농학석사학위논문

Comparative Genomics Reveals Insights into Viral Evolution of Mammalian Infections

유전체 비교분석을 통한 포유류 감염성 바이러스의 진화에 대한 통찰

2019년 8월

서울대학교 대학원

농생명공학부 동물생명공학전공

김 정 웅

Comparative Genomics Reveals Insights into Viral Evolution of Mammalian Infections

By

Jungwoong Kim

Supervisor: Professor Hee-bal Kim

July, 2019

Department of Agricultural Biotechnology

Seoul National University

Abstract

Comparative Genomics Reveals Insights into Viral Evolution of Mammalian Infections

Jungwoong Kim

Department of Agricultural Biotechnology

The Graduate School

Seoul National University

Infectious viruses infect many species of animal, including human, and cause irreversible consequence. They bring fetal death to human and cause massive economic losses to livestock industry due to the large-scale infection. Therefore, we need more research on infectious viruses. Viruses have faster and random genetic variable features than other organisms. Most viruses are susceptible to infection depending on the host species. However, since a single nucleotide and amino acid sequence variation leads infection to a new species or alter its toxicity, genomic level of virus research provides major commercial and scientific value. Therefore, many researchers focus on the single genetic variation for identification of a new virus species or vaccine study.

Chapter 1 | Zika virus (ZIKV) is known to be associated with a serious brain disease, fetal microcephaly in pregnant women, and has been explosively spread throughout the world over the last decade. Virologists of most countries attempted investigations of ZIKV molecular mechanisms to prevent the

worldwide proliferation. However, only few genetic variants in several regions were anticipated as targets of vaccines and medicines. Here, I analyzed all of available ZIKV complete genomes from the Virus Pathogen Resource (ViPR) database to identify novel genetic markers by considering geographical and temporal perspectives. By principal component and phylogenetic analysis, ZIKV strains formed four clusters according to collected continent. Focusing on the major groups in African, Asian, Central America and Caribbean, I found single nucleotide variants (SNVs) supported by statistical significance. From the dN/dS analysis, I identified the protein coding regions that were evolutionary accelerated in each group. Out of the intercontinental SNVs, non-synonymous and synonymous variants on functional protein domains and predicted B-cell and T-cell epitopes were suggested as regional markers. I believe these local genetic markers can improve medical strategies for ZIKV prevention, diagnosis, and treatment.

Chapter 2 | Influenza D virus (IDV), a new type of influenza, is a respiratory virus that infects ruminants, including cattle. Because the infection symptoms of IDV are mild, but, causes fatal infection of other respiratory viruses and have potential for infection in human, I conducted researches at the genomic level. Using the results of phylogeny and principal coordinate analysis (PCoA), we compared concatenated all of coding sequence dataset and each of genes coding sequence dataset. I confirmed that concatenated dataset results were more appropriately clustered into four groups with isolated region, and I selected the main three groups. Focusing on the main three groups, I found statistically significant genetic markers in comparison with dN/dS analysis, searching protein coding region, and B-cell epitope prediction analysis.

Through this study, I suggest local-specific genetic markers of infectious virus, and these markers will give a deep insight for further studies.

Key words: Infectious virus, Genetic marker, Genomic variants, dN/dS, Protein functional domain, Epitope prediction

Student number: 2017-29310

Contents

List of Tables

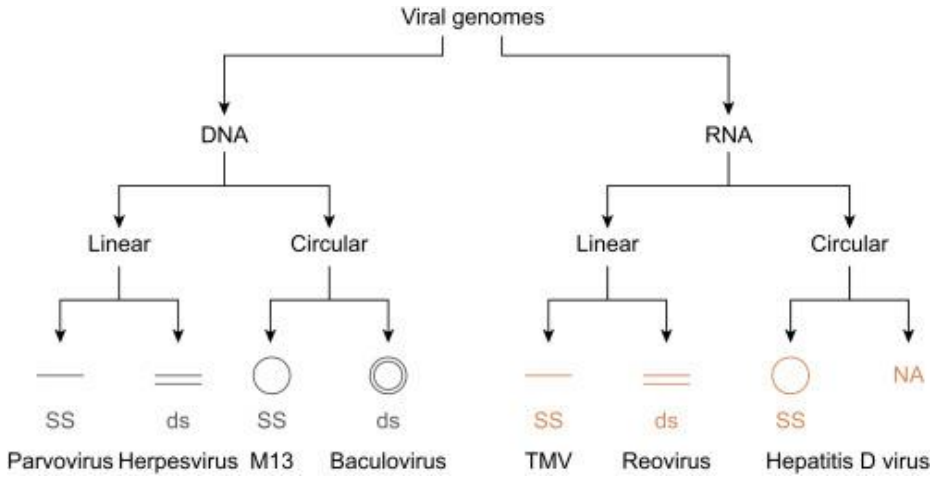
List of Figures

Chapter 1. Literature Review

1.1 Feature of Viruses

Viruses are small infectious agents that perform life activities only in the living cells of other organisms. Viruses can infect all kinds of organisms, from bacteria and animals, including mammals and plants to microorganisms. Studies conducted since Dmitry Ivanovich's 1892 study of cigarette mosaic viruses revealed that viruses are not in infected cells or exist in the form of independent particles during the process of infecting cells. These virus particles, sometimes called virions, consist of two or three parts that protect genetic material made of DNA or RNA. The shape of these viral particles, which are long molecules surrounding the gene coat and the protein envelope, varies from simple spiral and elliptical forms for some virus species to more complex structures for other species. Viral genomes consist of DNA or RNA only, never both. DNA and RNA molecules can be double stranded or single stranded, linear or circular, segmented (composed of multiple pieces of nucleic acid) or nonsegmented. Strictly speaking, a genome segment is an individual and unique piece of nucleic acid among multiple pieces comprising one whole viral genome. For example, the Zika virus and Influenza D virus have segmented genome comprised of ssRNA segments.

Figure 1.1 Diversity of viral genomes (Kaiser, Krieger et al. 2007)



1.2 Development of Sequence technology

Since the Human Genome Project, there has been a need for a sequencing technique for genomes that can eliminate biological anxieties. However, there were many difficulties due to the high price and the limited amount of data to process. However, with the breakthrough in sequencing technology over the past decade, the read length and number of DNA pairs that can be read at one time have been increased, enabling more accurate analysis. In addition, saving time and money, it is now possible to analyze one person's genome at 1 million won. However, the short-read length, which is considered to be the limit of the next generation sequencing, and the error rate due to this are a problem, and the nanopore sequencing technology which can read long length read length by one analysis is also an epoch-making sequencing technique. I did not get to the stage. Nevertheless, many studies have been done through DNA sequencing by overcoming many disadvantages by reducing the error rate by increasing the depth of short read.

1.3 Positive selection ‘branch-site model’

The nucleotide substitutions in gene coding for protein can be either synonymous, alternatively called silent substitutions, or non-synonymous. Usually, most non-synonymous changes would be expected to be eliminated by purifying selection, but under certain conditions Darwinian

selection may lead to their retention. Investigating the number of synonymous and non-synonymous substitutions may therefore provide information about the degree of selection operating on a system. dN/dS ratio is the ratio of non-synonymous substitution rates (dN) per synonymous substitutions rates (dS). The ratios are calculated by the ratio of number of substitutions per number of substitution sites. dN/dS ratio detects the selective force, with 'dN/dS = 1' meaning neutral selection, 'dN/dS < 1' meaning purifying selection, and 'dN/dS > 1' meaning positive selection (Yang 2007).

The branch-site model in positive selection analysis, allow dN/dS ratio to vary both among sites in the protein and across branches on the tree and aim to detect selection affecting a few sites along particular lineage (foreground branches). Although the original branch-site test was found to generate excessive false positive when its assumptions were violated, slight modifications introduced later appear to have made the test far more robust. Thus, the modified test is commonly used to detect evolutionary meaning.

