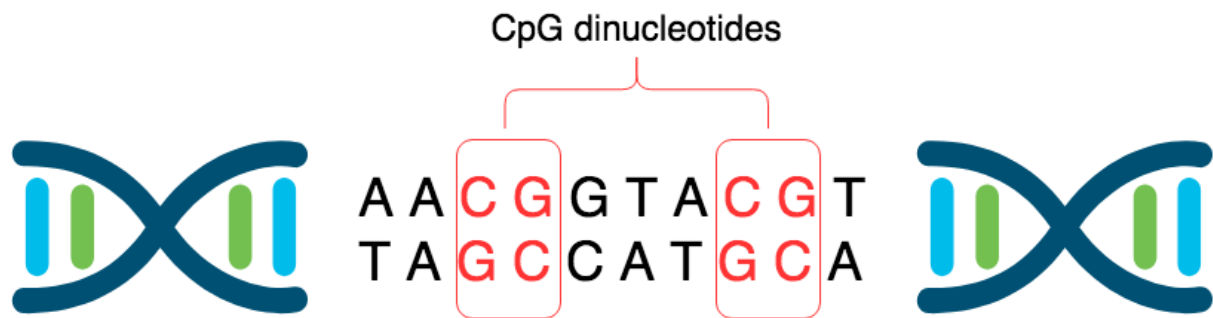


Figure 3 CpG dinucleotides. The 5'—C—phosphate—G—3' " sequence of nucleotides, is indicated on one DNA strand (upper side). On the reverse DNA strand (down side), the complementary 5'—CpG—3' site is shown.

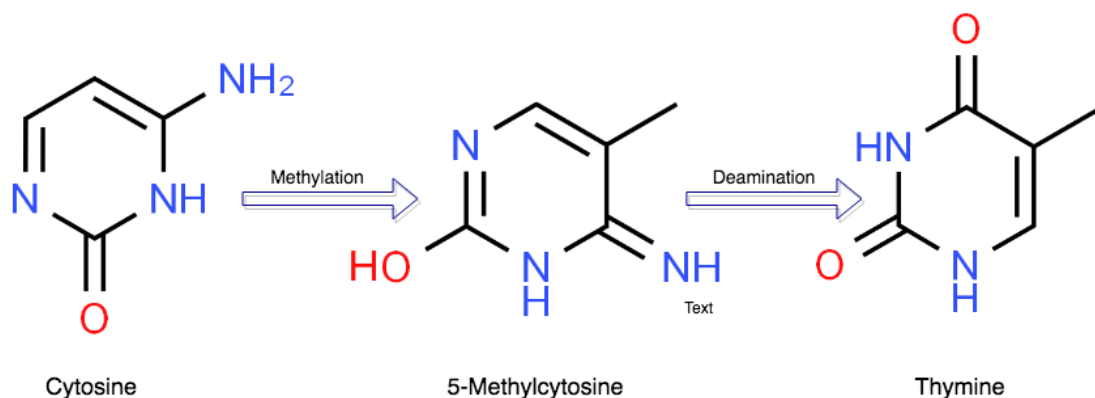


Figure 4 Methylation and Deamination of CpG dinucleotide. How methylation of CpG followed by spontaneous deamination leads to a lack of CpG sites in methylated DNA

Unsupervised Clustering and K-means algorithm

With the term “*unsupervised*”, we define a procedure that uses unlabeled data in its classification process. Unsupervised learning can be thought of as finding patterns in the data beyond what would be considered pure unstructured noise. With unsupervised learning it is possible to learn larger and more complex models than with supervised learning. This is because in supervised learning one is trying to find the connection between two sets of observations, while unsupervised learning tries to identify certain latent variables that caused a single set of observations.

The difference between supervised learning and unsupervised learning can be thought of as the difference between discriminant analysis from cluster analysis and K-means [17] is an

unsupervised clustering algorithm for partitioning unlabeled data into a pre-defined “k” number of distinct groupings. Otherwise stated, k-means algorithm discovers observations sharing important characteristics and classifies them together into clusters. If the algorithm can identify clusters such that the *inside-cluster* observations are more similar than the clusters themselves, then this is a good clustering solution. A plethora of existing algorithms make this job being one of the most widely used techniques for market or customer segmentation. In fact, ever the company’s data can be segregated into clusters and used to identify certain patterns which leads to a more customized approach. Cluster analysis is also widely used for exploratory data analysis to find hidden patterns or grouping in data and K-means is an algorithm that finds these groupings in big datasets. Choosing a value for k (the number of clusters) and randomly setting an initial centroid (center coordinates) for each cluster, the algorithm will assign each observation to its nearest center and update the centroids as being the center of their respective observation. Finally, reaching the step of *no-further-changes* in the clusters, the algorithm will converge providing the final clustering.

Objective of the research

The aim of the study undertaken is to understand if it is possible to characterize the vast family of orthohantaviruses using the odds ratio of the dinucleotide CpG as a marker. Obtaining confirmation that this dinucleotide odds ratio is so characterizing that it discriminates between groups of viruses belonging to the same family could provide useful information to better define the role of CpG islands in orthohantaviruses. This need is dictated both by the recurrent manifestation of acute pulmonary syndrome in America due to this virus, and by the urgency to understand why the Andes hantavirus is the only virus of the family with an anthroponotic transmission. In order to achieve this goal, we used an ANOVA statistical approach to verify the actual statistical difference between the different genomic segments and to focus the research on the most significant genomic group. This initial approach to the problem has provided us with a first index of characterization. The study of the correlation between the CpG dinucleotide ratio index relating to the entire genomic segment and that relating to the coding regions, confirmed the importance of CpG islands in CDS regions for the orthohantaviruses and provided us with a second characterization index. Given the nature of the information available, i.e. a collection of CpG dinucleotide frequency odds ratios on different orthohantaviruses (without any specific target), we used the characterization indices identified as features for the main unsupervised clustering algorithms, obtaining further confirmation the

importance of the CpG dinucleotide odds ratio to isolate Andes hantavirus as a group in its own right.

CHAPTER ONE: SEGMENTED GENOME AND STATISTICAL DIFFERENCE OF CpG ODDS RATIO

As already mentioned, three pieces of the genomes compose the Hantaviruses RNA repertoire: the large one (about 6.5 kb long), the medium (about 3.6 kb) and the small one (1.7 kb). As the first question to answer, we would like to know whether the CpG odds ratio could be used as a marker to discriminate the three groups. To address the question, we considered the samples of all OrthoHantaViruses from human hosts and computed the CpG odds ratio over the full-size genome of all the 236 segments (27 large, 29 medium, 170 small).

Statistical significance

Taking as null hypothesis (H_0) that the means values of the CpG odds ratio from the three groups (L, M and S) is equal, we wish to apply for the analysis of variance (ANOVA) to accept or reject H_0 . As a principle, the **normality** property (according to which the outcome variable must follow a normal distribution in each sub population) is the first assumption to use ANOVA.

To check the assumption, we based on the formality tests of Shapiro-Wilk with the $\alpha=0.05$, while the QQ plot-chart have been used as graphical method. Table 1 reports the results for the normality test.

Table 1 Normality test performed using Shapiro-Wilk approach

Group	Statistics	p-value
L	0.808	0.000192
M	0.878	0.000563
S	0.982	0.0290

To determine if the data is normally distributed by looking at the Shapiro-Wilk results, we just need to look at the “p-value” column and consider the two cases:

- P-value < 0.05, then this would indicate a significant result, i.e. the data is not normally distributed
- P-value > 0.05 in the Shapiro-Wilk test, this would suggest that the data is normally distributed

Considering the QQ plots, the vast majority of points should follow the theoretical normal reference line and fall within the curved 95% bootstrapped confidence bands to be considered

normally distributed. Figure 5 reports the distribution of the CpG odds ratio for all the three groups.

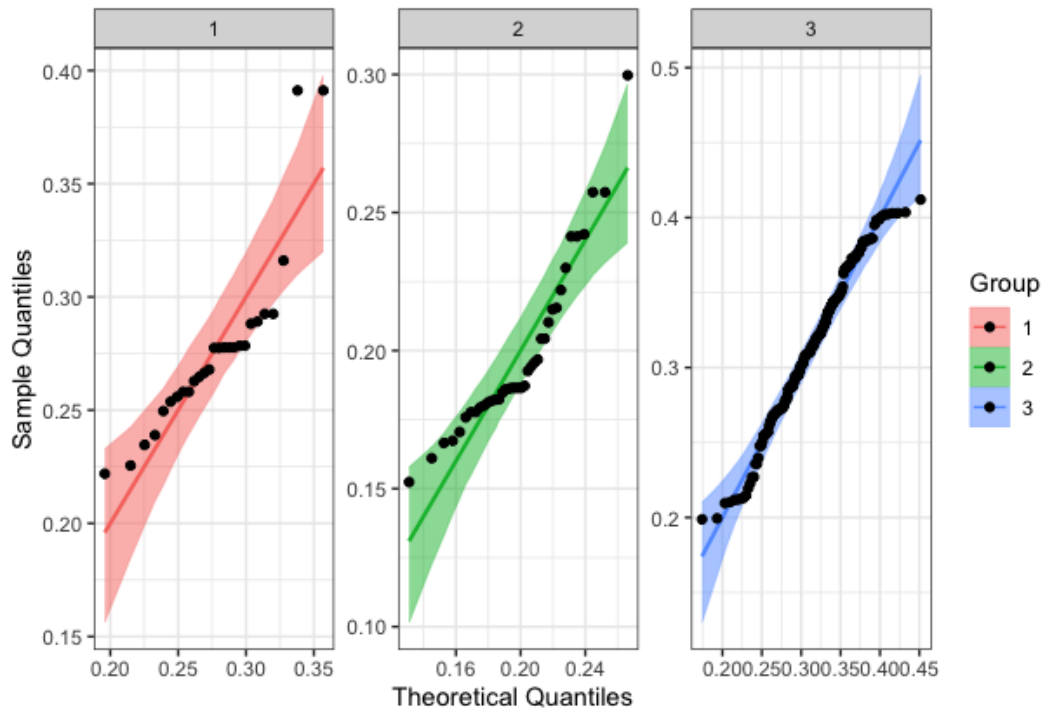


Figure 5 Normality QQ plots, 1 stay for group L, 2 for Medium and 3 for Small respectively

Even if the ANOVA test is considered robust for moderate violation of the normality assumption, both the Shapiro-Wilk and the QQ-plots suggest to perform an equivalent non-parametric test such as a Kruskal-Wallis Test that doesn't require the assumption of normality. The **homogeneity** of variances (according to which the variance within all subpopulations must be equal) is the second property to consider when using ANOVA. Levene's Test for Homogeneity of variance is performed using the traditional mean centered methodology and using R's default median centered methodology. The null hypothesis for this test is that variances are equal across groups. The alternative hypothesis is that variances are unequal for at least one of our treatment groups.

Table 2 Test of homogeneity of variance

Levene's Test for Homogeneity of Variance (center = "mean")		
Df	F value	Pr (>F)
2	8.356	0.0003128
Levene's Test for Homogeneity of Variance (center = "median")		
2	9.3875	0.0001199

Table 2 displays the test statistic for 2 different versions of Levene’s test. In our study, a p-value = 0.0003128 or 0.0001199 indicates that we reject the null hypothesis and conclude that variances are not equal. The boxplot reported in Figure 6 also indicate some major outliers, enough evidence to suggest we move to a different analysis method. Therefore, we will be using the **Kruskal-Wallis** ANOVA as non-parametric test results after checking the assumptions.

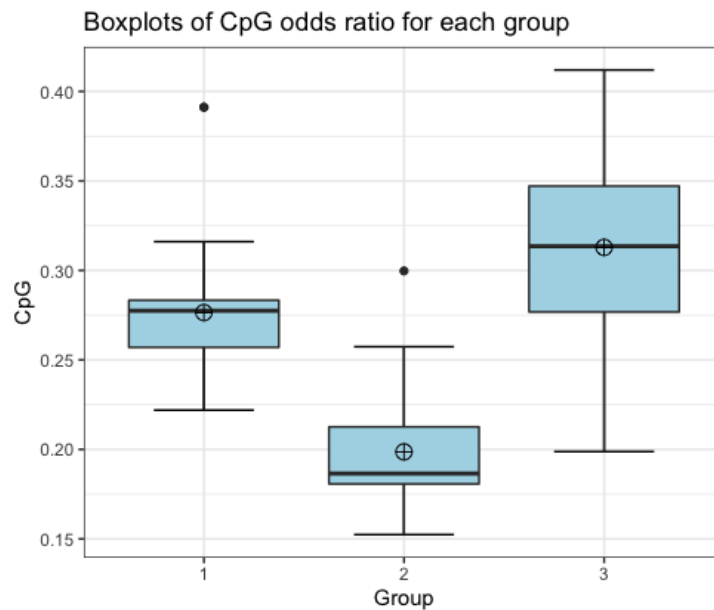


Figure 6 Boxplots to Visually Check for Outliers. 1 stay for group L, 2 for Medium and 3 for Small respectively

Decision about the null hypothesis

So far, we have determined that the data for each treatment group is not normally distributed, and we have major influential outliers. As a result, a Kruskal-Wallis test would be more appropriate than a one-way ANOVA to test for significant differences between genomic segments groups.

Performing the Kruskal-Wallis test, it is observed that $\chi^2 = 95.81 > \chi^2_{\alpha} = 5.991$, $p - value < 2.2e^{-16}$ and given our $\alpha=0.05$, we would reject our null hypothesis and conclude that there is a statistically significant difference in the CpG odds ratio that is calculated for each group of segmented genome.

Kruskal-Wallis Test Interpretation and Conclusions

We have concluded that the CpG odds ratio in genomic segments groups L and M are not normally distributed, and genomic group S is marginally non-normal. In addition, outliers exist for groups L and M. As a result, a Kruskal-Wallis [18] test is more appropriate than a

traditional one-way ANOVA to compare the CpG odds ratio over of three separate genomic groups.

The Kruskal-Wallis test results in a two-sided test $p - value < 2.2e^{-16}$. This indicates that we should reject the null hypothesis that mean ranks are equal across groups and conclude that there is a significant difference in CpG odds ratio distribution. Descriptive statistics indicate that the median value with 95% confidence intervals for group L is 0.277, group M is 0.187, and group S is 0.314. That is to say, the difference between the median values of each segments L and M is about 0.09 ($p=1.137969e-04$), segments L and S is about 0.037 ($p=7.471942e-04$), and segments M and S is about 0.127 ($p=2.173163e-21$).

CHAPTER TWO: THE SMALL GENOMIC SEGMENTS CLIQUE AS THE MORE INFORMATIVE GROUP

From the output of the Kruskal-Wallis test, we know that there is a significant difference between groups, but we do not know which pairs of groups are different neither which group will be more significative from the CpG odds ratio point of view. In this context, a post-hoc analysis can be performed to determine which groups differ from each other, and more measures can be collected to identify the group to focus on.

Dunn test for multiple comparisons of groups

Dunn's Multiple Comparison Test [19, 20] is a post hoc (i.e. it is run after an ANOVA) non parametric test (a "distribution free" test that does not assume your data comes from a particular distribution). In detail, it tests for stochastic dominance and reports the results among multiple pairwise comparisons after a Kruskal-Wallis test for stochastic dominance among k groups. The function we used (*dunn.test*) makes $m = \frac{k(k-1)}{2}$ multiple pairwise comparisons based on Dunn's z -test-statistic approximations to the actual rank statistics. The null hypothesis for each pairwise comparison is that the probability of observing a randomly selected value from the first group that is larger than a randomly selected value from the second group equals one half, and so rejecting H_0 based on $p \leq \alpha/2$. Several options are available to adjust p -values for multiple comparisons, including methods to control the family-wise error rate (FWER) and methods to control the false discovery rate (FDR). In our study, we used the Bonferroni adjustment (FWER) to control the Dunn's test, and adjusted p -values = $\max(1, pm)$. Table 3 reports results from the Dunn's test and those comparisons rejected with the Bonferroni adjustment at the α level (two-sided test) are starred. Figure 7 shows the test output between groups, suggesting that the difference between the group n. 3 (Small segments) and the other groups is significant.

Table 3 Kruskal-Wallis rank sum test. Comparison of x by group

<i>Pairwise comparisons</i>	<i>Z statistic</i>	<i>adjusted p-value</i>
<i>L-M</i>	4.025317	(0.0001)*
<i>L-S</i>	-3.371649	(0.0011)*
<i>M-S</i>	-9.610156	(0.0000)*

Kruskal-Wallis, $\chi^2(2) = 95.81, p = <0.0001, n = 236$

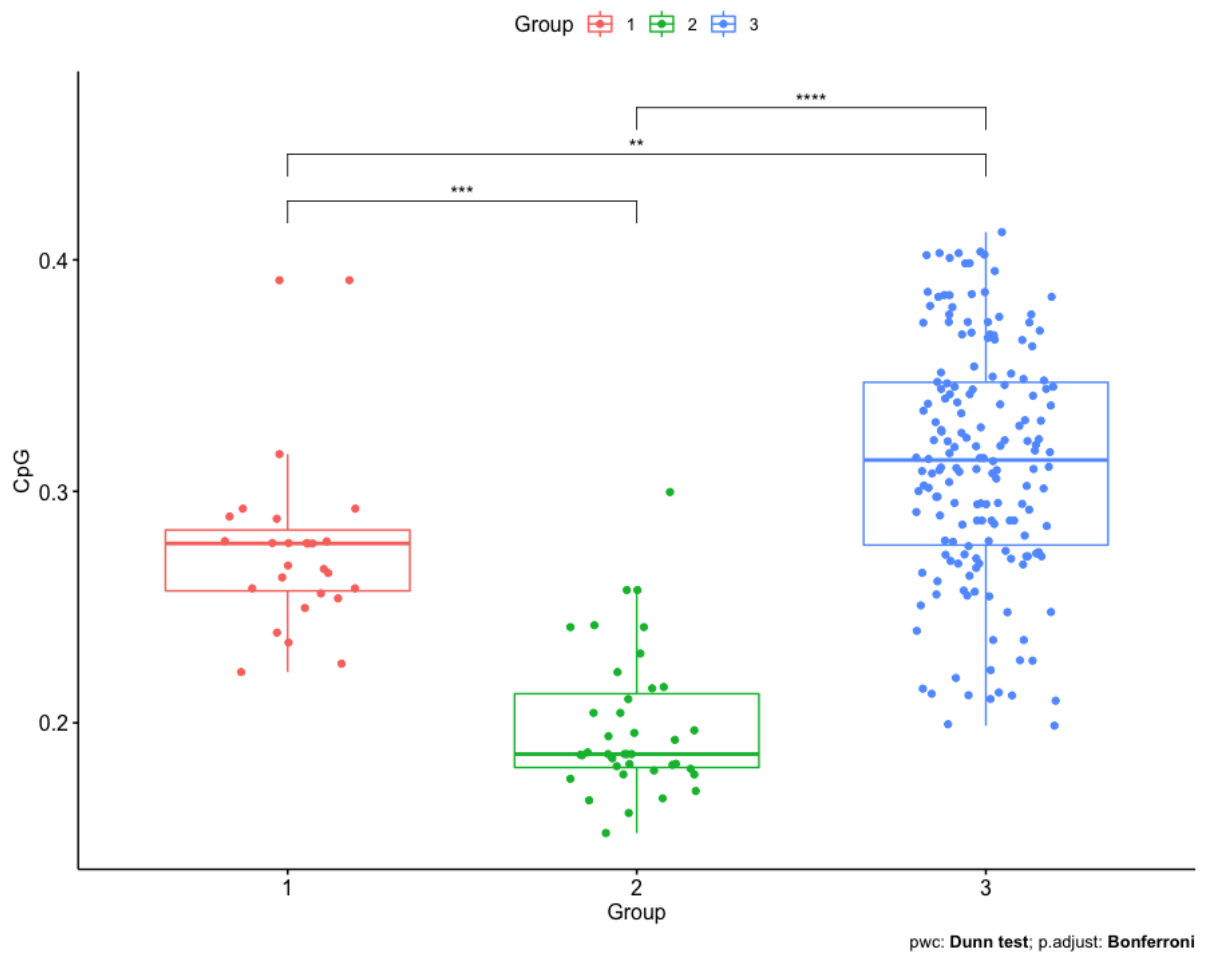


Figure 7 Boxplots representation of the Dunn's test

Statistical clues to identify the most significant group with respect to CpG frequency

We investigated for more statistical clues to identify the more meaningful group with respect to the CpG odds ratio. In our study, one group could be more meaningful than one other whether it presents a wider range of variation for the CpG odds ratio.

Variance of the dinucleotide odd ratio

Variance (σ^2) in statistics is a measurement of the spread between numbers in a data set. That is, it measures how far each number in the set is from the mean and therefore from every other number in the set. Variance is calculated by taking the differences between each number in the data set and the mean, then squaring the differences to make them positive, and finally dividing the sum of the squares by the number of values in the data set.

The Equation 1 reports the formula used to compute the variance

Equation 1 Definition of variance

$$\sigma^2 = \sum_{i=1}^n \frac{(x_i - \mu)^2}{n}$$

where:

- x_i is the i^{th} data point
- μ is the mean of all data points
- n is the number of data points

The Figure 8 reports the variance of the odds ratio for every di-nucleotide in each group of genomic segments, showing as the CpG tends to be more conservative in comparison with the other dinucleotides. In particular, the group of small genomic segments presents the lower value of variation for the CpG odds ratio close to 0.002, suggesting that variation inside of this group should be biologically relevant.

Average and median of variances for the dinucleotide odd ratio

Let us introduce two measures that we will use in the coming section. The Equation 2 defines the average value of the variances of the odds ratio over all the dinucleotides ($n=16$) into each group as:

Equation 2 Average of variances

$$AVG(\sigma_{TT}^2, \sigma_{TC}^2, \dots, \sigma_{GG}^2) = \begin{cases} i = 1 \rightarrow \text{dinucleotide} = TT \\ i = 2 \rightarrow \text{dinucleotide} = TC \\ \dots \\ i = 16 \rightarrow \text{dinucleotide} = GG \end{cases} \rightarrow \mu_{\sigma_{O/E}^2 \text{dinu}} = \frac{\sum_{i=1}^n \sigma_i^2}{n}$$

The Equation 3 defines the median value along the variances of all the odds ratio for all the dinucleotides ($n=16$) for each group as:

Equation 3 Median of variances

$$\begin{aligned} & Med(\sigma_{TT}^2, \sigma_{TC}^2, \dots, \sigma_{GG}^2) \\ & = \begin{cases} i = 1 \rightarrow \text{dinucleotide} = TT \\ i = 2 \rightarrow \text{dinucleotide} = TC \\ \dots \\ i = 16 \rightarrow \text{dinucleotide} = GG \end{cases} \rightarrow \text{Sort}(\sigma_i^2) \rightarrow idx_{median} = \frac{n+1}{2} \\ & \rightarrow M_{\sigma_{O/E}^2 \text{dinu}} = \sigma_{idx_{median}}^2 \end{aligned}$$

Equation 4 introduces the concept of distance between the variance of the odds ratio for the CpG dinucleotide and the average value of all the variances of all the frequencies for all the dinucleotides:

Equation 4 Distance between the average variance of a general dinucleotide and the CpG variance

$$\Delta_{\mu} = |\sigma_{O_{CG}}^2 - \mu_{\sigma_{O/E_{dinu}}^2}|$$

Finally, equation 5 represents the distance between the median of the odds ratio for the CpG dinucleotide and the median value of all the variances of all the frequencies for all the dinucleotides:

Equation 5 Distance between the median variance of a general dinucleotide and the CpG variance

$$\Delta_M = |\sigma_{O_{CG}}^2 - M_{\sigma_{O/E_{dinu}}^2}|$$

The diagram reported in Figure 9 shows that while the measures do not represent meaningful differences in case of large and medium genomic segments, the small genomic group looks to depict a more interesting situation. In fact, the value of the variance for the CpG is far away the median and average values of the dinucleotides from the other groups.

The observation becomes more evident from the diagram in Figure 10, where the distance values are indicated. For each group of genomic segments (L, M and S), we estimated the following measures:

1. $\mu_{\sigma_{O/E_{dinu}}^2}$, average of the variance for all the dinucleotides
2. Δ_{μ} , distance of CpG odds ratio variance from $\mu_{\sigma_{O/E_{dinu}}^2}$
3. Δ_M , distance of CpG odds ratio variance from $M_{\sigma_{O/E_{dinu}}^2}$

The comparison of the distances between the average of the variance for all the dinucleotides, the distance of CpG odds ratio variance from the average measure and the distance of CpG odds ratio variance from median of the variance for all the dinucleotides noticed the group of small genomic segments as that for which the Δ measures are bigger. These results together indicate the group of small genomic segments as the more informative from the CpG odds ratio point of view.

Statistical analysis of CpG odds ratio and Conclusion

The statistical analysis showed above pointed out two important results. The first one is that the value of the variance for the CpG islands into the group of small genomic segments is far away both from the median and average values of the dinucleotides from the large and medium genomic segment groups. Secondly, the group of small genomic segments has the bigger differential measures (Δ) compared to the other groups. These results together confirm the

statistical significance of the three groups and indicate the group of small genomic segments as the more informative from the CpG odds ratio point of view.

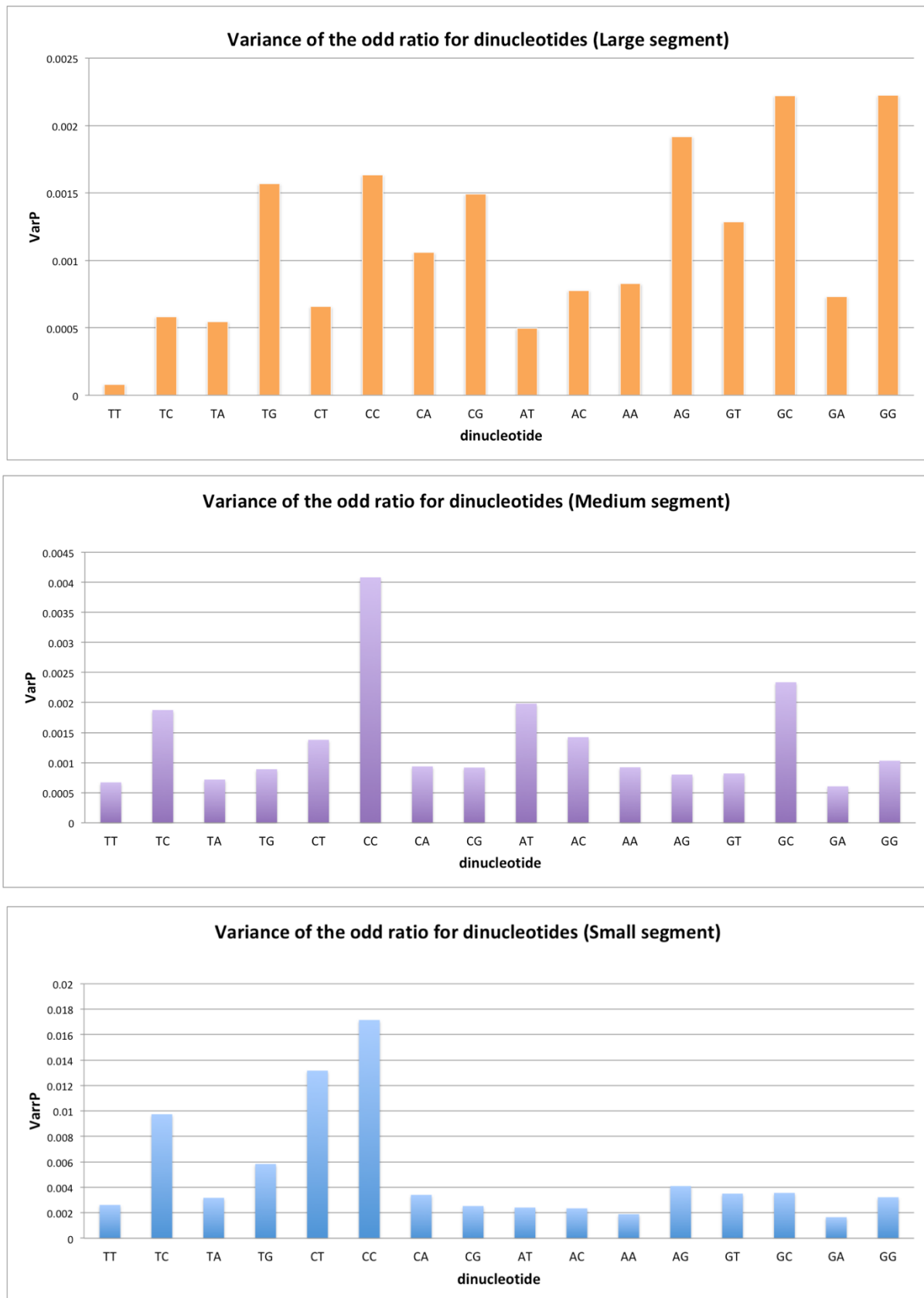


Figure 8 Variance of the dinucleotide frequency for the three genomic groups (L, M and S)

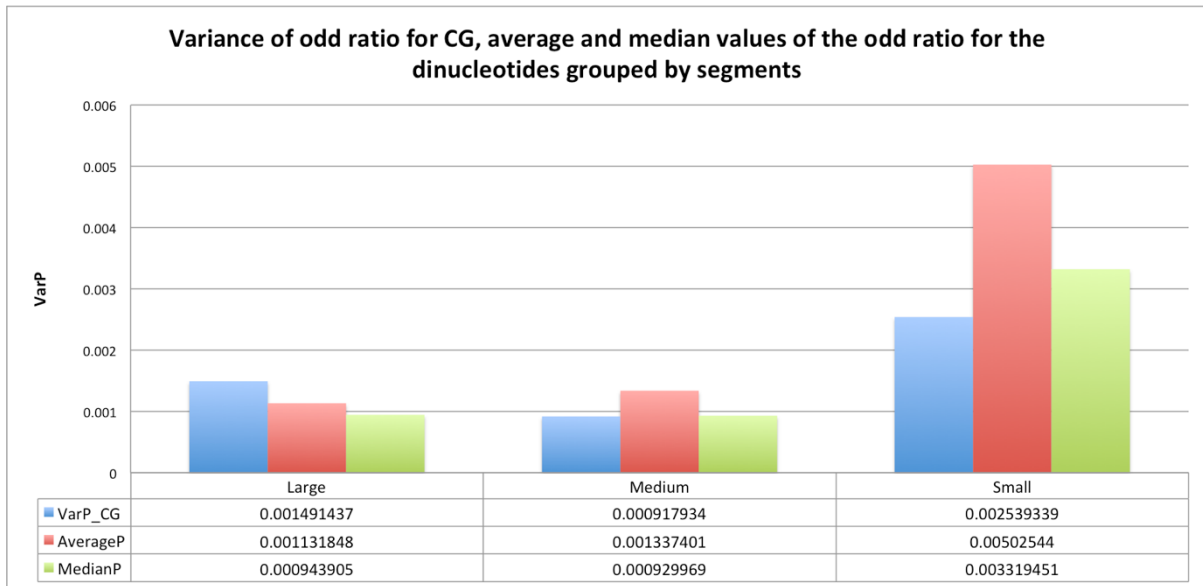


Figure 9 Comparison between the odds ratio variance of CpG dinucleotide and the average and median variance for generic dinucleotide grouped by genomic segments (L, M and S).