1.4 Epitope prediction

Epitope, also known as antigenic determinant, is the part of an antigen that is recognized by the immune system, specifically by antibodies, B cells, or T cells. For example, the epitope is the specific piece of the antigen to which an antibody binds. The part of an antibody that binds to the epitope is called

a paratope. Although epitopes are usually non-self proteins, sequences derived from the host that can be recognized (as in the case of autoimmune diseases) are also epitopes.

Adaptive immunity is mediated by T- and B-cells, which are immune cells capable of developing pathogen-specific memory that confers immunological protection. Memory and effector functions of B- and T-cells are predicated on the recognition through specialized receptors of specific targets (antigens) in pathogens. More specifically, B- and T-cells recognize portions within their cognate antigens known as epitopes. There is great interest in identifying epitopes in antigens for a number of practical reasons, including understanding disease etiology, immune monitoring, developing diagnosis assays, and designing epitope-based vaccines. Epitope identification is costly and time-consuming as it requires experimental screening of large arrays of potential epitope candidates. Fortunately, researchers have developed in silico prediction methods that dramatically reduce the burden associated with epitope mapping by decreasing the list of potential epitope candidates for experimental testing (Huang and Honda 2006).

Chapter 2. Identification of local-specific genetic markers of zika virus across the entire globe

2.1 Abstract

Zika virus (ZIKV) is known to be associated with a serious brain disease, fetal microcephaly in pregnant women, and has been explosively spread throughout the world over the last decade. Virologists of most countries attempted investigations of ZIKV molecular mechanisms to prevent the worldwide proliferation. However, only few genetic variants in several regions were anticipated as targets of vaccines and medicines. Here, we analyzed all available ZIKV complete genomes from the Virus Pathogen Resource (ViPR) database to identify novel genetic markers by considering geographical and temporal perspectives. By principal component and phylogenetic analysis, ZIKV strains formed four clusters according to collected continent. Focusing on the major groups in African, Asian, Central America and Caribbean, we found single nucleotide variants (SNVs) supported by statistical significance. From the dN/dS analysis, we identified the protein coding regions that were evolutionary accelerated in each group. Out of the intercontinental SNVs, nonsynonymous and synonymous variants on functional protein domains and predicted B-cell and T-cell epitopes were suggested as regional markers. We believe these local genetic markers can improve medical strategies for ZIKV prevention, diagnosis, and treatment.

2.2 Introduction

Zika virus (ZIKV) is a member of the Spondweni serogroup of mosquito-borne virus, transmitted by *Aedes* mosquitoes. It belongs to the *Flavivirus*, genus of viruses in the family *Flaviviridae,* which are positive-sense, single-stranded and enveloped ribonucleic acid (ssRNA). Its name, Zika, originates from the Zika Forest of Uganda where the virus was firstly isolated in 1947 (Dick, Kitchen et al. 1952, Malone, Homan et al. 2016). From its isolation, the ZIKV has been known to occur within a narrow equatorial belt between Africa and Asia. In Asia, the first isolation of its virus was in Malaysia in 1966 and subsequently circulated in the Philippine, Malaysia, Singapore, China and Thailand (Alera, Hermann et al. 2015, Buathong, Hermann et al. 2015, Deng, Zhao et al. 2016, Kindhauser, Allen et al. 2016, Ho, Hapuarachchi et al. 2017). It spread to Micronesia in Oceania in 2007, then to French Polynesia in 2013, and finally arrived in the Americas in 2014 which lead to a major ongoing pandemic (Foy, Kobylinski et al. 2011, Tappe, Rissland et al. 2014, Mehrjardi 2017). Despite the fact that the first human infection of ZIKV reported in 1954 outbreaks, in 2015, it became a global health emergency due to its ability to cause severe birth defect such as microcephaly in pregnant woman or trigger paralysis called Guillain-Barré syndrome in adults (Organization 2016, Rasmussen, Jamieson et al. 2016, White, Wollebo et al. 2016). However, there are no approved vaccines or specific treatment for ZIKV. Therefore, the physical isolation is the only way to prevent ZIKV infection in current.

Most purposes of ZIKV genome studies lied on discovering new genomic markers differentially evolved in each conventional region and understanding the evolutionary dynamics of the virus. According to recent reports, nucleotide substitutions have been noted as indicators of the major evolutionary mechanisms and the adaptation of the virus to the host (Israr-ul, Allen et al. 2013, Kaminski, Ohnemus et al. 2013, Schvoerer, Moenne-Loccoz et al. 2013). These different genomic variants among local environments and host-to-host infections can fine-tune the bio-interactions between viruses and new hosts (Pepin, Lass et al. 2010, Plotkin and Kudla 2011, Longdon, Brockhurst et al. 2014). It is necessary to conduct genome-wide comparative analyses to find new target genomic regions of viral vaccines to develop preventive drugs for ZIKV.

As a member of the family *Flaviviridae*, ZIKA virus (ZIKV) has approximately 10.8kb RNA genome with 11 protein coding genes. Its ORF encodes three structural proteins (capsid C, pre-membrane preM and envelope E) to form a regular spherical shell and to infect a host, and seven non-structural proteins (NS1, NS2A, NS2B, NS3, NS4A, 2k, NS4B, and NS5) to coordinate viral replication and inhibit immune response and participate in key mechanisms (Bahir, Fromer et al.). Out of these protein coding genes, only three were used as target genes of RT-PCR assay to identify ZIKV compared to the other *Flaviviridae* species (Musso and Gubler 2016). In classification of ZIKV strains, a previous research demonstrated based on the complete genome

sequences of three geographically and temporally distinct ZIKV strains: African lineage, Asian lineage, Asian lineage derived American-sub lineage (Yun, Song et al. 2016). However, they missed to check the major groups of each local region and to suggest candidate markers supported by evolutionary histories and possibilities of the markers as new vaccine targets.

Here, based on ZIKV complete genomes isolated from various countries, we tried to find novel genomic markers of ZIKV strains clustered by PCA and phylogenetic analysis, and we compared it with the previously reported local markers. Next, we illuminated candidate genomic markers specific to major groups in each cluster supported by the natural episodic adaptations, the functional domain search and the epitope prediction analysis. This study would provide a deeper insight to understand molecular mechanisms of ZIKV.

2.3 Materials and Methods

2.3.1 RNA genome sequences of ZIKV

In order to identify local genetic markers of ZIKV to provide against the rapid spread in the world, we collected complete coding sequences of ZIKV with their isolation information, such as isolation time and region and host species from Virus Pathogen Resource (ViPR) database (www.viprbrc.org, release date: Jul 14th 2017) (Pickett, Sadat et al. 2011). There were total of 450 strains. After filtering out 149 strains with InDels in coding regions and without isolation information of each sample, 301 ZIKV strains sampled in 33 countries remained (Table 2.1).

2.3.2 Multiple sequence alignment of orthologous gene sets

To match homologous sites of each strain in coding regions, we performed multiple sequence alignments for each orthologous gene set of ZIKV. Out of 10,254 – 10,272 nucleotide sequences of ZIKV, finally 10,227 nucleotide sequences of ZIKV strains were aligned in three steps. First, we eliminated data sets, which have a poor sequencing quality and then, translated codon sequences to amino acid sequences by using the 'biopython' library in Python 3.4.1 (Chapman and Chang 2000). Next, we respectively aligned 11 orthologous gene sets by using PRANK program (v.140110) (Löytynoja and Goldman 2008). Last, the translated amino acid sequence alignments were reversely translated to nucleotide sequences by using the PAL2NAL program (v.14) (Suyama, Torrents et al. 2006). In addition to improve accuracies of

analyses and extract clear results, we discarded poorly scored alignments to avoid bias of alignment step by using GBLOCKS program (v.0.91b) (Talavera and Castresana 2007). After each of these phases, a total of 10,227 nucleotide sequences of 11 orthologous genes were remained.

2.3.3 Principal Component Analysis and Phylogenetic tree construction of ZIKV

Principal component analysis (PCA) was used to obtain an understanding of the relationships between the aligned flavivirus data. For PCA, a similarity matrix was obtained from the aligned sequences by JalView (v.2.10.1) (Clamp, Cuff et al. 2004). From the matrix, the PCA plot results were visualized using R package ggplot2. For grouping based on the similarity of isolation information and genomic character we used the k-means algorithm (Hartigan and Wong 1979) and the optimal number of clusters showing the highest D index values in the 2nd differences which was determined from the NbClust R package (v.3.0) (Charrad, Ghazzali et al. 2012). Phylogenetic reconstruction was completed using maximum likelihood method with a bootstrap value of 1000 in the MEGA7 software base on the Tamura-Nei model (Kumar, Stecher et al. 2016). For this phylogenetic analysis, total of 301 ZIKV strains in 10,227 nucleotide sequences (Table 2.1) were used.

2.3.4 Comparative genomic analysis for local specific variants

By using the clustering analysis results, comparative genomic analysis was performed to identify nucleotide that are statistically different among the

groups by the metadata-driven Comparative Analysis for Sequences (meta-CATS) tool in ViPR. The meta-CATS performs a chi-square test to identify positions that are significantly different than the random distribution of residues between all metadata groups, and a Pearson's chi-square test in tandem to calculate a p-value to identify the specific pairs of groups, and shows a maximum probability cut off for statistical significance (default=0.05) (Pickett, Liu et al. 2013). In this study, we manually classified into major four groups (Africa, Singapore, Central America and Caribbean) based on the results of PCA and phylogeny analysis.

2.3.5 Group specific Amino Acid substitutions of ZIKV

To identify not only the statistical single nucleotide substitution but also mutually exclusive substitution in codon and amino acid, TSNV and TAAS analyses (Zhang, Li et al. 2014) were performed. Codon substitution sites, which are mutually exclusive between designated group and others, were examined to investigate convergent evolution at the molecular level. The TSNV analysis was performed with the four designated group datasets (Africa, Singapore, Central America and Caribbean) as the target groups. And the visualization of target group specific conserved variation on amino acid and nucleotide level used WebLogo3 (Crooks, Hon et al. 2004).

2.3.6 dN/dS Analysis of coding sequences for group specific positive selection

For protein coding genes, the most widely used method for identifying specific sites of positive selection is the nonsynonymous/synonymous rate ratio (dN/dS, ω) test. In order to identify evolutionary accelerated protein in ZIKV with the ratio values of the rate of nonsynonymous (dN) to synonymous (dS) substitutions of each protein gene sets (three structural and seven non-structural), we used site and branch-site models of the PAML package (v4.9a) (Yang 2007). Based on the PCA and phylogeny tree, four group lineages (Africa, Singapore, Central-America and Caribbean) were set as the foreground branches that particular lineages affected positive selection in each continental clade.

The option of codeml control files of branch-site model set was 'model = 2, NSsites = 2, and fix_omega = 0, CodonFreq = 2' in the alternative model. As a result, the p-value of each protein gene was calculated through compared the null and alternative model maximum likelihoods using likelihood ratio test (LTR) 'D = 2 * Δ *l*' and chi-square distribution. Hence, we identified sites affected positive selection on each lineage that were found based on the BEB inference, as well as the positive selective sites.

2.3.7 Conserved Domain search

To check whether the sequence variation was located in the functional domain of each protein, the oldest data among the ZIKV sequences was selected (strain: ZIKV/Macaca- mulatta/UGA/MR-766/1947) as a query. Prior to the domain search, the aligned sequences were translated into amino acid

sequences to find the domain of each protein. Also, the amino acid sequences were divided into 3 structure proteins (ancC, preM and E) and 7 non-structure proteins except protein '2k' (NS1, NS2A, NS2B, NS3, NS4A, NS4B and NS5). We searched for a total of 10 proteins searched for in the NCBI batch CD-search tool (https://www.ncbi.nlm.nih.gov/Structure/bwrpsb/bwrpsb.cgi) (Marchler-Bauer, Lu et al. 2010).

2.3.8 B-cell Antigenic & T-cell Cytotoxic T-lymphocyte (CTL) Epitope prediction

Using immune-informatic approach analysis and focusing on antigenicity, we predicted potential B-cell epitope to confirm the substitution regions, which the results of genomic comparative analysis shared with the epitope regions. The prediction analysis was conducted with the dataset representative strain of each group on domain search analysis. For antigenicity prediction we applied the 'Kolaskar and Tongaonkar' antigenicity prediction method which has been reported that it provides 75% experimental accuracy (Kolaskar and Tongaonkar 1990) to evaluate the analyzed antigenicity on the basis of the physicochemical properties of amino acids and their abundances in experimentally known epitopes. The NetCTL.1.2 server (http://www.cbs.dtu.dk/services/NetCTL/) was used to predict the Cytotoxic T-lymphocyte (CTL) epitopes. NetCTL is a predictor of T-cell epitopes along a protein sequence, which employs a neural network architecture (Larsen, Lundegaard et al. 2005, Kiepiela, Ngumbela et al. 2007). As an input, FASTA format sequences of each ZIKV amino acid sequence were provided (prediction score > 0.75000 threshold, HLA-A-3

supertype). As a result of the prediction analysis, we found the two types of candidates focusing on the three proteins (Envelope protein-E, NS1, NS3, NS5) that were (1) existing only the specific region and (2) commonly existing, but affecting the epitope score.

2.4 Results

2.4.1 Genome sequences of ZIKV

We collected complete genome sequences of 450 ZIKV strains registered in the ViPR database (www.viprbrc.org, release date: Jul 14th 2017). To improve accuracies of analyses, we excluded equivocal data sets, which either have a poor sequencing quality (with ambiguous nucleotide information as 'N' or without partial coding genes) or unconformable sampling information such as isolation region, date and host. Filtering 450 data sets based on the above criteria, we prepared the total of 301 strains of ZIKV complete coding sequences. In order to match homologous sites to the 301 strains of ZIKV, we performed multiple sequence alignments on 11 orthologous gene sets (ancC, preM, E, NS1, NS2A, NS2B, NS3, NS4A, 2k, NS4B and NS5), respectively. The concatenated alignments had 10,227 bases.

2.4.2 ZIKV strains clustered ty continental regions roughly

Most of the ZIKV genome studies classify the ZIKV strains into African and Asian clades based on their epidemical history (Hayes 2009, Haddow, Schuh et al. 2012, Ye, Liu et al. 2016, Jun, Wassenaar et al. 2017). To test how many groups could be clustered by genomic information, we performed Principal Component Analysis (PCA) based on the similarity matrix obtained from the aligned sequences as well as clustered the strains by the k-means method which are a procedure for clustering into a fixed number, k, of groups in each analysis (Hartigan and Wong 1979). Estimating the optimal number of

clusters by using the D index method (Dohnal, Gennaro et al. 2003, Charrad, Ghazzali et al. 2012), we checked the optimal D index values and the second differences D index values. It determined the optimal number of clusters as four in all 301 ZIKV strains (Figure 2.1 a). According to the optimal clustering number, the result of PCA using all ZIKV dataset formed four cluster groups (Africa, Singapore, Southeast Asia and Mixed group) (Figure 2.1 b). In the PCA plot, the 'Africa group' was distant from the other major clusters, and the 'Singapore group' with the highest density was distant from the 'Southeastern Asian group' (Figure 2.1 b).