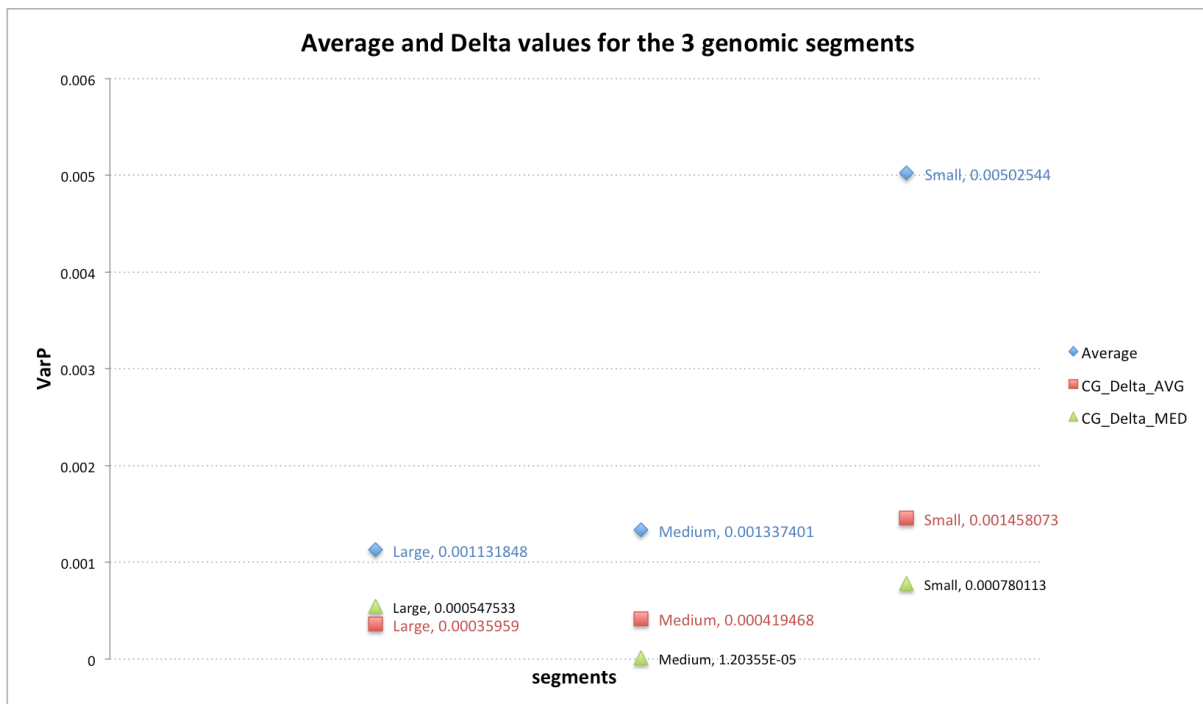


Figure 10 Comparison of the distances between the average of the variance for all the dinucleotides (Average, blue diamond), the distance of CpG odds ratio variance from the Average measure (CG_Delta_AVG, red square) and the distance of CpG odds ratio variance from Median of the variance for all the dinucleotides (CG_Delta_MED, green triangle). The values are grouped by genomic segment type (L, M and S)

CHAPTER THREE: INFLUENCE OF THE CpG ODDS RATIO FROM NON-CODING REGIONS

To avoid the influence of the CpG odds ratio from the non-coding regions, we firstly calculated the CpG odds ratio into the coding regions for the small segments of all viruses. Secondly, we considered the correlation coefficient between the CpG odds ratio and the CpG odds ratio of the coding regions from the group of small genomic group of all the viruses keeping in mind that positive correlation implies a more significative CpG odds ratio from the small genomic segment group.

Odds ratio inside CDS regions

Previous studies already underlined the CpG odds ratio as the lowest compared to those of the other dinucleotides, even in case of RNA viruses [16]. The calculation of the odds ratio for all the dinucleotides around into the CDS regions, restricted our study to 10 different RNA viruses from the *Hantaviridae* family: Andes, Tunari, Bayou, Choclo, Dobrava-Belgrade, Hantaan, Hantaanvirus, Puumala, Seoul and Tula. Furthermore, it confirmed the CpG odds ratio into CDS as the lowest also for group of small genomic segments. In fact, as showed by the Figure 11, the odds ratio for CpG in CDS regions is the lowest compared to the odds ratio of other dinucleotides for the 10 viruses considered.

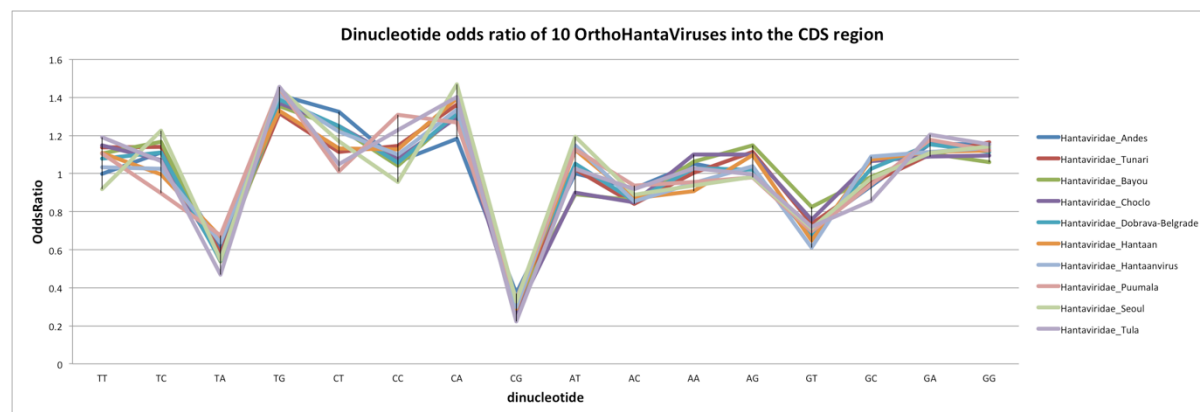


Figure 11 Dinucleotide odds ratio into CDS regions for the 10 viruses. The CDS regions belong to the group of small genomic segments

Andes hantavirus and CpG frequency from CDS regions

Considering independently the frequency of CpG inside the coding regions, it is evident how *Hantaviridae* Andes can present itself as a particular case. Starting from the data reported in Table 4, we calculated the Pearson correlation coefficient obtaining a value close to 0.98. Such as result confirms the positive correlation between the CpG odds ratio over the full genome and the CpG odds ratio into the CDS, highlighting the possible function of the CpG dinucleotides into the coding regions. Analyzing more carefully the data contained into the

Table 4, it immediately catches the eye how the CpG frequency in CDS for the Andes *Hantaviridae* represents the highest value, 7.58% greater than the second highest value (*Hantaviridae* Dobrava-Belgrade). The odds ratio bars depicted in Figure 12 show even more as the Andes hantavirus detains the highest CpG odds ratio into CDS regions compared to the other hantaviruses.

Table 4 CpG odds ratio from CDS regions and from full genome into the group of small genomic segments

Virus	Odds ratio CpG into CDS	Odds CpG ratio from full genome
<i>Hantaviridae</i> Andes	0.369086166	0.357064072
<i>Hantaviridae</i> Tunari	0.272528294	0.272530915
<i>Hantaviridae</i> Bayou	0.311789101	0.309152507
<i>Hantaviridae</i> Choclo	0.266112427	0.24787315
<i>Hantaviridae</i> Dobrava-Belgrade	0.341706719	0.327283795
<i>Hantaviridae</i> Hantaan	0.288624107	0.265946324
<i>Hantaviridae</i> Hantaanvirus	0.298984901	0.282896747
<i>Hantaviridae</i> Puumala	0.338483857	0.346475985
<i>Hantaviridae</i> Seoul	0.326166667	0.326014792
<i>Hantaviridae</i> Tula	0.22244768	0.199395228

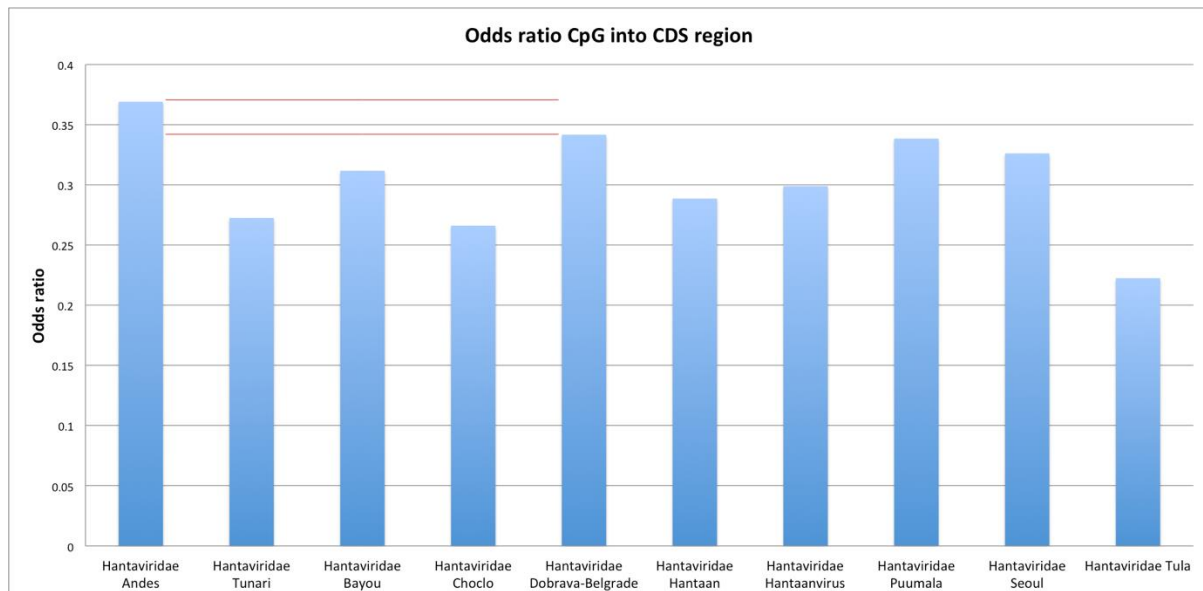


Figure 12 Odds ratio of CpG into CDS regions

Furthermore, considering both the odds ratio of CpG along the full genome, the CpG odds ratio into the CDS regions, and the median value for hantaviruses mentioned above, the *Hantaviridae* Andes holds on the highest values for all three measures. Table 5 and Figure 13 show that the

CpG odds ratio value into CDS of *Hantaviridae* Andes is 7.58% greater than the same value from *Hantaviridae* Dobrava-Belgrade (the second virus sorted by CpG odds ratio into CDS value). And also, *Hantaviridae* Andes is 3.08% and 5.78% greater than *Hantaviridae* Puumala (the second virus for CpG into full genome and CpG median values), regard to the CpG into full genome and CpG median value, respectively.

Table 5 Comparison (Δ) of the CpG odds ratio in CDS, CpG odds ratio from full genome and Median values for the viruses with the top frequencies

	Andes	Dobrava-Belgrade	Puumala	Δ
<i>CpG into CDS</i>	0.369	0.341		0.028
<i>CpG full genome</i>	0.357		0.346	0.011
<i>Median</i>	0.363		0.342	0.021

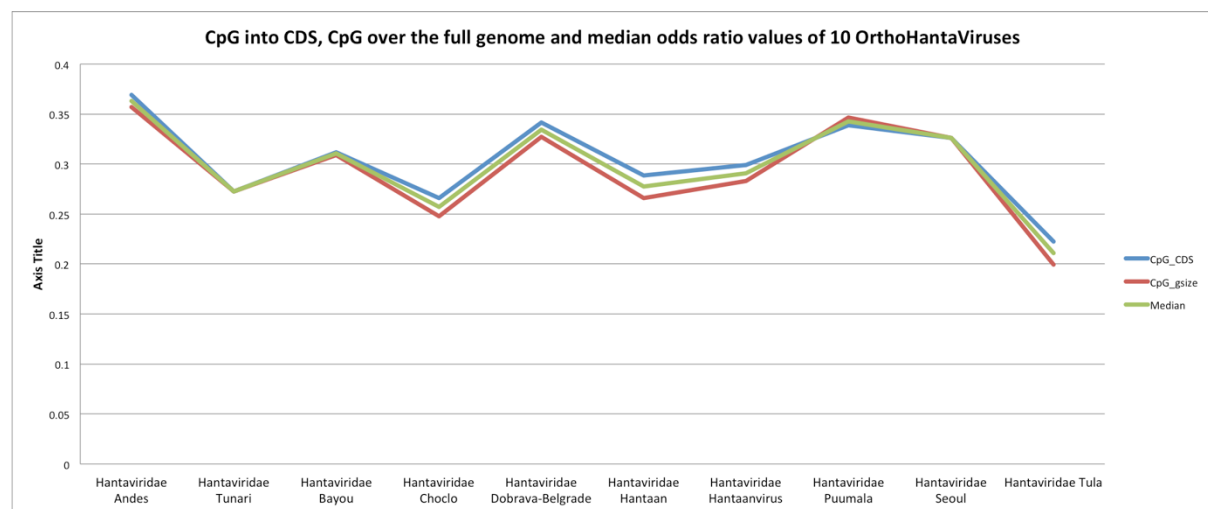


Figure 13 The Andes *Hantaviridae* shows the highest values in all the three cases (CpG odds ratio into CDS, CpG odds ratio from full genome and Median values)

CpG odds ratio in CDS regions and Conclusions

The analysis of the CpG odds ratio into CDS regions of small genomic segments led to several results. Firstly, has been confirmed that also in case of the small genomic segments from Hantaviruses the CpG odds ratio is the lowest one compared to the other dinucleotides placed into the CDS regions. In detail, the Andes *Hantaviridae* brings the highest value of CpG odds ratio into CDS regions. Moreover, the Pearson correlation close to 0.98 confirms the positive correlation between the CpG odds ratio along the full small genomic segment and the CpG odds ratio into the CDS regions of the same genomic segment, stressing the possible roles carried out by the CpG islands into the coding regions. Lastly, the comparison of the CpG odds ratio from the full genome, from the CDS regions and the median values, draw attention to a

stronger concentration of CpG islands both along the full small genomic segment and into the CDS regions for the Andes Hantaviridae.

CHAPTER FOUR: UNSUPERVISED LEARNING TO CLUSTERIZE

HANTAVIRIDAE FAMILY

Introduction

In the current chapter we will move to introduce the clustering problem, the most used techniques to organize data into clusters and the available strategies to compare the results from different clustering algorithms. Then, we will apply the unsupervised clustering technique to the group of small genomic segments from Hantaviridae family to identify eventually subgroups and visualize the position of the Andes Hantaviridae with respect to the other clusters.

What's clustering?

Clustering is a set of techniques used to partition data into groups or clusters. Clusters are defined as groups of data objects that are more similar to other objects in their cluster than they are to data objects in other clusters. Clustering is important because it determines the intrinsic grouping among the present unlabeled data, finding similarity based on features as well as the relationship patterns among data samples.

The Figure 14 reports an example on what does it mean to cluster different objects based on one feature as the shape.

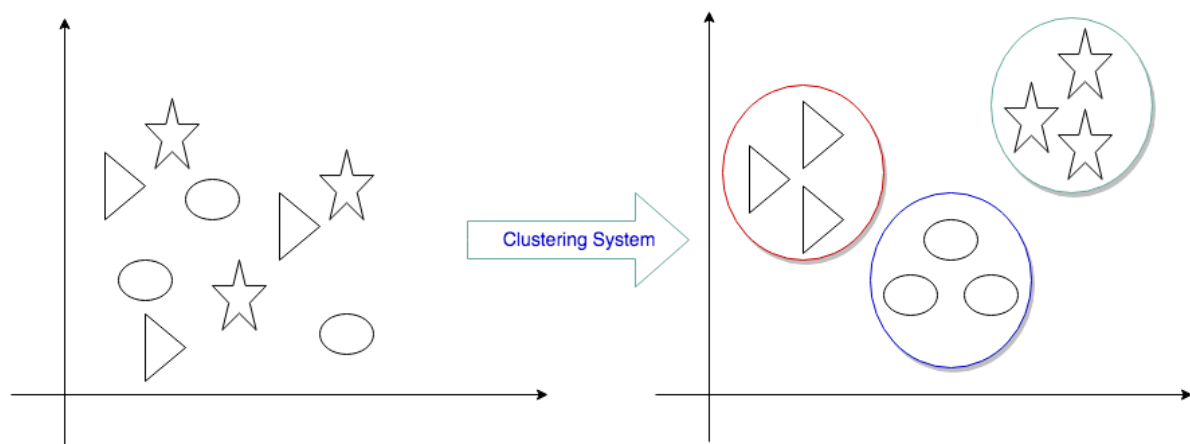


Figure 14 Example of Clustering based on the shape feature

Overview of clustering techniques

Spherical dataset is a form of non-linear dataset in which observational data are modeled by a function which is a non-linear combination of the model parameters and depends on one or more independent variables. In other words, we say that a dataset has a spherical form if

literally its mean data distribution on X, Y is roughly a sphere. Different clustering algorithms work better on different distributions and it is not necessary that clusters will be formed in spherical form. Followings are three popular categories of clustering algorithms.

Partitional clustering and k-means algorithm

The cluster are formed by partitioning the objects into k clusters. It divides data objects into non-overlapping groups through an iterative process to assign subsets of data points into k clusters. This kind of algorithm is defined as **non-deterministic** because it could produce different results from different running on the same data input. They have several strengths as working well when clusters have a spherical shape and being scalable with respect to problem complexity. Example of partitional clustering algorithms are k-means, k-medoids and CLARANS. In details, k-means clustering algorithm (also called flat clustering algorithm) [17] computes the centroids and iterates until it finds optimal centroid, assuming that the number of clusters are already known. **Centroids** are data points representing the center of the cluster. The main element of the k-means algorithm is the **expectation-maximization** approach used to solve the problem. The Expectation-step is used for assigning the data points to the closest cluster and the Maximization-step is used for computing the centroid of each cluster. Algorithm 1 reports the conventional version of the k-means algorithm:

Algorithm 1 *k-means algorithm*

```
Specify the number k of clusters to assign
Randomly initialize k centroids
repeat
    expectation: Assign each point to its closest centroid
    maximization: Compute the new centroid (mean) of each cluster
until The centroid positions do not change
```

Algorithm 1 K-means algorithm

Hierarchical clustering and Agglomerative algorithm

In these methods, the clusters are formed as a tree structure based on the hierarchy called *dendrogram*. This is implemented by either a bottom-up or a top-down approach. Namely, the *agglomerative* clustering that merges the two points more similar until all points have been merged into a single cluster, and the *divisive* clustering that starts with all points into the same cluster and splits the least similar clusters at each step until only single data points remain. The

process is deterministic, so the cluster assignments will not change after running the algorithm on the same data input. The dendrogram often is easy to be interpreted and reveals fine details about the relationships between data objects. Examples of hierarchical clustering algorithms are CURE (Clustering Using REpresentative), BIRCH (Balanced Iterative Reducing Clustering using Hierarchies) [21]. The Algorithm 2 shows as the canonical hierarchical agglomerative cluster (HAC) algorithm works.