In contrast to the above three major clusters from geographically adjacent regions, the 'Mixed group' included ZIKV strains isolated various regions including Asia, Europe and America. However, predominant strains that consisted of the mixed cluster were isolated from American continent which were known to be originated from Asia continent in recent years (Ye, Liu et al. 2016). We tried to confirm the 'Asia-derived America group' in the mixed cluster. Estimating the optimal number of clusters as five by applying the D index method (Figure 2.1 c), we found the mixed cluster was composed of five sub-cluster groups (Caribbean, Central-America, China, French Polynesia and Mixed America group) (Figure 2.1 d). From the PCA result, we confirmed that the ZIKV strains isolated from America continent originated from two independent main group: Central America (Nicaragua, Guatemala, South Mexico and Honduras) and Caribbean (Florida of USA, Dominican Republic and Haiti) (Grubaugh, Faria et al. 2018).

2.4.3 Phylogenetic relationships of ZIKV

Phylogenetic analysis was performed to investigate the genetic distances and historical relatedness among ZIKV strains. Based on multiple sequence alignments of coding sequences of 301 ZIKV strains, phylogenetic analysis was performed by the Maximum Likelihood (ML) methods (Kumar, Stecher et al. 2016). As a previous research (Hayes 2009), ZIKV represented two lineages related to the geographic distribution and isolated time as an African lineage and Asian lineage (Figure 2.2). The African lineage was composed of the ZIKV strains that were isolated from African countries (Uganda, Central Africa, Nigeria and Senegal), whereas the Asian lineage was isolated not only from Asia countries, but also from America and some of European countries (Figure 2.2). Within the Asian lineage, there is a monophyletic clade of the ZIKV strains isolated from American countries and several countries showed as a new American lineage (Figure 2.2), which was known to be originated from French Polynesia strains (Haddow, Schuh et al. 2012). When dividing the phylogenetic tree by countries, the ZIKV strains isolated from Singapore constructed a specific distinct Singapore group in the Asian lineage, and the ZIKV strains isolated from Caribbean (Florida of USA, Dominican Republic and Haiti) and Central America (Nicaragua, Guatemala, South Mexico and Honduras) constructed distinct sub-group each in the American lineage (Figure 2.2). Based on the clustering results of PCA and phylogenetic analysis, 301 ZIKV strains were manually classified into four major group (Africa, Singapore, Caribbean and Central America group).

2.4.4 Local-specific genomic markers of ZIKV

Focusing on the four major groups, we performed comparative genomic analyses to find new genomic markers specific local groups. Nucleotide variations have long been noticed as genomic markers to understand evolutionary relationships between viruses to give insights into viral adaption to host (Wichman, Badgett et al. 1999). To identify statistically significant genomic variants specific to each major group in clustered groups different from the rest strains, we applied metadata-driven Comparative Analysis Tool for Sequences (meta-CATS) analyses (Pickett, Liu et al. 2013). We divided ZIKV strains into one of major group and the other strains in each analysis. In the nucleotide alignments of the ZIKV datasets, we identified 1,674 Africa group-specific nucleotide variations (Figure 2.3 a), 438 Singapore group-specific nucleotide variations (Figure 2.3 b), 192 Central America group specific nucleotide variations (Figure 2.3 c) and 131 Caribbean group specific nucleotide variations (Figure 2.3 d).

Although nucleotide variants supported by statistical significance can explain the genomic differences among ZIKV strains, it is hard to say all of the nucleotide variants can contribute ZIKV evolution equally. In coding regions, nucleotide variants are classified into synonymous and nonsynonymous substitutions. Out of both type of substitutions, only nonsynonymous substitutions can alter protein sequences and their functions. If a major group would have a different protein function compared to the other strains at a homologous site, the major group have to get amino acid substitutions mutually exclusive to the other strains. To identify mutually exclusive single nucleotide and amino acid variants to each major group, we performed target-specific

single nucleotide variants (TSNV) analysis and target -specific amino acid substitution (TAAS) analysis. First, by applying the TSNV analysis with nucleotide alignments, we identified 339 TSNVs of Africa group (20.2%, compared with meta-CATS results), 8 TSNVs of Singapore group (1.8%), 4 TSNVs of Central America group (2.0%) and 1 TSNV of Caribbean group (0.7%). Next, we performed the TAAS analysis with amino acid sequence data to identify each group-specific conserved nonsynonymous substitution that cause amino acid variation. We discovered 36 TAASs of Africa group, 2 TAASs of Central America group and 1 TAAS of Caribbean group (Supplementary Table S2.4 and S2.5). However, there was no Singapore group specific TAAS. Our 36 amino acid markers, specific to Africa group specific markers, were less than 46 markers previously reported by Ye, Qing, et al. and Wang, Lulan, et al. (Wang, Valderramos et al. 2016, Ye, Liu et al. 2016). Out of 36 African group markers in this study, 33 were shared with previous results, but 3 amino acid substitutions (I/V208L in NS2A, I519V and N524S in NS5) were initially identified. In addition, we also identified initially Central America (A100G in NS1 and L572M in NS3) and Caribbean (E877D in NS5) group specific nonsynonymous substitution sites.

2.4.5 Positive selection of local markers of ZIKV

To estimate positive selection on group markers associated with survival and propagation of ZIKV, we performed dN/dS analysis with branch-site model of codeml in PAML package (Yang 2007). In this study, the most recent common ancestral branches of four each major group was used as a foreground branch under positive selection, while the rest branches were used as

background branches under neutral or purifying selection in the phylogenetic tree (Figure 2.2). Among the 10 ZIKV proteins of four groups, the two NS1 and NS3 protein of Central America group and one NS5 protein of Caribbean group were identified as evolutionarily accelerated proteins. Based on posterior probabilities to detect positively selected sites, the Caribbean group had a marker in NS5 and the Central America group had two markers in NS1 and NS3 under positive selection on their common ancestral branch ($D > 0$, ω of foreground branches > 1, BEB > 0.5) (Table 2.2). On the other hand, the Africa group did not show any markers under positive selection. We assumed the positively selected markers could be a key role in adaptation of ZIKV stains in various countries and hosts.

2.4.6 Local markers causing gene product alterations on functional domains

In order to detect the functional domain organizations on three structural proteins and seven non-structural proteins, NCBI conserved domain search was conducted. After the domain search, we overlapped the four major group specific synonymous and nonsynonymous substitution regions on all domains. As a result, Africa group specific mutually exclusive amino acid substitutions existed in all of each domain, except Flavi_NS4A domain. Out of four ZIKV proteins that play a physiologically important role, E protein had three domains Flavi_glycoprotein, Flavi_E_C and Flavi_E_stem, which contained 2, 1, and 1 amino acid substitutions, respectively (Figure 2.4) (Ye, Liu et al. 2016). In addition, there were two amino acid substitutions in Flavi_NS1 domain of NS1, but among the five amino acid substitutions in NS3, only 2, 1 and 1

substitutions located in Prptidase_S7, Flavi_DEAD and HELICc domains each. The amino acid substitutions were most abundant in the NS5. Two domains of NS5, FtsJ and Flavi_NS5, contained 1 and 14 amino acid substitutions each. Between the two mutually exclusive Central America group specific amino acid substitutions, only one substitution was located in Flavi_NS1 domain of NS1. On the other hand, only one Caribbean group specific amino acid substitution existed in Flavi_NS5 domain of NS5. As we described group markers, there was no Singapore group specific amino acid substitution (Figure 2.4 b).

2.4.7 Epitope prediction of local genetic markers

Epitope is an antigenic determinant, which is distinguished by antigenic specificity, especially the antibody B-cells and T-cells. In addition, epitope binds antibody with its free form and as membrane bound B-cell receptor called as B-cell epitope, it binds to divide with T-cell epitope, which is proteolytically cleaved peptides of the antigen that interacts with the receptors of T-cells. In this paper, we predicted potential B-cell and T-cell epitopes to identify which four group markers affect the prediction of epitopes using immunoinformatics approach. These epitope prediction analyses were performed based on the immunologically and physiologically important three ZIKV proteins E, NS3 and NS5 (Malet, Massé et al. 2008, Perera, Khaliq et al. 2008).

The method of 'Kolarskar and Tongaonkar' based on the physicochemical characteristics of amino acids (Kolaskar and Tongaonkar 1990), B-cell antigenic epitopes were predicted (Supplementary Figure S2.1) and identified that the E protein had 20 antigenic peptides in the Africa group and 19 antigenic peptides in others which shares same amino acid sequences. Likewise, in the

NS1 protein, the analysis predicted 14 antigenic peptides in the Africa group and 15 antigenic peptides in others. Moreover, in the NS3 protein and NS5 protein, 22 and 33 antigenic peptides were predicted respectively, in all groups.

Cytotoxic T-cell is a kind of T-lymphocyte (CTL, cytotoxic T-cell lymphocyte) that destroys virus-infected cell or damaged cell from other factors. Most cytotoxic T-cell contains receptors that have functions, which distinguish specific antigenic, bounded to the type I MHC molecule located on the surface of all cells. In addition, a glycoprotein called CD8 is located on the surface of cytotoxic T-cells and bind to MHC type I molecules to perform cytotoxic activities. Therefore, CTL epitope prediction is a useful in silico tool for its vaccine design. CTL epitope prediction through neural network architecture of proteins was performed using the NetCTL.1.2 server with human leukocyte antigen (HLA) A-3 supertype that were known to have important locally-clustered interactions that synergistically stabilizes the peptide-MHC complexes (Ragoza, Hochuli et al. 2017). As a result, the one Africa group marker were contained in the predicted CTL epitope which the prediction scores were above the >0.75 threshold from ZIKV E protein. Moreover, one CTL epitope which containing the Africa group marker was predicted from the NS3 protein also contained the Africa group markers each. In particular, CTL epitopes predicted from NS5 protein contained the most Africa group markers. In case of the Caribbean group, one Caribbean group marker was contained in the two CTL epitopes, which predicted from the NS5 protein. In contrast, there was no CTL epitope, which containing the Central America group markers. These Africa and Caribbean group markers affected the CTL epitopes, resulting in differences in the number of epitopes and in the epitope scores between the groups. (Supplementary Figure S2.1 and Supplementary Figure S2.2)

2.5 Discussion

Our research aimed to investigate the spread of local-specific Zika virus (ZIKV) genomic variations in global regions to identify genetic markers as epitopes for therapeutic vaccine development. Particularly, due to the collected ZIKV coding sequence data was roughly clustered on the continental region basis, four major local groups were designated geographically and genetically adjacent. Focusing on each of these four major groups, we identified significantly different local-specific single nucleotide variants (SNVs) from genomic information of ZIKV in the other groups and uncovered mutually exclusive SNVs between each of the major groups and the rest of ZIKV strains. Out of these SNVs, several synonymous and nonsynonymous variants harbored functional domains of protein coding genes. In addition, we conducted dN/dS analysis to estimate positive selection on the local-specific variants associated with viral survival and propagation. Furthermore, we performed epitope predictions to illuminate candidate sites with SNVs specific to each of the local group for vaccine developments. Through the findings of our research, we suggest the useful local markers of ZIKV.

In this study, it was important to compare and analyze the genomic features of each local-specific ZIKV. For the PCA analysis, geographically and genetically adjacent four major groups were clustered as Africa, Singapore, Central America and Caribbean group. The PCA performed using the different two types of dataset. The first of PCA used the total of 301 global ZIKV strains dataset and the second PCA used the 170 ZIKV strains, which clustered as ‘Mixed group’ in PCA result using the 301 global ZIKV dataset. According to the PCA result using the 301 global ZIKV strains dataset, the ZIKV strains

were classified as four major groups by the k-means clustering analysis. As expected, the Africa strains formed a cluster specifically apart from the out of Africa strains, and Singapore group formed a dense and distinguish group apart from the 'Southeast Asia group'. This observation shows that the genome feature of Africa strains differs from the out of Africa strains, and Singapore strains are geographically including Southeast Asia but genetically differed (Figure 2.1 b). The second PCA dataset mostly contained the America ZIKV strains, which is causing neurological disorder and rapidly spread in recent, were classified as five sub-groups. In particular, two of the five sub-groups (Central America and Caribbean group) show distinguish identical clusters. From this result, it can be inferred that the ZIKV strains were evolved forming geographically independent strains and containing local-specific genetic feature.

To investigate the genetic distance and historical relatedness between ZIKV populations, we performed the phylogenetic analysis and identified two viral lineages: African and Asian (Hayes 2009). It also showed that the evolutionary epidemiology was circulated outside of Africa until early 2017. Based on these results, we can infer the geographical pathway of ZIKV. When considering the evolutionary aspects and the time of the epidemic through phylogenetic tree, ZIKV, which was circulated in early Southeast Asia, indicated that it has spread to French Polynesia and Americas (Figure 2.2) (Enfissi, Codrington et al. 2016, Wang, Valderramos et al. 2016). Similar to the PCA results, phylogenetic analysis was generated a distinct Singapore clade within the Asian strains, whereas all of the isolated ZIKV in Americas were clustered closely within the Asian lineage, forming a new American sub-clade which is identical to the French Polynesia (Figure 2.2) (Ye, Liu et al. 2016).

This result supports the uniqueness of the imported ZIKV clade from the French Polynesia that was spreading through the Americas (Zanluca, Melo et al. 2015). Especially, in the American lineage, there were two distinct sub-clades composed of the ZIKV strains isolated from Central America and Caribbean, showing the similar result drawn by the PCA (Figure 2.1 d). As a result, we confirmed a strong correlation between the genetic characterization of the root of the local ZIKV phylogeny and the region of sequence isolated for local dataset. Besides, these correlations suggest a possibility of that the local characterization of specific ZIKV coding sequences, especially the four major groups, and its effectiveness from adaptation of environment and host immune pressure. The America strains become phylogenetically diversified and at the same time generated independent strains, Central America and Caribbean.

As important evolutionary mechanisms, investigation of genetic variation provides deeper insights into viral pathogenesis and reveals functional changes in the virus (Weaver 2017). Based on the evolutionary analysis, we conducted a statistical genomic comparison analysis that were able to identify the genomic sites, which were significantly different between continental strains (Figure 2.3). Moreover, it also identified mutually exclusive nucleotide and amino acid substitutions of its local group specifics using 301 ZIKV strains.

In particular, compared with the previous researches, the identification of fewer results as in 36 strains of African lineage specific simultaneously Asia lineage specific amino acid substitutions caused from the number of ZIKV strains. Because of the greater number of ZIKV strains used as Asian lineage, the more various amino acid substitutions were existed in Asian lineage, which lead to differences in total number of mutually exclusive amino acid variation

among local group specifics. Moreover, increased in total number of ZIKV strains made it possible to identify the amino acid variation markers for Central America and Caribbean group. Therefore, the more the strains, the less mutually exclusive amino acid substitutions are identified. On the other hand, the number of ZIKV used as African lineage was reduced than previous researches which happen to have less amino acid variation sites within the local groups. In addition, the identification of amino acid variation markers for Central America and Caribbean group specific were verified and evolved through independent root within the Americas. This result reflects that ZIKV established independent strains in each region, and significant parts of the differences of nucleotide and amino acid sequences were conserved over time. The group-specific variation sites have potential to be used as markers that identify not only the Africa lineage from non-African lineages, but also each local group identification marker. In addition, dN/dS analysis was demonstrated to confirm whether the suggested group markers were under positive selection within the continental sets. We found the local-specific positive selected sites that shares the amino acid substitution marker. Although the results of dN/dS analysis were not statistically sufficient, it is important to note that the group markers suggest the under positive selected sites by matching with the results of the genome comparative analysis. Particularly, the proposed candidates of group-specific markers reflect that there were two independent infection process routs from southern America to northern America each with individual positive selection.