Algorithm 2 *HAC algorithm*

1. Make each data point a single-point cluster → forms N clusters
2. Take the two closest data points and make them one cluster → forms N-1 clusters
3. Take the two closest clusters and make them one cluster → Forms N-2 clusters.
4. Repeat step-3 until you are left with only one cluster.

Algorithm 2 HAC algorithm

Density-based clustering and DBSCAN

In these methods, the clusters are formed as the *dense* region, assignments are based on the density of data points in a region. So, clusters are assigned where there are high density of data points separated by low-density regions. The advantage of these methods is that they have a good accuracy as well as a good ability to merge two clusters. They do not need to know a priori the k value, but a specific threshold will determine how close points must to be considered a cluster member. These kinds of algorithms excel at identifying clusters of non-spherical shapes and are resistant to outliers. Examples of density-based clustering algorithms are DBSCAN [22] and OPTICS [23]. The Algorithm 3 reports the steps executed by DBSCAN to perform the clustering.

Algorithm 3 DBSCAN algorithm

1. Arbitrary select a point P
2. Retrieve all points density-reachable from P wrt Eps and MinPts
3. If P is a core point
4. A cluster is formed
5. if P is border point
6. No points are density-reachable from P
7. Visit the next point in the database
8. Continue the process until all the points have been processed

Algorithm 3 DBSCAN algorithm

Choose the appropriate number of clusters

Determining the optimal number of clusters in a data set is a fundamental issue in partitioning clustering, such as k-means clustering, which requires the user to specify the number of clusters k to be generated. The optimal number of clusters is somehow subjective and depends on the method used for measuring similarities and the parameters used for partitioning. A simple and popular solution consists of inspecting the dendrogram produced using hierarchical clustering to see if it suggests a particular number of clusters. These methods include direct methods and statistical testing methods:

1. Direct methods: consists of optimizing a criterion, such as the within cluster sums of squares or the average silhouette. The corresponding methods are named *elbow curve* and *silhouette score* methods, respectively.
2. Statistical testing methods: consists of comparing evidence against null hypothesis. An example is the *gap statistic*.

Having already used the statistical approach in the previous chapters, here we will focus to the direct methods.

Elbow curve method

The main idea of the elbow curve method [24] is to define clusters such that the total within-cluster sum of square (WSS) is minimized. It measures the compactness of the clustering and we want it to be as small as possible. The idea is to choose a number of clusters (k) so that adding another cluster doesn't improve much better the total WSS. Basically, WSS is the sum of squared distance (usually Euclidean distance) from its nearest centroid (center point of cluster). Of course, it decreases with increasing number of clusters(k) and usually an aim is to find the bend (like an *elbow joint*) point in the graph. Figure 15 represents the elbow output where $k=4$ is the optimal number of clusters, while the Algorithm 4 shows as it works.

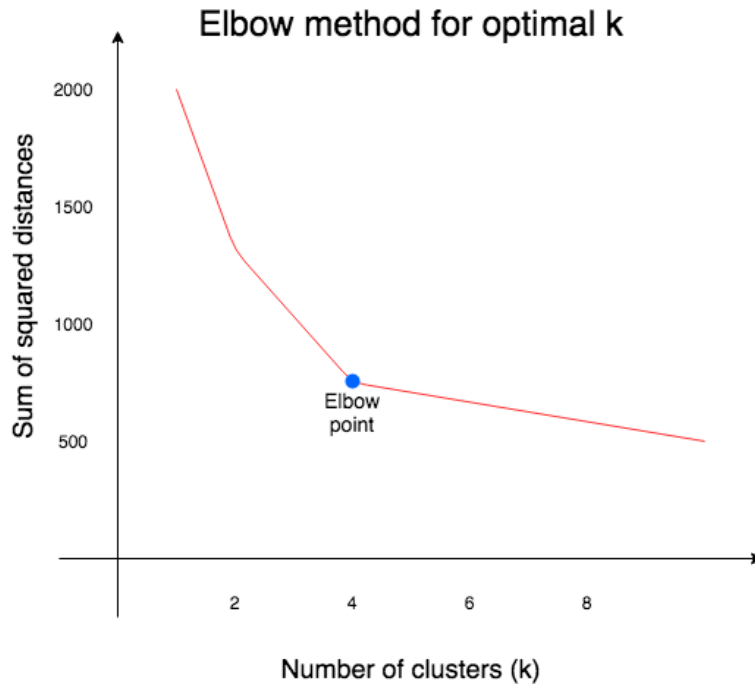


Figure 15 Elbow curve method

Algorithm 4 Elbow method

Compute clustering algorithm (e.g., k-means clustering) for different values of k.

For each k

 Calculate the total within-cluster sum of square (wss).

Plot the curve of wss according to the number of clusters k.

The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters.

Algorithm 4 Elbow curve method

Silhouette score method

Silhouette Score [25] is calculated using mean of intra-cluster distance (a) and the mean of nearest-cluster distance (b) for each sample. The Silhouette Coefficient for a sample is given by $\frac{(b-a)}{\max(a,b)}$. For better clarification, intra-cluster distance (a) is distance of sample point to its centroid and (b) is the distance of sample point to nearest cluster that it is not a part of. Hence, because we want the silhouette score to be maximum, we must find a global maximum for this method (as described by the Algorithm 5). Silhouette coefficient exhibits a *peak* characteristic as compared to the gentle bend in the elbow method. This is easier to visualize and reason with as showed in Figure 16.

Algorithm 5 Silhouette score method

Compute clustering algorithm (e.g., k-means clustering) for different values of k.

For each k, calculate the average silhouette of observations (avg.sil).

Plot the curve of avg.sil according to the number of clusters k.

The location of the maximum is considered as the appropriate number of clusters.

Algorithm 5 Silhouette score method

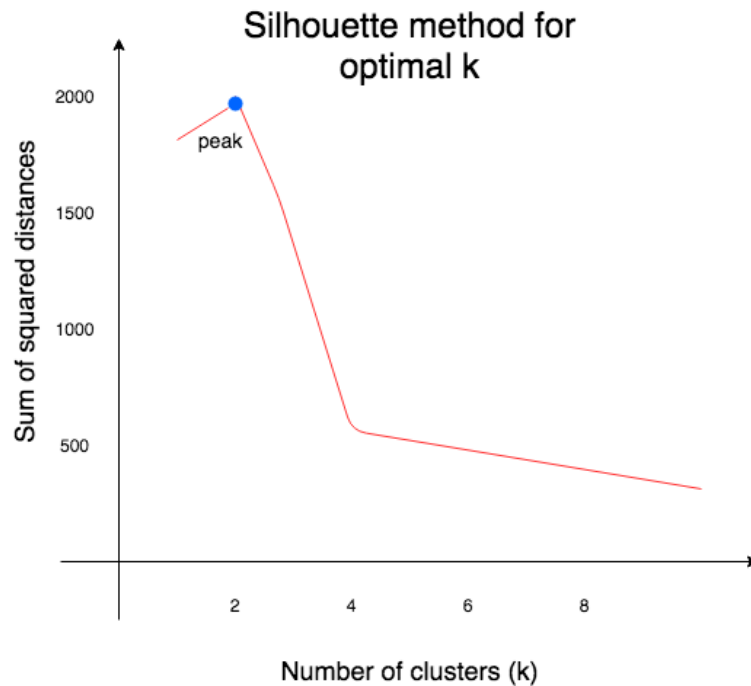


Figure 16 Silhouette score optimal k point

Unsupervised Clustering and Hantaviruses

According to the results obtained in the previous chapters and after the preprocessing standardization of the data, we are going to use both the CpG odds ratio measurements (based on the CDS and on full genome size) from the group of small genomic segments as features for the unsupervised cluster analysis.

Optimal number of clusters for Hantaviruses

In order to find the optimal number of clusters, we used the following three approaches:

1. Elbow curve method
2. Silhouette score method
3. Gap statistic method

Remembering that the Elbow curve method looks at the total within-cluster sum of square (WSS) as a function of the number of clusters, the location of a knee in the plot is usually considered as an indicator of the appropriate number of clusters because it means that adding another cluster does not improve much better the partition. This method seems to suggest $k=4$ as the optimal number of clusters. The Silhouette score method measures the quality of a clustering and determines how well each point lies within its cluster and in our case, it suggests $k=2$ as optimal number of clusters. The optimal number of clusters is the one that maximizes the gap statistic. Approaching the problem by the use of the GAP statistical method, it suggests only 1 cluster (which is therefore a useless clustering). Figure 17 reports all the three results. Giving that all the three approaches suggest a different number of clusters, we chosen to use an alternative approach by considering how samples change groupings as the number of clusters increases. This is useful for showing which clusters are distinct and which are unstable. It does not explicitly tell us which choice of *optimal* clusters is but it is useful for exploring possible choices.

In Figure 18 the size of each node corresponds to the number of samples in each cluster, and the arrows are colored according to the number of samples each cluster receives. In this graph we see that as we move from $k=2$ to $k=3$ a number of viruses from the lookers-left cluster are reassigned to the third cluster on the right. As we move from $k=4$ to $k=5$ we see two nodes with multiple incoming edges an indicator that we over-clustered the data. This is a good indication that we have over clustered the data and that we have reason to set $k=4$ as the optimal number of clusters for our dataset.

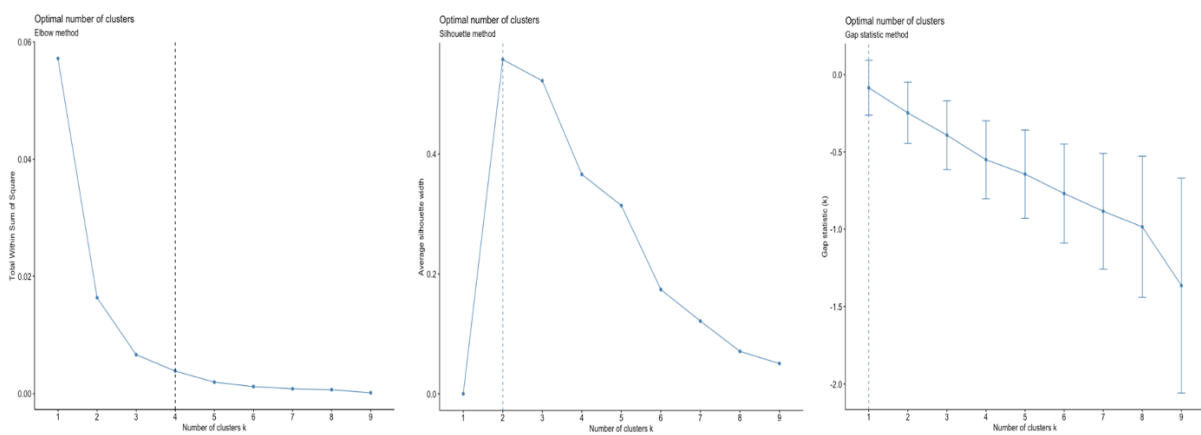


Figure 17 Optimal number of clusters according to Elbow, Silhouette and GAP methods

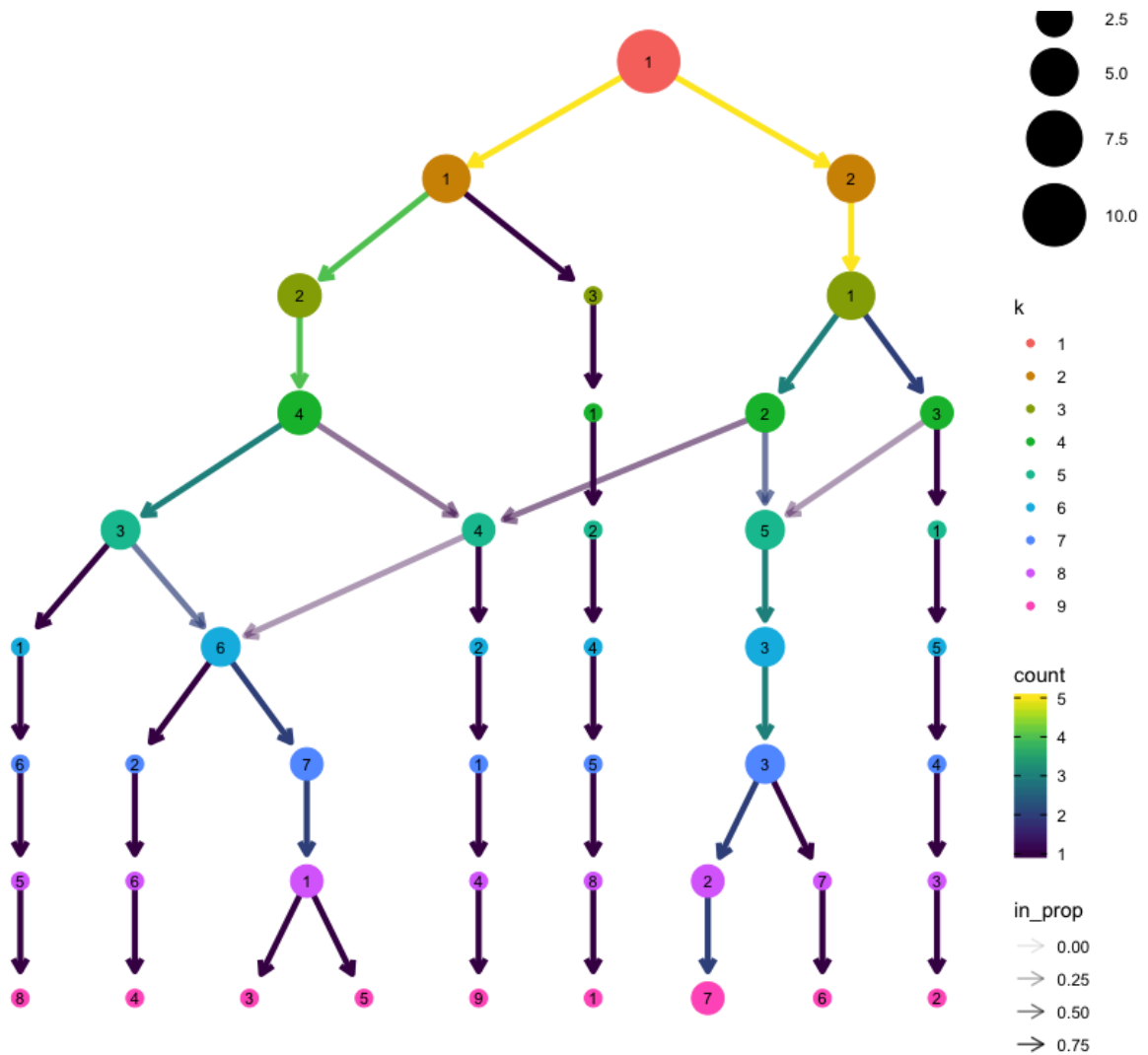


Figure 18 Cluster tree representation

K-means, DBSCAN and HCA vs Hantavirus

Using the number of clusters $k=4$, we executed the three mentioned algorithms of unsupervised clustering to identify the groups of Hantaviruses more similar according to the CpG odds ratio both from full genome and from the CDS regions and to their median values from the group of small genomic segments. We focused attention to the Andes Hantavirus, being the unique hantavirus able to pass from human to human. K-mean algorithm showed the Andes H. as an element of the 4th cluster with the Puumala H., however showing a relevant distance from it (see Figure 19). DBSCAN algorithm showed four groups of viruses, even if the distance between them is not well demarked (see Figure 20). HCA agglomerative and divisive reported the same dendrogram, showing Andes H. as a “border line” virus as the Tula H., even if belonging to two different clusters (see Figure 21). Making a representation of the clustering obtained by the hierarchical methods, we got again evidence that Andes H. looks like an

isolated cluster (as also the Tula H.), suggesting some important difference with the other viruses from the same *Hantaviridae* family (see Figure 22).

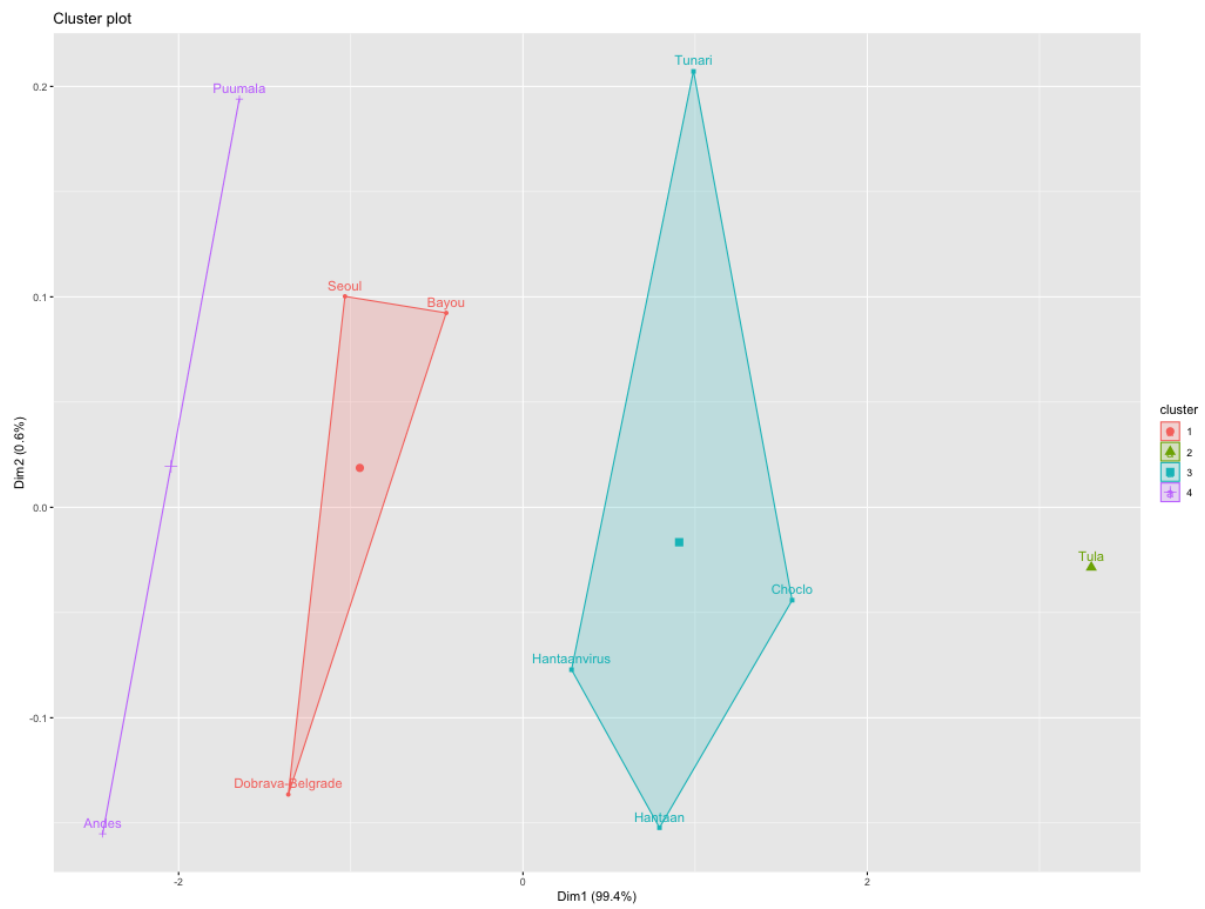


Figure 19 K-means with $k=4$

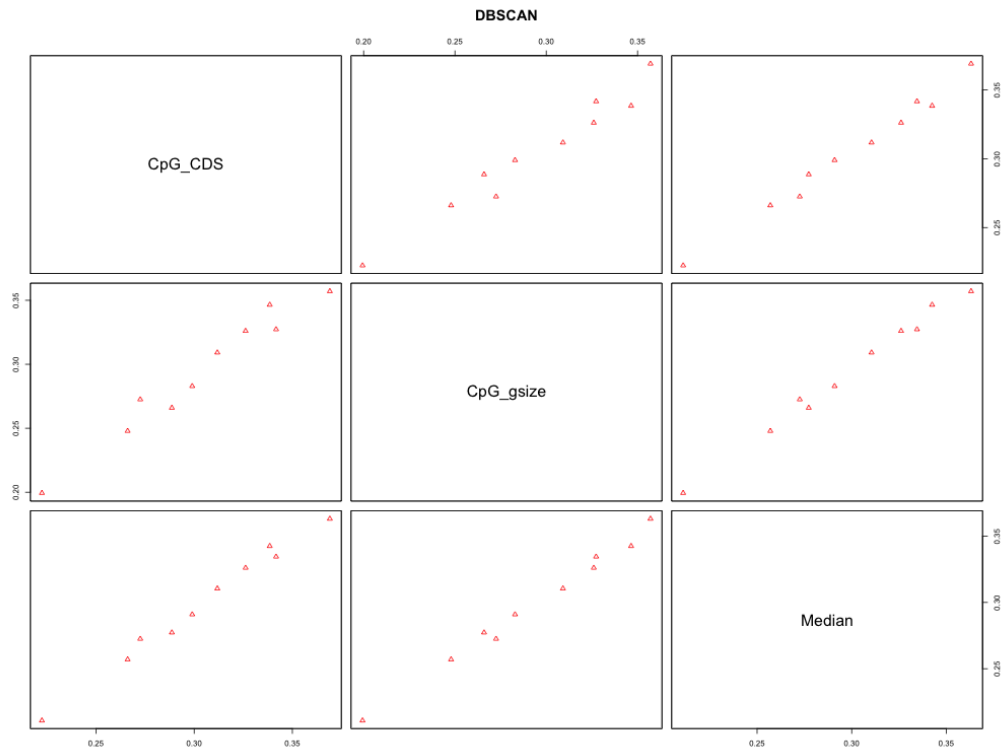


Figure 20 DBSCAN and four groups of viruses

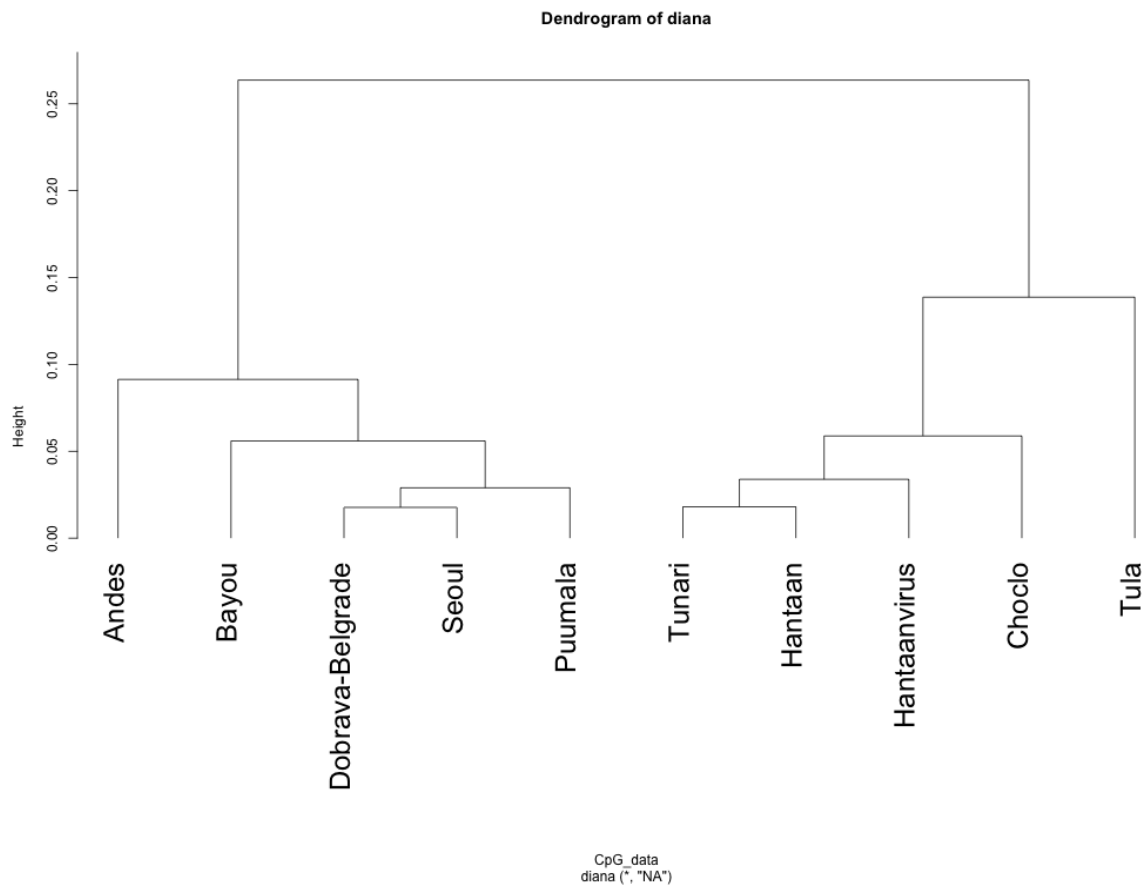


Figure 21 HCA divisive (AGNES)

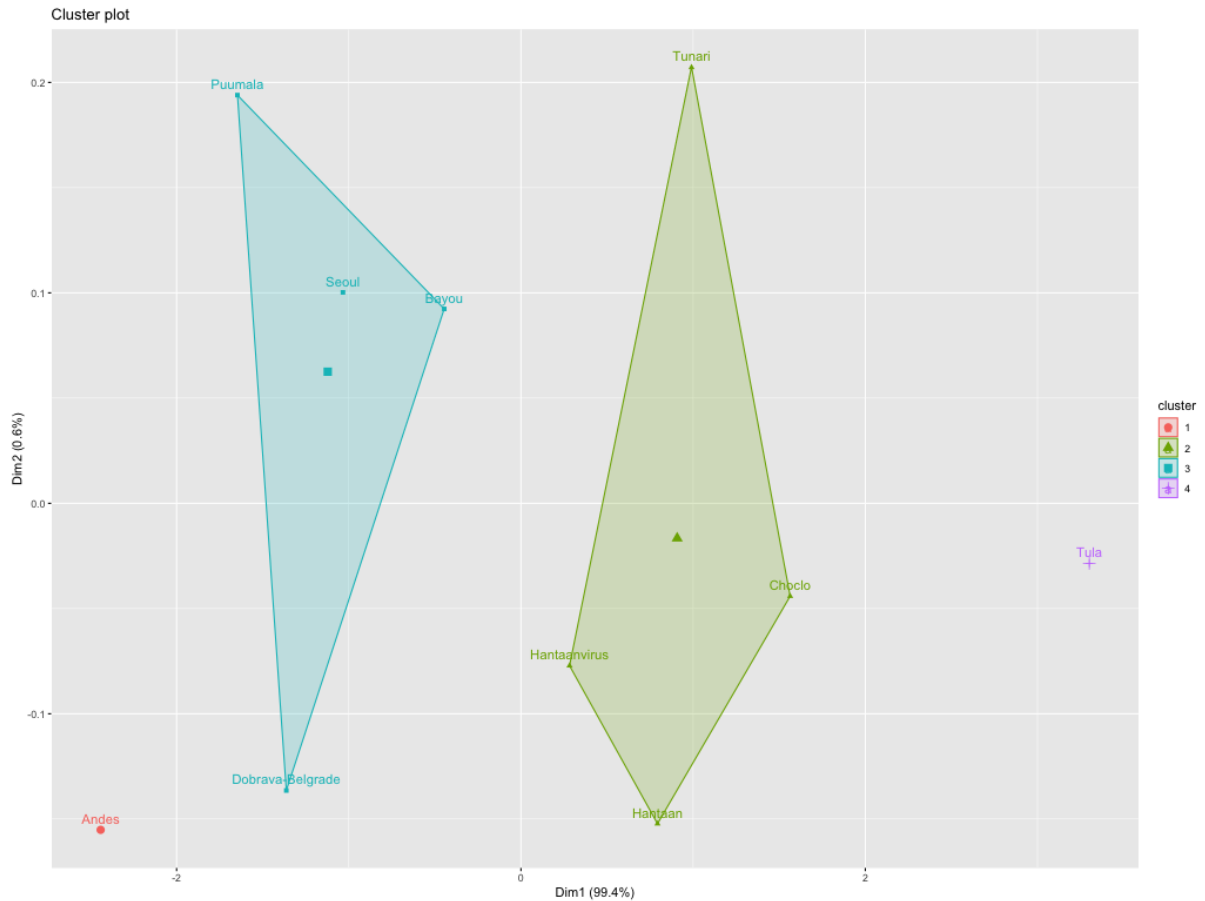
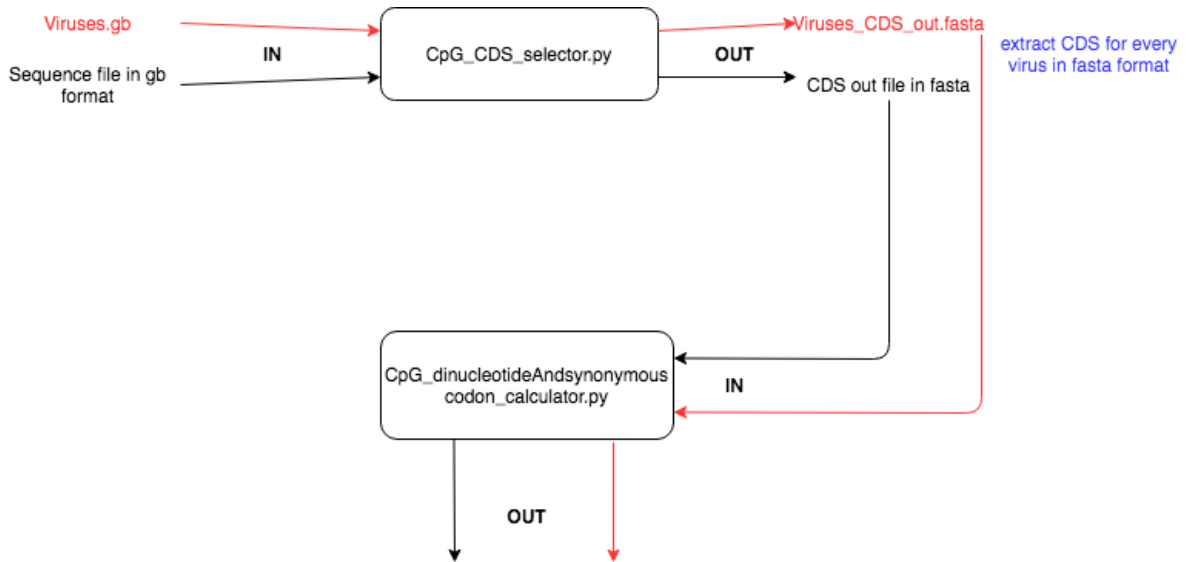


Figure 22 HCA clustering

Methods and materials

The genomic data to accomplish the current study have been downloaded from the ViPR [26] database. The Tables 6-8 in the Appendix section report the complete list of the RNA sequences we treated: 27 RNA sequences of large genomic sequences, 39 sequences of RNA from the medium sized genomic segments and 170 of small genomic RNA sequences, for a total of 236 genomic segments from *Hantaviridae* family. We used R version 3.6.2 and Bio Python version 1.71 to conduct the statistical analysis and make the calculation of the CpG odds ratio, respectively. Figure 23 depicts the steps followed to obtain the CpG odds ratio for all the segmented genomic sequences. Figures 24-25 report the scripts used to conduct the ANOVA analysis and the unsupervised clustering in R.



dinu

Viruses_CDS_dinu.txt

dinucleotide percentage
distribution over CDS (CpG
odd ratio into CDS)

txt

Viruses_CDS_info.txt

informative file (as a log file)

percentage

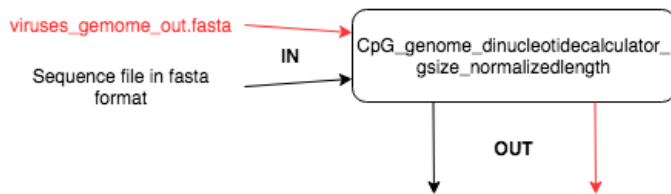
Viruses_CDS_Npercentage.txt

codon percentage frequency

RSCU

Viruses_codon.txt

relative synonymous codon
usage



dinu

Virus_genome_dinu_withNormalized_gsize.txt

dinucleotide percentage
normalized to the genome
size (CpG odd ratio relative to
the genome size)

txt

Virus_genome_info_withNormalized_gsize.txt

informative file (as a log file)