A nonsynonymous substitution is a nucleotide mutation that alters the protein sequence, which most likely change the resulting protein that is expressed. Using the group specific nonsynonymous substitution sites

identified from the genomic comparative analysis, we proposed the sites as candidates that identify the belonging local group, as well as induce domain functional changes by overlapping the sequence variant sites to domains found with the NCBI batch CD-search tool in proteins related to ZIKV major functions (Figure 2.4). Compared with other strains, the mutually exclusive nonsynonymous substitution sites in the Africa group were relatively distributed in the proteins except for the NS4 protein. This result showed that the Africa group is gene widely independent compared with the non-African group. Especially, the four Africa group specific nonsynonymous substitution sites were in E protein which is known for including glycosylation sites that play important roles in the infectivity, maturation and virulence (Ye, Liu et al. 2016). In addition, NS1 contained the RNA replication and immune system modulation domain 'Flavi_NS1', and NS3 contained the Helicase/NTPase, which is a member of the P-loop NTPase domain superfamily. NS5 contained the N-terminal methyltransferase domain which methylates the 5′cap structure of genomic RNA (Ftsj) and the C-terminal terminal RNA dependent RNA polymerase (Flavi_NS5) domain which includes Africa group specific markers. Interestingly, there was no common nonsynonymous substitution site found in the Singapore group, whereas the Central America and the Caribbean groups shared a single site in Flavi_NS1 domain of NS1 protein and Flavi_NS5 domain of NS5 protein, respectively. These results have potential of the rapid epidemic of ZIKV in America, which caused by the local specific amino acid substitutions on host infectivity or viral survival domains. However, the epidemic of ZIKV in Singapore was caused by the amino acid substitution on functional protein domain rather than other factors.

Furthermore, we tried to confirm whether the candidates identified by the previous genomic comparative analysis share the potential B-cell and CTL epitope, which is an immuno-intelligent approach, thus genetic variation in the functional genes could affect the local-specific peptide vaccine design. For antigenicity prediction, we used the 'Kolaskar and Tongaonkar' semi-empirical method to evaluate the frequencies of amino acid residues in experimentally determined epitopes. Additionally, we also predicted epitope candidates that can be used for vaccine design using the CTL epitope prediction. Among the ZIKV proteins, we focused on the E, NS3 and NS5 protein, which is known as potential targets for therapeutics and vaccines development, because these targets have been recognized to have a major role in viral entry into the cell and in viral replication. In the predicted B-cell epitopes, the most distinguished differences were between Africa group and non-Africa group epitopes. Non-Africa groups (Singapore, Central America and Caribbean) shared B-cell epitopes, whereas there were Africa group specific B-cell epitopes.

Among the CTL epitope candidates, only the Africa and Caribbean group markers were contained in the predicted CTL epitopes (Table 2.4). Through the comparison of epitope candidates among the four groups, it would be possible to help find and design a vaccine candidate for a group-specific vaccine and to cope with the emergence of a new type due to the evolutionary mutation of ZIKV. Although it is possible to utilize the vaccine design as a group-specific marker through epitope candidates located in the functional domain, there is a limit that it cannot confirm whether the markers affect the functional change of the virus. To overcome the limitations, we need further studies on molecular structural levels.

In conclusion, we have confirmed the local specific substitutions being used as an epitope candidate through observing a strong correlation between the genomic variation and the isolated regions. As we studied to identify genomic structural features and candidates against the pathogen for clarifying local specific antiviral drugs and vaccine designs, we observed the local specific sequence variation in the protein domain accounted for an important role in the viruses of ZIKV by performing genomic comparative analysis and evolutionary analysis. The variation regions identified as a marker in the particular group-specific epitope candidates where it has the potential to cause the changes in major mechanism in ZIKV; however, it is unknown whether for these amino acid substitutions would affect the viral pathogenicity, fitness, transmission and reproductive functions of the particular region. Despite the fact that the amino acid variations we identified were not directly related to the fact of circulation among local groups, these changes were located in the structural and biochemical functional domains of ZIKV, which explains the trend of the virus. Therefore, we suggest that the variations may have influenced the epitope formation and the decisive action between local groups compared with B-cell and CTL epitope candidates. This study would provide not only some insight into the evolution of ZIKV and virulence determination, but also valuable information for future vaccine and drug development.

Table 2.1 Summary of ZIKV complete genome

Country	# strains	Host (# strains)	Year
Uganda	4	Monkey (4)	1947-1947
Malaysia	3	Mosquito (3)	1966-1966
Nigeria	1	Human (1)	1968-1968
Central African Republic	1	Mosquito (1)	1976-1976
Senegal	6	Mosquito (6)	1984-1984
Cambodia	1	Human (1)	2010-2010
Philippines	1	Human (1)	2012-2012
French Polynesia	11	Human (11)	2013-2014
Thailand	4	Human (4)	2013-2016
Haiti	8	Human (7), Mosquito (1)	2014-2016
Guatemala	2	Human (2)	2015-2015
Martinique	1	Human (1)	2015-2015
Puerto Rico	4	Human (4)	2015-2015
Brazil	25	Human (25)	2015-2016
Colombia	9	Human (9)	2015-2016
Honduras	5	Human (5)	2015-2016
Mexico	11	Human (7), Mosquito (4)	2015-2016
Panama	4	Human (4)	2015-2016
Suriname	3	Human (3)	2015-2016
China	20	Human (20)	2016-2016
Dominican Republic	7	Human (7)	2016-2016
Ecuador	2	Human (2)	2016-2016
French Guiana	1	Human (1)	2016-2016
Italy	4	Human (4)	2016-2016
Japan	3	Human (3)	2016-2016
Nicaragua	12	Human (12)	2016-2016
Peru	2	Human (2)	2016-2016
Singapore	107	Human (93), Mosquito (14)	2016-2016
Taiwan	1	Human (1)	2016-2016
Tonga	1	Human (1)	2016-2016
USA	32	Human (23), Mosquito (9)	2016-2016
Venezuela	1	Human (1)	2016-2016
Russia	4	Human (4)	2016-2017

Table 2.2 Positive selection on group markers estimated by using branch-site model

Foreground branch group	Protein	H0 lnL	H1 lnL	LRT	P-value	W_2	Protein Position	Posterior probability
Caribbean	NS5	-6198.48	-6198.27	0.41	0.51	3.84	877	0.757
Central America	NS1	-2508.02	-2508.38	0.71	0.39	108.77	100	0.967
	NS3	-3958.87	-3958.47	0.79	0.37	8.37	572	0.875

Table 2.3 Comparison of predicted antigenic B-cell epitopes of ZIKV between 'Africa group' and 'out of Africa group'

Protein	Marker Position	Africa Start-End	Africa Epitopes	Africa Epitope value	Out of Africa Start-End	Out of Africa Epitopes	Out of Africa Epitope value
E	120	110-121	KGSLVTCAKF**T**C	1.088333	110-122	KGSLVTCAKF**A**CS	1.0975
	429	425-431	VGGV**F**NS	1.053	-	-	-
NS3	407	404-409	DFV**I**TT	1.038667	403-409	WDFV**V**TT	1.065571
NS5	449	445-454	RGEC**H**SCVYN	1.0976	445-454	RGEC**Q**SCVYN	1.0886
	519	513-519	LQRLGY**I**	1.061571	513-521	LQRLGY**V**LE	1.071
	569	564-572	TLALA**V**IKY	1.104556	563-572	RALALA**I**IKY	1.0822
	640	635-641	LWLLR**K**P	1.043143	-	-	-
	783	778-789	NAICS**A**VPVDWV	1.115833	778-789	NAICS**S**VPVDWV	1.111667
	812	-	-	-	810-816	ML**V**VWNR	1.07257