Figure 23 Flowchart of executed steps to calculate the CpG odds ratio

Discussion and Conclusions

In the current study, we analyzed the *Orthohantaviridae* family from the CpG odds ratio point of view. As first result, we got evidence of the statistical difference between the three groups of segmented genomes and identified the group of small genomic segments as the more informative, giving us the chance to reduce the research space. Considering the CpG odds ratio from the CDS regions, we obtained the confirmation that the CpG frequency is the lowest compared to the other dinucleotides and the Andes Hantavirus showed its highest CpG odds ratio in CDS. The analysis of correlation between the CpG odds ratio considering the full size of the segmented small genome and the CDS regions, resulted into a positive index and underlined the possible function of the CpG islands inside of the coding regions. The comparison between the CpG over the full genome, the CpG over the CDS and the median values over the ten viruses suggested a stronger concentration of the CpG islands both along the full-size genome and the CDS regions into the Andes virus. Using both the CpG odds ratio measurements (based on the CDS and on full genome size) from the group of small genomic segments as features, the unsupervised clustering analysis identified four different sub-groups inside of the *Orthohantaviridae* family and corroborated the evidence that the Andes Hantavirus (similar, in some way, to Tula H.) exhibits a peculiar CpG odds ratio distribution, perhaps linked to its unique prerogative to pass from human-to-human. Previous research already pointed out the huge variations of CpG bias in RNA viruses and brought out the observed under-representation of CpG in RNA viruses as not caused by the biased CpG usage in the non-coding regions but determined mainly by the coding regions [13]. In our study, through the calculation of the odds ratio for all the dinucleotides around into the CDS regions from 10 different RNA viruses from the *Hantaviridae* family, confirmed the CpG odds ratio into CDS as the lowest also for group of small genomic segments. Also, the examination of the correlation index between the distribution of CpG dinucleotides along the entire genomic segment and only the coding regions, confirmed what has already been observed in general for RNA viruses and highlighted the importance assumed by this dinucleotide in the case of orthohantavirus. The use of these indices as features for unsupervised clustering algorithms has highlighted how Andes H. and Tula H. somehow constitute “particular” cases within the family. A peculiarity linked to Andes H. could be its anthroponotic transmission capacity. The current study suggests that the prerogative of Andes H. to be transmitted from human to human could be linked to its distribution of CpG dinucleotides, or that in any case its frequency of CpG islands is such as to be identified as a cluster in its own right. In case of Tula orthohantavirus,

infections being only rarely found in humans [27-29] and even if (at the moment) there is no evidence to suggest a diversification of this virus from the rest of the family, it is questionable whether this similarity suggests a potential anthroponotic capacity in this virus. We can certainly assert that even in its case the distribution of CpG dinucleotides suggests greater attention. As a possible step forward in the research carried out, surely the use of further features related to the distribution of CpG dinucleotides as a relationship index with the CpG distribution of the host or with the distribution of the CpG islands in the regions internal to the codons and between the codons, could provide more detailed clustering results. The research carried out has already presented many important results, such as the significant statistical difference between the distributions of CpG dinucleotides in the different genomic segments (S, M and L), the identification of numerical indices useful for the application of unsupervised clustering algorithms and the identification of subgroups within the family of orthohantaviruses, including Andes H. and Tula H. as cases worthy of particular attention, especially in the case of Andes H. whose peculiar *anthroponicity* is particularly dangerous for humans.

Appendix

List of genomic sequences

Table 6 List of large RNA sequences

<i>HortoHantaVirus – Large RNA sequences</i>	
gb:KY659431	Organism:Andes orthohantavirus Strain Name:ANDV LS-CH-2016 Segment:L Host:Human
gb:JF920148	Organism:Dobrava-Belgrade orthohantavirus Strain Name:Ap/Sochi/hu Segment:L Host:Human
gb:MH251336	Organism:Dobrava-Belgrade orthohantavirus Strain Name:DOB-SOCHI Segment:L Host:Human
gb:MH251330	Organism:Hantaan orthohantavirus Strain Name:HTN-P88 Segment:L Host:Human
gb:KP896316	Organism:Hantaan orthohantavirus Strain Name:JS10 Segment:L Host:Human
gb:KP896317	Organism:Hantaan orthohantavirus Strain Name:JS11 Segment:L Host:Human
gb:KP896318	Organism:Hantaan orthohantavirus Strain Name:JS12 Segment:L Host:Human
gb:KP896314	Organism:Hantaan orthohantavirus Strain Name:JS8 Segment:L Host:Human
gb:KP896315	Organism:Hantaan orthohantavirus Strain Name:JS9 Segment:L Host:Human
gb:KU207198	Organism:Hantaan orthohantavirus Strain Name:ROKA13-8 Segment:L Host:Human
gb:KU207199	Organism:Hantaan orthohantavirus Strain Name:ROKA14-11 Segment:L Host:Human
gb:MH598466	Organism:Hantaan orthohantavirus Strain Name:ROKA16-9 Segment:L Host:Human
gb:MH598467	Organism:Hantaan orthohantavirus Strain Name:ROKA17-3 Segment:L Host:Human
gb:MH598468	Organism:Hantaan orthohantavirus Strain Name:ROKA17-5 Segment:L Host:Human
gb:MH598469	Organism:Hantaan orthohantavirus Strain Name:ROKA17-7 Segment:L Host:Human
gb:MH598470	Organism:Hantaan orthohantavirus Strain Name:ROKA17-8 Segment:L Host:Human
gb:MN608086	Organism:Hantaan orthohantavirus Strain Name:Tianmen1 Segment:L Host:Human
gb:MN608087	Organism:Hantaan orthohantavirus Strain Name:Tianmen15 Segment:L Host:Human
gb:MN608088	Organism:Hantaan orthohantavirus Strain Name:Tianmen35 Segment:L Host:Human
gb:MN608089	Organism:Hantaan orthohantavirus Strain Name:Tianmen39 Segment:L Host:Human
gb:MN608090	Organism:Hantaan orthohantavirus Strain Name:Tianmen51 Segment:L Host:Human
gb:MH251333	Organism:Puumala orthohantavirus Strain Name:PUU-TKD Segment:L Host:Human
gb:JN831952	Organism:Puumala orthohantavirus Strain Name:PUUV/Pieksamaki/human_kidney/2008 Segment:L Host:Human
gb:JN831949	Organism:Puumala orthohantavirus Strain Name:PUUV/Pieksamaki/human_lung/2008 Segment:L Host:Human
gb:MF149951	Organism:Seoul orthohantavirus Strain Name:Hu02-258/NGS Segment:L Subtype:Seoul Host:Human
gb:L37901	Organism:Sin Nombre orthohantavirus Strain Name:NM H10 Segment:L Host:Human
gb:NC_005217	Organism:Sin Nombre orthohantavirus Strain Name:NM H10 Segment:L Host:Human

Table 7 List of medium RNA sequences

<i>HortoHantaVirus – Medium RNA sequences</i>	
gb:AY228238	Organism:Andes orthohantavirus Strain Name:CHI-7913 Segment:M Host:Human
gb:KY604962	Organism:Andes orthohantavirus Strain Name:LS-CH2016 Segment:M Host:Human
gb:L36930	Organism:Bayou orthohantavirus Strain Name:UNKNOWN-L36930 Segment:M Host:Human
gb:NC_038300	Organism:Bayou orthohantavirus Strain Name:UNKNOWN-NC_038300 Segment:M Host:Human
gb:JF920149	Organism:Dobrava-Belgrade orthohantavirus Strain Name:Ap/Sochi/hu Segment:M Host:Human
gb:MH251335	Organism:Dobrava-Belgrade orthohantavirus Strain Name:DOB-SOCHI Segment:M Host:Human
gb:MH251329	Organism:Hantaan orthohantavirus Strain Name:HTN-P88 Segment:M Host:Human
gb:JQ665881	Organism:Hantaan orthohantavirus Strain Name:HubeiHu02 Segment:M Host:Human
gb:KP970569	Organism:Hantaan orthohantavirus Strain Name:JS10 Segment:M Host:Human

gb:KP970570 Organism:Hantaan orthohantavirus Strain Name:JS11 Segment:M Host:Human
gb:KP970571 Organism:Hantaan orthohantavirus Strain Name:JS12 Segment:M Host:Human
gb:KP970567 Organism:Hantaan orthohantavirus Strain Name:JS8 Segment:M Host:Human
gb:KP970568 Organism:Hantaan orthohantavirus Strain Name:JS9 Segment:M Host:Human
gb:KU207202 Organism:Hantaan orthohantavirus Strain Name:ROKA13-8 Segment:M Host:Human
gb:KU207203 Organism:Hantaan orthohantavirus Strain Name:ROKA14-11 Segment:M Host:Human
gb:MH598480 Organism:Hantaan orthohantavirus Strain Name:ROKA16-9 Segment:M Host:Human
gb:MH598481 Organism:Hantaan orthohantavirus Strain Name:ROKA17-3 Segment:M Host:Human
gb:MH598482 Organism:Hantaan orthohantavirus Strain Name:ROKA17-5 Segment:M Host:Human
gb:MH598483 Organism:Hantaan orthohantavirus Strain Name:ROKA17-7 Segment:M Host:Human
gb:MH598484 Organism:Hantaan orthohantavirus Strain Name:ROKA17-8 Segment:M Host:Human
gb:MN608075 Organism:Hantaan orthohantavirus Strain Name:Tianmen1 Segment:M Host:Human
gb:MN608076 Organism:Hantaan orthohantavirus Strain Name:Tianmen15 Segment:M Host:Human
gb:MN608077 Organism:Hantaan orthohantavirus Strain Name:Tianmen35 Segment:M Host:Human
gb:MN608078 Organism:Hantaan orthohantavirus Strain Name:Tianmen39 Segment:M Host:Human
gb:MN608079 Organism:Hantaan orthohantavirus Strain Name:Tianmen51 Segment:M Host:Human
gb:KU207204 Organism:Hantaan orthohantavirus Strain Name:US8A14-2 Segment:M Host:Human
gb:KU207205 Organism:Hantaan orthohantavirus Strain Name:US8A15-1 Segment:M Host:Human
gb:EU092222 Organism:Hantaanvirus CGHu1 Strain Name:CGHu1 Segment:M Host:Human
gb:EU363819 Organism:Hantaanvirus CGHu2 Strain Name:CGHu2 Segment:M Host:Human
gb:EU363818 Organism:Hantaanvirus CGHu3 Strain Name:CGHu3 Segment:M Host:Human
gb:EF990923 Organism:Hantaanvirus CGHu3612 Strain Name:CGHu3612 Segment:M Host:Human
gb:EF990922 Organism:Hantaanvirus CGHu3614 Strain Name:CGHu3614 Segment:M Host:Human
gb:MK496163 Organism:Puumala orthohantavirus Strain Name:H46/Ufa Segment:M Host:Human
gb:MK496160 Organism:Puumala orthohantavirus Strain Name:P-360 Segment:M Host:Human
gb:MH251332 Organism:Puumala orthohantavirus Strain Name:PUU-TKD Segment:M Host:Human
gb:JN831951 Organism:Puumala orthohantavirus Strain Name:PUUV/Pieksamaki/human_kidney/2008 Segment:M Host:Human
gb:JN831948 Organism:Puumala orthohantavirus Strain Name:PUUV/Pieksamaki/human_lung/2008 Segment:M Host:Human
gb:MF149946 Organism:Seoul orthohantavirus Strain Name:Hu02-258/NGS Segment:M Subtype:Seoul Host:Human
gb:NC_005215 Organism:Sin Nombre orthohantavirus Strain Name:NM H10 Segment:M Host:Human

Table 8 List of small RNA sequences

<i>HortoHantaVirus – Small RNA sequences</i>
gb:KY659432 Organism:Andes orthohantavirus Strain Name:ANDV LS-CH-2016 ex Chile Segment:S Host:Human
gb:AY228237 Organism:Andes orthohantavirus Strain Name:CHI-7913 Segment:S Host:Human
gb:JF750419 Organism:Tunari virus Strain Name:FVB554 Segment:S Host:Human
gb:JF750418 Organism:Tunari virus Strain Name:FVB640 Segment:S Host:Human
gb:JF750417 Organism:Tunari virus Strain Name:FVB799 Segment:S Host:Human
gb:L36929 Organism:Bayou orthohantavirus Strain Name:UNKNOWN-L36929 Segment:S Host:Human
gb:NC_038298 Organism:Bayou orthohantavirus Strain Name:UNKNOWN-NC_038298 Segment:S Host:Human
gb:KM597161 Organism:Choclo virus Strain Name:Uk (ex Panama) Segment:S Host:Human
gb:KP878313 Organism:Dobrava-Belgrade orthohantavirus Strain Name:10752/hu Segment:S Host:Human
gb:JF920150 Organism:Dobrava-Belgrade orthohantavirus Strain Name:Ap/Sochi/hu Segment:S Host:Human
gb:MH251334 Organism:Dobrava-Belgrade orthohantavirus Strain Name:DOB-SOCHI Segment:S Host:Human

gb:KC570384 Organism:Hantaan orthohantavirus Strain Name:DandongHu-22 Segment:S Host:Human
gb:KC570385 Organism:Hantaan orthohantavirus Strain Name:DandongHu-28 Segment:S Host:Human
gb:KC570386 Organism:Hantaan orthohantavirus Strain Name:DandongHu-32 Segment:S Host:Human
gb:KC570387 Organism:Hantaan orthohantavirus Strain Name:DandongHu-34 Segment:S Host:Human
gb:KC570388 Organism:Hantaan orthohantavirus Strain Name:DandongHu-44 Segment:S Host:Human
gb:KC570389 Organism:Hantaan orthohantavirus Strain Name:DandongHu-89 Segment:S Host:Human
gb:KC570390 Organism:Hantaan orthohantavirus Strain Name:DandongHu-91 Segment:S Host:Human
gb:MH251328 Organism:Hantaan orthohantavirus Strain Name:HTN-P88 Segment:S Host:Human
gb:MN478382 Organism:Hantaan orthohantavirus Strain Name:HTNV-HN2017/70 Segment:S Host:Human
gb:MN478383 Organism:Hantaan orthohantavirus Strain Name:HTNV-HN2017/76 Segment:S Host:Human
gb:MN478384 Organism:Hantaan orthohantavirus Strain Name:HTNV-HN2017/79 Segment:S Host:Human
gb:MN478385 Organism:Hantaan orthohantavirus Strain Name:HTNV-HN2017/80 Segment:S Host:Human
gb:MN478386 Organism:Hantaan orthohantavirus Strain Name:HTNV-HN2017/81 Segment:S Host:Human
gb:MN478387 Organism:Hantaan orthohantavirus Strain Name:HTNV-HN2017/82 Segment:S Host:Human
gb:MN478388 Organism:Hantaan orthohantavirus Strain Name:HTNV-HN2017/87 Segment:S Host:Human
gb:MN478389 Organism:Hantaan orthohantavirus Strain Name:HTNV-HN2018/106 Segment:S Host:Human
gb:MN478390 Organism:Hantaan orthohantavirus Strain Name:HTNV-HN2018/131 Segment:S Host:Human
gb:MN478391 Organism:Hantaan orthohantavirus Strain Name:HTNV-HN2018/134 Segment:S Host:Human
gb:MN478392 Organism:Hantaan orthohantavirus Strain Name:HTNV-HN2018/138 Segment:S Host:Human
gb:MN478393 Organism:Hantaan orthohantavirus Strain Name:HTNV-HN2018/146 Segment:S Host:Human
gb:MN478394 Organism:Hantaan orthohantavirus Strain Name:HTNV-HN2018/150 Segment:S Host:Human
gb:MN478395 Organism:Hantaan orthohantavirus Strain Name:HTNV-HN2018/152 Segment:S Host:Human
gb:MN478396 Organism:Hantaan orthohantavirus Strain Name:HTNV-HN2018/154 Segment:S Host:Human
gb:MN478397 Organism:Hantaan orthohantavirus Strain Name:HTNV-HN2018/157 Segment:S Host:Human
gb:JQ665905 Organism:Hantaan orthohantavirus Strain Name:HubeiHu02 Segment:S Host:Human
gb:KP970581 Organism:Hantaan orthohantavirus Strain Name:JS10 Segment:S Host:Human
gb:KP970582 Organism:Hantaan orthohantavirus Strain Name:JS11 Segment:S Host:Human
gb:KP970583 Organism:Hantaan orthohantavirus Strain Name:JS12 Segment:S Host:Human
gb:KP970579 Organism:Hantaan orthohantavirus Strain Name:JS8 Segment:S Host:Human
gb:KP970580 Organism:Hantaan orthohantavirus Strain Name:JS9 Segment:S Host:Human
gb:KY283955 Organism:Hantaan orthohantavirus Strain Name:MN2009P-M3 Segment:S Host:Human
gb:KY283956 Organism:Hantaan orthohantavirus Strain Name:MN2009P-M6 Segment:S Host:Human
gb:KU207206 Organism:Hantaan orthohantavirus Strain Name:ROKA13-8 Segment:S Host:Human
gb:KU207207 Organism:Hantaan orthohantavirus Strain Name:ROKA14-11 Segment:S Host:Human
gb:MH598494 Organism:Hantaan orthohantavirus Strain Name:ROKA16-9 Segment:S Host:Human
gb:MH598495 Organism:Hantaan orthohantavirus Strain Name:ROKA17-3 Segment:S Host:Human
gb:MH598496 Organism:Hantaan orthohantavirus Strain Name:ROKA17-5 Segment:S Host:Human
gb:MH598497 Organism:Hantaan orthohantavirus Strain Name:ROKA17-7 Segment:S Host:Human
gb:MH598498 Organism:Hantaan orthohantavirus Strain Name:ROKA17-8 Segment:S Host:Human
gb:KC844226 Organism:Hantaan orthohantavirus Strain Name:SXHu2012B1 Segment:S Host:Human
gb:KC844227 Organism:Hantaan orthohantavirus Strain Name:SXHu2012B3 Segment:S Host:Human
gb:MN608064 Organism:Hantaan orthohantavirus Strain Name:Tianmen1 Segment:S Host:Human
gb:MN608065 Organism:Hantaan orthohantavirus Strain Name:Tianmen15 Segment:S Host:Human
gb:MN608066 Organism:Hantaan orthohantavirus Strain Name:Tianmen35 Segment:S Host:Human
gb:MN608067 Organism:Hantaan orthohantavirus Strain Name:Tianmen39 Segment:S Host:Human

gb:MN608068 Organism:Hantaan orthohantavirus Strain Name:Tianmen51 Segment:S Host:Human
gb:KU207208 Organism:Hantaan orthohantavirus Strain Name:US8A14-2 Segment:S Host:Human
gb:KU207209 Organism:Hantaan orthohantavirus Strain Name:US8A15-1 Segment:S Host:Human
gb:KM355414 Organism:Hantaan orthohantavirus Strain Name:WCL Segment:S Host:Human
gb:KY357324 Organism:Hantaan orthohantavirus Strain Name:XA2009P-M18 Segment:S Host:Human
gb:KY357325 Organism:Hantaan orthohantavirus Strain Name:XA2011P-Z21 Segment:S Host:Human
gb:KY357323 Organism:Hantaan orthohantavirus Strain Name:XA2012P-Z22 Segment:S Host:Human
gb:KY357326 Organism:Hantaan orthohantavirus Strain Name:XA2012P133 Segment:S Host:Human
gb:KY357327 Organism:Hantaan orthohantavirus Strain Name:XA2012P148 Segment:S Host:Human
gb:KY357322 Organism:Hantaan orthohantavirus Strain Name:XA2012P160 Segment:S Host:Human
gb:HQ834507 Organism:Hantaan virus P09072 Strain Name:P09072 Segment:S Host:Human
gb:EU092218 Organism:Hantaanvirus CGHu1 Strain Name:CGHu1 Segment:S Host:Human
gb:EU363813 Organism:Hantaanvirus CGHu2 Strain Name:CGHu2 Segment:S Host:Human
gb:EU363809 Organism:Hantaanvirus CGHu3 Strain Name:CGHu3 Segment:S Host:Human
gb:EF990909 Organism:Hantaanvirus CGHu3612 Strain Name:CGHu3612 Segment:S Host:Human
gb:EF990908 Organism:Hantaanvirus CGHu3614 Strain Name:CGHu3614 Segment:S Host:Human
gb:MG923671 Organism:Puumala orthohantavirus Strain Name:AISNE-02/Hu/FRA/2016.00467 Segment:S Host:Human
gb:MG923604 Organism:Puumala orthohantavirus Strain Name:ALFORTVILLE-94/Hu/FRA/2015.00456 Segment:S Host:Human
gb:MG923608 Organism:Puumala orthohantavirus Strain Name:ANGIREY-70/Hu/FRA/2015.00410 Segment:S Host:Human
gb:MG923656 Organism:Puumala orthohantavirus Strain Name:ANOR-59/Hu/FRA/2015.00422 Segment:S Host:Human
gb:MG923652 Organism:Puumala orthohantavirus Strain Name:ARBOIS-39/Hu/FRA/2014.00622 Segment:S Host:Human
gb:MG923647 Organism:Puumala orthohantavirus Strain Name:ATHIES-SOUS-LAON-02/Hu/FRA/2014.00135 Segment:S Host:Human
gb:MG923665 Organism:Puumala orthohantavirus Strain Name:AULNOYE-AYMERIES-59/Hu/FRA/2016.00325 Segment:S Host:Human
gb:MG923605 Organism:Puumala orthohantavirus Strain Name:BAR-LE-DUC-55/Hu/FRA/2012.00123 Segment:S Host:Human
gb:MG923627 Organism:Puumala orthohantavirus Strain Name:BOGNY-SUR-MEUSE-08/Hu/FRA/2015.00329 Segment:S Host:Human
gb:MG923660 Organism:Puumala orthohantavirus Strain Name:BOULZICOURT-08/Hu/FRA/2016.00182 Segment:S Host:Human
gb:MG923618 Organism:Puumala orthohantavirus Strain Name:BUIRONFOSSE-02/Hu/FRA/2014.00153 Segment:S Host:Human
gb:MG923640 Organism:Puumala orthohantavirus Strain Name:CESSIERES-02/Hu/FRA/2016.00353 Segment:S Host:Human
gb:MG923623 Organism:Puumala orthohantavirus Strain Name:CHAMBLY-60/Hu/FRA/2014.00540 Segment:S Host:Human
gb:MG923600 Organism:Puumala orthohantavirus Strain Name:CHAMPIGNY-SUR-MARNE-94/Hu/FRA/2014.00499 Segment:S Host:Human
gb:MG923654 Organism:Puumala orthohantavirus Strain Name:CHARLEVILLE-MEZIERES-08/Hu/FRA/2015.00402 Segment:S Host:Human
gb:MG923611 Organism:Puumala orthohantavirus Strain Name:CHEVROCHES-58/Hu/FRA/2012.00086 Segment:S Host:Human
gb:MG923631 Organism:Puumala orthohantavirus Strain Name:CILLY-02/Hu/FRA/2015.00657 Segment:S Host:Human
gb:MG923606 Organism:Puumala orthohantavirus Strain Name:COISERETTE-39/Hu/FRA/2012.00102 Segment:S Host:Human
gb:MG923612 Organism:Puumala orthohantavirus Strain Name:COLOMBEY-LES-BELLES-54/Hu/FRA/2012.00307 Segment:S Host:Human
gb:MG923663 Organism:Puumala orthohantavirus Strain Name:CORNY-MACHEROMENIL-08/Hu/FRA/2016.00295 Segment:S Host:Human
gb:MG923641 Organism:Puumala orthohantavirus Strain Name:COUSOLRE-59/Hu/FRA/2012.00057 Segment:S Host:Human
gb:MG923655 Organism:Puumala orthohantavirus Strain Name:DOUZY-08/Hu/FRA/2015.00419 Segment:S Host:Human
gb:MG923644 Organism:Puumala orthohantavirus Strain Name:ENGLANCOURT-02/Hu/FRA/2012.00349 Segment:S Host:Human
gb:MG923624 Organism:Puumala orthohantavirus Strain Name:ETEIGNIERES-08/Hu/FRA/2015.00019 Segment:S Host:Human
gb:MG923626 Organism:Puumala orthohantavirus Strain Name:FELLERING-68/Hu/FRA/2015.00185 Segment:S Host:Human

gb:MG923649 Organism:Puumala orthohantavirus Strain Name:FOURMIES-59/Hu/FRA/2014.00184 Segment:S Host:Human
gb:MG923650 Organism:Puumala orthohantavirus Strain Name:FOURMIES-59/Hu/FRA/2014.00233 Segment:S Host:Human
gb:MG923622 Organism:Puumala orthohantavirus Strain Name:FOURMIES-59/Hu/FRA/2014.00321 Segment:S Host:Human
gb:MG923601 Organism:Puumala orthohantavirus Strain Name:FOURMIES-59/Hu/FRA/2014.00598 Segment:S Host:Human
gb:MG923651 Organism:Puumala orthohantavirus Strain Name:FOURMIES-59/Hu/FRA/2014.00613 Segment:S Host:Human
gb:MG923625 Organism:Puumala orthohantavirus Strain Name:FOURMIES-59/Hu/FRA/2015.00045 Segment:S Host:Human
gb:MG923666 Organism:Puumala orthohantavirus Strain Name:FOURMIES-59/Hu/FRA/2016.00333 Segment:S Host:Human
gb:MG923667 Organism:Puumala orthohantavirus Strain Name:FOURMIES-59/Hu/FRA/2016.00345 Segment:S Host:Human
gb:MG923669 Organism:Puumala orthohantavirus Strain Name:FOURMIES-59/Hu/FRA/2016.00427 Segment:S Host:Human
gb:MG923615 Organism:Puumala orthohantavirus Strain Name:GIVET-08/Hu/FRA/2012.00638 Segment:S Host:Human
gb:MG923614 Organism:Puumala orthohantavirus Strain Name:GOUVIEUX-60/Hu/FRA/2012.00402 Segment:S Host:Human
gb:MG923653 Organism:Puumala orthohantavirus Strain Name:GREZY-SUR-ISERE-73/Hu/FRA/2015.00153 Segment:S Host:Human
gb:MK496162 Organism:Puumala orthohantavirus Strain Name:H46/Ufa Segment:S Host:Human
gb:MG923668 Organism:Puumala orthohantavirus Strain Name:HIRSON-02/Hu/FRA/2016.00357 Segment:S Host:Human
gb:MG923633 Organism:Puumala orthohantavirus Strain Name:JALLANGES-21/Hu/FRA/2016.00275 Segment:S Host:Human
gb:MG923635 Organism:Puumala orthohantavirus Strain Name:LA-NEUVILLE-SUR-RESSONS-60/Hu/FRA/2016.00293 Segment:S Host:Human
gb:MG923645 Organism:Puumala orthohantavirus Strain Name:LA-PESSE-39/Hu/FRA/2012.00536 Segment:S Host:Human
gb:MG923609 Organism:Puumala orthohantavirus Strain Name:LANISCOURT-02/Hu/FRA/2012.00061 Segment:S Host:Human
gb:MG923636 Organism:Puumala orthohantavirus Strain Name:LAON-02/Hu/FRA/2016.00311 Segment:S Host:Human
gb:MG923639 Organism:Puumala orthohantavirus Strain Name:LAON-02/Hu/FRA/2016.00326 Segment:S Host:Human
gb:MG923670 Organism:Puumala orthohantavirus Strain Name:LAON-02/Hu/FRA/2016.00452 Segment:S Host:Human
gb:MG923607 Organism:Puumala orthohantavirus Strain Name:LE-MOUTARET-38/Hu/FRA/2014.00120 Segment:S Host:Human
gb:MG923621 Organism:Puumala orthohantavirus Strain Name:LILLE-59/Hu/FRA/2014.00276 Segment:S Host:Human
gb:MG923628 Organism:Puumala orthohantavirus Strain Name:MONTCORNET-02/Hu/FRA/2015.00430 Segment:S Host:Human
gb:MG923630 Organism:Puumala orthohantavirus Strain Name:MONTHERME-08/Hu/FRA/2015.00526 Segment:S Host:Human
gb:MG923634 Organism:Puumala orthohantavirus Strain Name:MORBECQUE-59/Hu/FRA/2016.00282 Segment:S Host:Human
gb:MG923610 Organism:Puumala orthohantavirus Strain Name:MOUTHE-25/Hu/FRA/2012.00301 Segment:S Host:Human
gb:MK496159 Organism:Puumala orthohantavirus Strain Name:P-360 Segment:S Host:Human
gb:MG923672 Organism:Puumala orthohantavirus Strain Name:PREMONTRE-02/Hu/FRA/2016.00469 Segment:S Host:Human
gb:MG923661 Organism:Puumala orthohantavirus Strain Name:PRESLES-ET-THIERNY-02/Hu/FRA/2016.00268 Segment:S Host:Human
gb:MH251331 Organism:Puumala orthohantavirus Strain Name:PUU-TKD Segment:S Host:Human
gb:JN831950 Organism:Puumala orthohantavirus Strain Name:PUUV/Pieksamaki/human_kidney/2008 Segment:S Host:Human
gb:JN831947 Organism:Puumala orthohantavirus Strain Name:PUUV/Pieksamaki/human_lung/2008 Segment:S Host:Human
gb:MG923643 Organism:Puumala orthohantavirus Strain Name:REIMS-51/Hu/FRA/2012.00278 Segment:S Host:Human
gb:MG923674 Organism:Puumala orthohantavirus Strain Name:REIMS-51/Hu/FRA/2015.00665 Segment:S Host:Human
gb:MG923658 Organism:Puumala orthohantavirus Strain Name:REMILLY-AILLICOURT-08/Hu/FRA/2015.00498 Segment:S Host:Human
gb:MG923629 Organism:Puumala orthohantavirus Strain Name:REVIGNY-SUR-ORNAIN-55/Hu/FRA/2015.00457 Segment:S Host:Human
gb:MG923598 Organism:Puumala orthohantavirus Strain Name:RIOZ-70/Hu/FRA/2015.00567 Segment:S Host:Human
gb:MG923673 Organism:Puumala orthohantavirus Strain Name:ROCROI-08/Hu/FRA/2012.00018 Segment:S Host:Human
gb:MG923638 Organism:Puumala orthohantavirus Strain Name:RONCHAMP-70/Hu/FRA/2015.00504 Segment:S Host:Human
gb:MG923613 Organism:Puumala orthohantavirus Strain Name:SAINT-CLAUDE-39/Hu/FRA/2012.00396 Segment:S Host:Human
gb:MG923646 Organism:Puumala orthohantavirus Strain Name:SAINT-MICHEL-02/Hu/FRA/2014.00097 Segment:S Host:Human

gb:MG923619 Organism:Puumala orthohantavirus Strain Name:SAINT-SAULVE-59/Hu/FRA/2014.00171 Segment:S Host:Human
gb:MG923637 Organism:Puumala orthohantavirus Strain Name:SAINT-VIT-25/Hu/FRA/2016.00320 Segment:S Host:Human
gb:MG923603 Organism:Puumala orthohantavirus Strain Name:SAINTE-MENEHOULD-51/Hu/FRA/2012.00025 Segment:S Host:Human
gb:MG923659 Organism:Puumala orthohantavirus Strain Name:SAULES-25/Hu/FRA/2014.00637 Segment:S Host:Human
gb:MG923617 Organism:Puumala orthohantavirus Strain Name:SECHEVAL-08/Hu/FRA/2014.00053 Segment:S Host:Human
gb:MG923657 Organism:Puumala orthohantavirus Strain Name:SEDAN-08/Hu/FRA/2015.00488 Segment:S Host:Human
gb:MG923642 Organism:Puumala orthohantavirus Strain Name:SIGNY-LE-PETIT-08/Hu/FRA/2014.00488 Segment:S Host:Human
gb:MG923648 Organism:Puumala orthohantavirus Strain Name:ST-ERME-OUTRE-ET-RAMECOURT-02/Hu/FRA/2014.00174 Segment:S Host:Human
gb:MG923664 Organism:Puumala orthohantavirus Strain Name:THIN-LE-MOUTIER-08/Hu/FRA/2016.00310 Segment:S Host:Human
gb:MG923620 Organism:Puumala orthohantavirus Strain Name:TREMBLOIS-LES-ROCROI-08/Hu/FRA/2014.00209 Segment:S Host:Human
gb:MG923662 Organism:Puumala orthohantavirus Strain Name:TRUCY-02/Hu/FRA/2016.00286 Segment:S Host:Human
gb:MG923616 Organism:Puumala orthohantavirus Strain Name:VENDIN-LES-BETHUNE-62/Hu/FRA/2013.00250 Segment:S Host:Human
gb:MG923599 Organism:Puumala orthohantavirus Strain Name:VIC-SUR-AISNE-02/Hu/FRA/2015.00660 Segment:S Host:Human
gb:MG923632 Organism:Puumala orthohantavirus Strain Name:VIREUX-MOLHAIN-08/Hu/FRA/2016.00239 Segment:S Host:Human
gb:MG923602 Organism:Puumala orthohantavirus Strain Name:VRIGNE-MEUSE-08/Hu/FRA/2015.00328 Segment:S Host:Human
gb:GQ279395 Organism:Seoul orthohantavirus Strain Name:CUI Segment:S Host:Human
gb:KX064275 Organism:Seoul orthohantavirus Strain Name:ERIZE-ST-DIZIER/Hu/FRA/2014/2014.00479 Segment:S Host:Human
gb:MF149954 Organism:Seoul orthohantavirus Strain Name:Hu02-258/NGS Segment:S Subtype:Seoul Host:Human
gb:MF149955 Organism:Seoul orthohantavirus Strain Name:Hu02-294/NGS Segment:S Subtype:Seoul Host:Human
gb:MF149956 Organism:Seoul orthohantavirus Strain Name:Hu02-529/NGS Segment:S Subtype:Seoul Host:Human
gb:GQ279390 Organism:Seoul orthohantavirus Strain Name:HuBJ15 Segment:S Host:Human
gb:GQ279380 Organism:Seoul orthohantavirus Strain Name:HuBJ16 Segment:S Host:Human
gb:GQ279389 Organism:Seoul orthohantavirus Strain Name:HuBJ19 Segment:S Host:Human
gb:GQ279394 Organism:Seoul orthohantavirus Strain Name:HuBJ20 Segment:S Host:Human
gb:GQ279379 Organism:Seoul orthohantavirus Strain Name:HuBJ22 Segment:S Host:Human
gb:GQ279391 Organism:Seoul orthohantavirus Strain Name:HuBJ3 Segment:S Host:Human
gb:GQ279381 Organism:Seoul orthohantavirus Strain Name:HuBJ7 Segment:S Host:Human
gb:GQ279384 Organism:Seoul orthohantavirus Strain Name:HuBJ9 Segment:S Host:Human
gb:KC902522 Organism:Seoul orthohantavirus Strain Name:REPLONGES/Hu/FRA/2012/12-0882 Segment:S Host:Human
gb:KX064270 Organism:Seoul orthohantavirus Strain Name:TURCKHEIM/Hu/FRA/2016/2016.00044 Segment:S Host:Human
gb:KT946591 Organism:Tula orthohantavirus Strain Name:CHEVRU/Hu/FRA/2015/15.00453 Segment:S Host:Human