Table 2.4 Comparison of predicted Cytotoxic T-lymphocyte (CTL) epitopes of ZIKV

Protein	Marker Position	Epitope Position	Epitope Sequence	Epitope Score	Epitope Sequence	Epitope Score	Epitope Sequence	Epitope Score
			Africa		**Central America & Caribbean**		**Singapore**	
E	120	115	TCAKF**T**CSK	1.0377	TCAKF**A**CSK	0.8211	TCAKF**A**CSK	0.8211
NS3	215	207	EIVREAIK**K**	0.9729	-	-	-	-
	583	575	SVPAEVWT**K**	1.5618	SVPAEVWT**R**	0.9576	SVPAEVWT**R**	0.9317
	583	579	EVWT**K**YGEK	0.8778	EVWT**R**HGEK	0.9441	EVWT**R**YGEK	0.9745
	583	583	**K**YGEKRVLK	1.6245	**R**HGEKRVLK	0.9534	**R**YGEKRVLK	1.5103
NS5	195	189	STMMET**M**ER	1.8339	STMMET**L**ER	1.8748	STMMET**L**ER	1.8748
	195	193	ET**M**ERLQRR	1.3697	-	-	-	-
	280	276	KIIG**R**RIER	1.5154	KIIG**N**RIER	1.5601	KIIG**N**RIER	1.5601
	389	389	**R**KRPRVCTK	0.9097	-	-	-	-
	524,526	522	EM**N**R**A**PGGK	0.7874	-	-	-	-
	569	563	RTLALA**V**IK	1.5191	RALALA**I**IK	1.141	RALALA**I**IK	1.141
	569	564	TLALA**V**IKY	0.9928	ALALA**I**IKY	1.0791	ALALA**I**IKY	1.0791
	569	568	A**V**IKYTQNK	1.5925	A**I**IKYTQNK	1.3178	A**I**IKYTQNK	1.3178
	586	580	VLRPAE**G**GK	0.766	-	-	-	-
	640	632	MQDLWLLR**K**	1.6658	MQDLWLLR**R**	1.2026	MQDLWLLR**R**	1.2026
			Caribbean		**Africa & Singapore**		**Central America**	
NS5	877	872	RRIIG**E**EEK	0.8783	RRIIG**D**EEK	0.8841	RRIIG**D**EEK	0.8841
	877	873	RIIG**E**EEKY	0.756	RIIG**D**EEKY	0.7985	RIIG**D**EEKY	0.7985

Supplementary Table S2.1 List of predicted functional protein domains

Protein	PSSM-ID	From (AA)	To (AA)	E-Value	Bitscore	Accession	Short name	Funtion & Characteristic
ancC	307236	5	118	6.49E-28	98.1671	cl03064	Flavi_capsid	Multiple copies of the C protein form the nucleocapsid, which contains the ssRNA molecule.
preM	307624	5	92	1.04E-26	96.3089	cl03269	Flavi_propep	The genome encodes one large ORF a polyprotein which undergos proteolytic processing into mature viral peptide chains. This family consists of a propeptide region of approximately 90 amino acid length.
	307237	95	168	1.67E-20	80.0367	cl03065	Flavi_M	The envelope glycoprotein M is made as a precursor, called prM. The precursor portion of the protein is the signal peptide for the proteins entry into the membrane. prM is cleaved to form M in a late-stage cleavage event. Associated with this cleavage is a change in the infectivity and fusion activity of the virus.
E	279241	2	293	5.75E-138	399.43	cl02995	Flavi_glycoprot	Flavivirus glycoprotein, central and dimerisation domains; Flavivirus glycoprotein, central and dimerization domains.
	213392	302	294	3.35E-49	163.63	cd12149	Flavi_E_C	The C-terminal domain (domain III) of Flavivirus glycoprotein E appears to be involved in low-affinity interactions with negatively charged glycoaminoglycans on the host cell surface. Domain III may also play a role in interactions with alpha-v-beta-3 integrins in West Nile virus, Japanese encephalitis virus, and Dengue virus. The interface between domain I and domain III appears to be destabilized by the low-pH environment of the endosome, and domain III may play a vital role in the conformational changes of envelope glycoprotein E that follow the clathrin-mediated endocytosis of viral particles and are a prerequisite to membrane fusion.
	213897	399	495	2.21E-48	161.652	TIGR04240	flavi_E_stem	This model describes the C-terminal domain, containing a stem region followed by two transmembrane anchor domains, of the envelope protein E. This protein is cleaved from the large flavivirus

								polyprotein, which yields three structural and seven nonstructural proteins.
NS1	279316	3	352	3.62E-170	477.611	cl03032	Flavi_NS1	It contains 12 cysteines, and undergoes glycosylation in a similar manner to other NS proteins. Mutational analysis has strongly implied a role for NS1 in the early stages of RNA replication.
NS2A	279359	12	161	9.16E-15	70.2556	cl03066	Flavi_NS2A	NS2A is a hydrophobic protein about 25 kDa is size. NS2A is cleaved from NS1 by a membrane bound host protease. NS2A has been found to associate with the dsRNA within the vesicle packages. It has also been found that NS2A associates with the known replicase components and so NS2A has been postulated to be part of this replicase complex.
NS2B	279357	5	130	7.55E-31	106.628	cl03063	Flavi_NS2B	This is cleaved into three structural and seven non-structural proteins. All, but two, are cleaved by the NS2B-NS3 protease complex.
NS3	307204	17	167	1.29E-100	118.697	pfam00949	Peptidase_S7	Processing of the polyprotein precursor into mature proteins is carried out by the host signal peptidase and by NS3 serine protease, which requires NS2B (pfam01002) as a cofactor.
	284962	186	331	1.52E-31	301.946	pfam07652	Flavi_DEAD	-
	238034	357	473	9.91E-08	51.0845	cd00079	HELICc	Members of the P-loop NTPase domain superfamily are characterized by a conserved nucleotide phosphate-binding motif, also referred to as the Walker A motif, and the Walker B motif. The Walker A and B motifs bind the beta-gamma phosphate moiety of the bound nucleotide and the Mg2+ cation, respectively. The P-loop NTPases are involved in diverse cellular functions, and they can be divided into two major structural classes: the KG (kinase-GTPase) class which includes Ras-like GTPases and its circularly permutated YlqF-like; and the additional strand catalytic E class which includes ATPase Binding Cassette, DExD/H-like helicases, 4Fe-4S iron sulfur cluster binding proteins of NifH family, RecA-like F1-ATPases, and ATPases Associated with a wide variety of Activities. Also included are a diverse set of nucleotide/nucleoside kinase families.
NS4A	279666	6	127	7.31E-25	91.8876	cl03176	Flavi_NS4A	NS4A contains multiple hydrophobic potential membrane spanning regions. NS4A has only been found in cells infected by Kunjin virus.
NS4B	279665	1	243	2.30E-70	216.047	cl03175	Flavi_NS4B	NS4B contains multiple hydrophobic potential membrane spanning regions. NS4B may form membrane components of the viral

								replication complex and could be involved in membrane localization of NS3 and pfam00972.
NS5	307718	55	226	1.16E-16	78.3902	pfam01728	FtsJ	This family consists of FtsJ from various bacterial and archaeal sources FtsJ is a methyltransferase, but actually has no effect on cell division. FtsJ's substrate is the 23S rRNA. The 1.5 A crystal structure of FtsJ in complex with its cofactor S-adenosylmethionine revealed that FtsJ has a methyltransferase fold. This family also includes the N terminus of flaviviral NS5 protein. It has been hypothesized that the N-terminal domain of NS5 is a methyltransferase involved in viral RNA capping.
	279336	254	891	0	884.071	cl03045	Flavi_NS5	This RNA-directed RNA polymerase possesses a number of short regions and motifs homologous to other RNA-directed RNA polymerases