List of R scripts

```
#Loading the libraries
library("gmodels")
library("car")
library("ggplot2")
library("ggplotr")
library("dplyr")
library("emmeans")
library("FSA")

#set working path
setwd("Documents/Research/Hantavirus/Anova-OneWay/")

#load data
dat<-read.csv("CpG_Values.csv")

#Designate Group as a categorical factor
dat$Group<-as.factor(dat$Group)

#Produce descriptive statistics by treatment
dat %>% select(CpG, Group) %>% group_by(Group) %>%
  summarise(n = n(),
            mean = mean(CpG, na.rm = TRUE),
            sd = sd(CpG, na.rm = TRUE),
            stderr = sd/sqrt(n),
            LCL = mean - qt(1 - (0.05 / 2), n - 1) * stderr,
            UCL = mean + qt(1 - (0.05 / 2), n - 1) * stderr,
            median = median(CpG, na.rm = TRUE),
            min = min(CpG, na.rm = TRUE),
            max = max(CpG, na.rm = TRUE),
            IQR = IQR(CpG, na.rm = TRUE))

#Perform the Shapiro-Wilk Test for Normality on each group
dat %>%
  group_by(Group) %>%
  summarise(`W Stat` = shapiro.test(CpG)$statistic,
            `p-value` = shapiro.test(CpG)$p.value)

#Perform QQ plots by group
ggplot(data = dat, mapping = aes(sample = CpG, color = Group, fill = Group)) +
  stat_qq_band(alpha=0.5, conf=0.95, qtype=1, bandType = "boot", B=5000) +
  stat_qq_line(identity=TRUE) +
  stat_qq_point(col="black") +
  facet_wrap(~ Group, scales = "free") +
  labs(x = "Theoretical Quantiles", y = "Sample Quantiles") + theme_bw()

#Perform Levene's Test of Equality of Variances
lev1<-leveneTest(CpG ~ Group, data=dat, center="mean")
lev2<-leveneTest(CpG ~ Group, data=dat, center="median")
print(lev1)

#Produce boxplots and visually check for outliers
ggplot(dat, aes(x = Group, y = CpG, fill = Group)) +
  stat_boxplot(geom = "errorbar", width = 0.5) +
  geom_boxplot(fill = "light blue") +
  stat_summary(fun.y=mean, geom="point", shape=10, size=3.5, color="black") +
  ggtitle("Boxplots of CpG odds ratio for each group") +
  theme_bw() + theme(legend.position="none")

#Perform the Kruskal-Wallis test
m1<-kruskal.test(CpG ~ Group, data=dat)

#Dunn's Kruskal-Wallis post-hoc test
posthocsl<-dunnTest(CpG ~ Group, data=dat, method="holm")
print(posthocsl)

library(rcompanion)
PT = posthocsl$res
cldList(P.adj ~ Comparison,
        data = PT,
        threshold = 0.05)

library(tidyverse)
library(ggpubr)
library(rstatix)

pwc <- dunn_test(CpG~Group, data=dat, p.adjust.method = "bonferroni")
pwc <- pwc %>% add_xy_position(x = "group")
res.kruskal <- dat %>% kruskal_test(CpG ~ Group)

ggboxplot(dat, x = "Group", y = "CpG", color = "Group", add = "jitter") +
  stat_pvalue_manual(pwc, hide.ns = TRUE) +
  labs(
    subtitle = get_test_label(res.kruskal, detailed = TRUE),
    caption = get_pwc_label(pwc)
  )

library(dunn.test)
dunn.test(dat$CpG, dat$Group, "bonferroni", list=TRUE)
```

Figure 24 Script to conduct ANOVA analysis in R

```

# Import data
CpG_data <- read.csv(
file = "data_CpG.csv",
sep = ",", dec = ".", header = TRUE, row.names = 1
)
head(CpG_data)
library(factoextra)
library(NbClust)

# Elbow method
fviz_nbclust(CpG_data, kmeans, method = "wss", k.max = 9) +
geom_vline(xintercept = 4, linetype = 2) + # add line for better visualisation
labs(subtitle = "Elbow method") # add subtitle

# Silhouette method
fviz_nbclust(CpG_data, kmeans, method = "silhouette", k.max = 9) +
labs(subtitle = "Silhouette method")

# Gap statistic
set.seed(42)
fviz_nbclust(CpG_data, kmeans,
nstart = 25,
method = "gap_stat",
nboot = 500, k.max = 9
) + # reduce it for lower computation time (but less precise results)
labs(subtitle = "Gap statistic method")
library(clustree)
tmp <- NULL
for (k in 1:9){
tmp[k] <- kmeans(CpG_data, k, nstart = 30)
}
df <- data.frame(tmp)

# add a prefix to the column names
colnames(df) <- seq(1:9)
library(dplyr)
colnames(df) <- paste0("k",colnames(df))

# get individual PCA
df.pca <- prcomp(df, center = TRUE, scale. = FALSE)
ind.coord <- df.pca$x
ind.coord <- ind.coord[,1:2]
df <- bind_cols(as.data.frame(df), as.data.frame(ind.coord))
png(filename="clustree.png", width = 1024, height = 768)
clustree(df, prefix = "k")
dev.off()

#Kmeans k=4
km_res <- kmeans(CpG_data, centers = 4, nstart = 20)
png(filename="Kmeans_K4.png", width = 1024, height = 768)
fviz_cluster(km_res, CpG_data)
dev.off()

#DBSCAN
library("fpc")
# Compute DBSCAN using fpc package
set.seed(444)
db <- fpc::dbscan(CpG_data, eps = 0.15, MinPts = 3, method = "dist", scale = TRUE)
# Plot DBSCAN results
png(filename="DBSCAN.png", width = 1024, height = 768)
plot(db, CpG_data, main = "DBSCAN", frame = TRUE)
dev.off()
fviz_cluster(db, CpG_data, stand = FALSE, frame = FALSE, geom = "point")

##HCA
##Agglomerative
#Ward's method gets us the highest agglomerative coefficient. Let us look at its dendrogram.
hc3 <- agnes(CpG_data, method = "ward")
png(filename="HCA Agglomerative-AGNES.png", width = 1024, height = 768)
pltree(hc3, cex = 2, hang = -1, main = "Dendrogram of agnes")
dev.off()
# Dissimilarity matrix
d <- dist(CpG_data, method = "euclidean")

# Hierarchical clustering using Complete Linkage
hc1 <- hclust(d, method = "complete" )
plot(hc1, cex = 0.6, hang = -1)

##Divisiveive
hc4 <- diana(CpG_data)
png(filename="HCA Divisive-AGNES.png", width = 1024, height = 768)
pltree(hc4, cex = 2, hang = -1, main = "Dendrogram of diana")
dev.off()

#Visualize cluster from HCA
clust <- cutree(hc4, k = 4)
png(filename="HCA Clustering_K4.png", width = 1024, height = 768)
fviz_cluster(list(data = CpG_data, cluster = clust))
dev.off()

```

Figure 25 Script to conduct the unsupervised clustering in R

REFERENCES

- [1] P. Kaukinen, A. Vaheri, and A. Plyusnin, "Hantavirus nucleocapsid protein: a multifunctional molecule with both housekeeping and ambassadorial duties," *Arch Virol*, vol. 150, no. 9, pp. 1693-713, Sep 2005.10.1007/s00705-005-0555-4
- [2] E. S. Travassos da Rosa *et al.*, "Pygmy rice rat as potential host of Castelo dos Sonhos Hantavirus," *Emerg Infect Dis*, vol. 17, no. 8, pp. 1527-30, Aug 2011.10.3201/eid1708.101547
- [3] D. B. Medeiros *et al.*, "Circulation of hantaviruses in the influence area of the Cuiaba-Santarem Highway," *Mem Inst Oswaldo Cruz*, vol. 105, no. 5, pp. 665-71, Aug 2010.10.1590/s0074-02762010000500011
- [4] B. Knust, A. Macneil, and P. E. Rollin, "Hantavirus pulmonary syndrome clinical findings: evaluating a surveillance case definition," *Vector Borne Zoonotic Dis*, vol. 12, no. 5, pp. 393-9, May 2012.10.1089/vbz.2011.0764
- [5] B. Hjelle and F. Torres-Perez, "Hantaviruses in the americas and their role as emerging pathogens," *Viruses*, vol. 2, no. 12, pp. 2559-86, Dec 2010.10.3390/v2122559
- [6] C. B. Jonsson, L. T. Figueiredo, and O. Vapalahti, "A global perspective on hantavirus ecology, epidemiology, and disease," *Clin Microbiol Rev*, vol. 23, no. 2, pp. 412-41, Apr 2010.10.1128/CMR.00062-09
- [7] D. C. Watson, M. Sargianou, A. Papa, P. Chra, I. Starakis, and G. Panos, "Epidemiology of Hantavirus infections in humans: a comprehensive, global overview," *Crit Rev Microbiol*, vol. 40, no. 3, pp. 261-72, Aug 2014.10.3109/1040841X.2013.783555
- [8] M. Ferres *et al.*, "Prospective evaluation of household contacts of persons with hantavirus cardiopulmonary syndrome in chile," *J Infect Dis*, vol. 195, no. 11, pp. 1563-71, Jun 1 2007.10.1086/516786
- [9] P. J. Padula, A. Edelstein, S. D. Miguel, N. M. Lopez, C. M. Rossi, and R. D. Rabinovich, "Hantavirus pulmonary syndrome outbreak in Argentina: molecular evidence for person-to-person transmission of Andes virus," *Virology*, vol. 241, no. 2, pp. 323-30, Feb 15 1998.10.1006/viro.1997.8976
- [10] N. Lopez, P. Padula, C. Rossi, M. E. Lazaro, and M. T. Franze-Fernandez, "Genetic identification of a new hantavirus causing severe pulmonary syndrome in Argentina," *Virology*, vol. 220, no. 1, pp. 223-6, Jun 1 1996.10.1006/viro.1996.0305
- [11] M. D. Nieves Parisi, D. A. Enria, N. C. Pini, and M. S. Sabattini, "[Retrospective detection of hantavirus clinical infections in Argentina]," *Medicina (B Aires)*, vol. 56, no. 1, pp. 1-13, 1996. Deteccion retrospectiva de infecciones clinicas por hantavirus en la Argentina., <https://www.ncbi.nlm.nih.gov/pubmed/8734923>
- [12] H. Razuri *et al.*, "Andes hantavirus variant in rodents, southern Amazon Basin, Peru," *Emerg Infect Dis*, vol. 20, no. 2, pp. 257-60, Feb 2014.10.3201/eid2002.131418
- [13] B. K. Rima and N. V. McFerran, "Dinucleotide and stop codon frequencies in single-stranded RNA viruses," *J Gen Virol*, vol. 78 (Pt 11), pp. 2859-70, Nov 1997.10.1099/0022-1317-78-11-2859
- [14] S. Karlin, W. Doerfler, and L. R. Cardon, "Why is CpG suppressed in the genomes of virtually all small eukaryotic viruses but not in those of large eukaryotic viruses?," *J Virol*, vol. 68, no. 5, pp. 2889-97, May 1994.10.1128/JVI.68.5.2889-2897.1994
- [15] S. Jimenez-Baranda *et al.*, "Oligonucleotide motifs that disappear during the evolution of influenza virus in humans increase alpha interferon secretion by plasmacytoid dendritic cells," *J Virol*, vol. 85, no. 8, pp. 3893-904, Apr 2011.10.1128/JVI.01908-10

- [16] X. Cheng *et al.*, "CpG usage in RNA viruses: data and hypotheses," *PLoS One*, vol. 8, no. 9, p. e74109, 2013.10.1371/journal.pone.0074109
- [17] S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129-137, 1982.10.1109/TIT.1982.1056489
- [18] W. H. Kruskal and W. A. Wallis, "Use of Ranks in One-Criterion Variance Analysis," *Journal of the American Statistical Association*, vol. 47, no. 260, pp. 583-621, 1952/12/01 1952.10.1080/01621459.1952.10483441
- [19] O. J. Dunn, "Multiple Comparisons Using Rank Sums," *Technometrics*, vol. 6, no. 3, pp. 241-252, 1964/08/01 1964.10.1080/00401706.1964.10490181
- [20] O. J. Dunn, "Multiple Comparisons among Means," *Journal of the American Statistical Association*, vol. 56, no. 293, pp. 52-64, 1961/03/01 1961.10.1080/01621459.1961.10482090
- [21] A. Saxena *et al.*, "A review of clustering techniques and developments," *Neurocomputing*, vol. 267, pp. 664-681, 2017/12/06/ 2017.<https://doi.org/10.1016/j.neucom.2017.06.053>
- [22] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," presented at the Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Portland, Oregon, 1996.
- [23] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, "OPTICS: ordering points to identify the clustering structure," *SIGMOD Rec.*, vol. 28, no. 2, pp. 49-60, 1999.10.1145/304181.304187
- [24] R. L. Thorndike, "Who belongs in the family?," *Psychometrika*, vol. 18, no. 4, pp. 267-276, 1953/12/01 1953.10.1007/BF02289263
- [25] P. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, no. 1, pp. 53-65, 1987.10.1016/0377-0427(87)90125-7
- [26] B. E. Pickett *et al.*, "ViPR: an open bioinformatics database and analysis resource for virology research," *Nucleic Acids Res*, vol. 40, no. Database issue, pp. D593-8, Jan 2012.10.1093/nar/gkr859
- [27] V. Vrbovska, P. Chalupa, P. Strakova, Z. Hubalek, and I. Rudolf, "[Human hantavirus diseases - still neglected zoonoses?]," *Epidemiol Mikrobiol Imunol*, vol. 64, no. 4, pp. 188-96, Oct 2015. Onemocneni cloveka zpusobena hantaviry - stale opomijene zoonozy?, <https://www.ncbi.nlm.nih.gov/pubmed/26795222>
- [28] J. M. Reynes, D. Carli, N. Boukezia, M. Debruyne, and S. Herti, "Tula hantavirus infection in a hospitalised patient, France, June 2015," *Euro Surveill*, vol. 20, no. 50, 2015.10.2807/1560-7917.ES.2015.20.50.30095
- [29] H. Zelena, J. Mrazek, and T. Kuhn, "Tula hantavirus infection in immunocompromised host, Czech Republic," *Emerg Infect Dis*, vol. 19, no. 11, pp. 1873-5, Nov 2013.10.3201/eid1911.130421