Supplementary Table S2.2 Africa group specific TAAS analysis result

Gene	Position Nuc	AF CODON	Others CODON	Position AA	AF AA	Others AA
ancC	166	ACA	ACG	56	T	T
ancC	298	AGG	AAG	100	R	K
ancC	325	ATT,GTC	GTT	109	IV	V
preM	7	ATC,ATT	GTC	3	I	V
preM	46	AGG	AGA	16	R	R
preM	67	ATC,ATT	ATA	23	I	I
preM	76	GCC,GCT,GTT	CCA	26	AV	P
preM	94	AAC	AAT	32	N	N
preM	100	TGC	TGT	34	C	C
preM	121	CTC	CTT	41	L	L
preM	124	GGG	GGA	42	G	G
preM	136	GAC	GAT	46	D	D
preM	148	AGT	AGC	50	S	S
preM	175	GGA	GGG	59	G	G
preM	268	CGA	AGA	90	R	R
preM	304	ACA,ACG	ACC,ACT	102	T	T
preM	337	CTA,TTA	CTG,TTG	113	L	L
preM	370	AAG	AGA	124	K	R
preM	397	CCC	CCT	133	P	P
preM	400	GGG	GGC,GGT	134	G	G
preM	415	GCC,GCT	GCA	139	A	A
preM	418	GTA,GTT	GCT	140	V	A
preM	427	GCC	GCG,GCT	143	A	A
preM	445	TCG	TCA	149	S	S
E	25	AGA	AGG	9	R	R
E	31	TTC	TTT	11	F	F
E	40	GGC	GGT	14	G	G
E	97	GTG,GTT	GCA,GTA	33	V	AV
E	118	ACA	ACT	40	T	T
E	139	ACG	ACA,TCA	47	T	ST
E	160	GCC	GCA,GCG	54	A	A
E	283	ACA	ACG	95	T	T
E	310	GGG,GGT	GGA	104	G	G
E	346	TGT	TGC	116	C	C

E	358	ACA,ACG	GCA,GCT	120	T	A
E	394	CCG	CCA	132	P	P
E	427	GTG	GTC,GTT	143	V	V
E	457	GAA	GAG	153	E	E
E	472	GTC	GTT	158	V	V
E	478	GTC,GTT	ATA	160	V	I
E	499	GCA,GCG	GCC	167	A	A
E	505	GCA	GCC	169	A	A
E	547	CCA	CCG	183	P	P
E	625	TTT	TTC	209	F	F
E	634	ATC	ATT	212	I	I
E	715	GCC	GCA	239	A	A
E	718	CAC	CAT	240	H	H
E	724	AAG	AAA	242	K	K
E	844	TGC	TGT	282	C	C
E	850	CTA,TTA	CTG,TTG	284	L	L
E	862	AAG	AAA	288	K	K
E	886	TAT	TAC	296	Y	Y
E	922	GTC,GTT	ATC	308	V	I
E	925	CCA	CCG	309	P	P
E	943	GGA	GGG	315	G	G
E	1150	GAC	GAA,GAG,GAT	384	D	DE
E	1156	AAA	AAG	386	K	K
E	1282	GTG	GCT,GTT	428	V	AV
E	1285	TTC,TTT	CTC,CTT	429	F	L
E	1306	ATT,GTT	ATC	436	IV	I
E	1309	CAC	CAT	437	H	H
E	1312	CAG	CAA	438	Q	Q
E	1366	CAG	CAA	456	Q	Q
E	1378	GGC	GGA,GGG	460	G	G
E	1423	ATC	ATT	475	I	I
E	1432	ACA	ACG,ATG,GTG	478	T	MTV
E	1465	CTC,CTT	CTA,TTA,TTG	489	L	L
E	1471	ACG	ACA	491	T	T
NS1	154	GAG	GAT	52	E	D
NS1	205	AAA	AGA,AGC	69	K	RS
NS1	235	CTA,TTA	CTG,TTG	79	L	L
NS1	238	GAG	GAA	80	E	E

NS1	259	ACA	ACG	87	T	T
NS1	262	GTT	GTC	88	V	V
NS1	313	CCA	CCC,CCT	105	P	P
NS1	403	GTC	GTG	135	V	V
NS1	427	TGT	TGC	143	C	C
NS1	451	AAT	AAC	151	N	N
NS1	484	ATC,GTC	GTA	162	IV	V
NS1	493	ACC	ACT	165	T	T
NS1	538	GAC	GAT	180	D	D
NS1	682	TCT	TCC	228	S	S
NS1	745	GGT	GGA,GGG	249	G	G
NS1	781	AGA	AGG	261	R	R
NS1	865	GAG	GAA	289	E	E
NS1	877	ACT	ACA,ACG	293	T	T
NS1	919	GTC	GTG	307	V	V
NS1	937	TGT	TCC,TGC	313	C	CS
NS1	979	GAC	GAT	327	D	D
NS1	985	TGC	TGT	329	C	C
NS1	1036	GTG	GTA	346	V	V
NS1	1051	ACA	ACC,ACT,AGT	351	T	ST
NS2A	7	ACC	ACT	3	T	T
NS2A	49	CTA	CTG,TTG	17	L	L
NS2A	100	ATG	ACA,ATA,GTA	34	M	ITV
NS2A	172	GTG	GCA,GTA	58	V	AV
NS2A	196	GCA	GCG,GCT	66	A	A
NS2A	226	CAC	CAT	76	H	H
NS2A	232	GCA	GCG	78	A	A
NS2A	235	TTG	CTG	79	L	L
NS2A	247	TTT	TTC	83	F	F
NS2A	262	GCC	GCG	88	A	A
NS2A	319	CTA	CTG,TTG	107	L	L
NS2A	328	GCT	GCC	110	A	A
NS2A	358	GCT	GCC	120	A	A
NS2A	361	CTC,CTT	CTG,TTG	121	L	L
NS2A	379	GTC	GTT	127	V	V
NS2A	385	ATT,GTC,GTT	ATC	129	IV	I
NS2A	391	GGA	AGT,GGG,GGT	131	G	GS
NS2A	421	GCA	GCG,GCT	141	A	A

NS2A	427	GCC	GCT,GTT	143	A	AV
NS2A	430	GTG	GTC,GTT	144	V	V
NS2A	451	GCT	ACC	151	A	T
NS2A	481	CTA,TTA	CTG	161	L	L
NS2A	505	GCA	GCG	169	A	A
NS2A	514	GCG	GCA,GCT	172	A	A
NS2A	520	CTC,CTG	CTT	174	L	L
NS2A	532	GGA	GGG	178	G	G
NS2A	550	TCC	TCT	184	S	S
NS2A	556	AAA	AAG	186	K	K
NS2A	583	CTG	CTA,TTA	195	L	L
NS2A	607	TTG	CTA,CTG	203	L	L
NS2A	622	ATA,GTA	CTA,CTG	208	IV	L
NS2A	625	GTA,GTG	GTC,GTT	209	V	V
NS2A	637	AAT	AAC	213	N	N
NS2A	649	CTA	CTG	217	L	L
NS2B	61	TTT	TTC	21	F	F
NS2B	73	GAC	GAT	25	D	D
NS2B	103	GTA,GTG	GTC,GTT	35	V	V
NS2B	190	GAC	GAT	64	D	D
NS2B	265	GAA,GAG	GAT	89	E	D
NS2B	349	GCT	GCA	117	A	A
NS2B	352	GCA	GCT	118	A	A
NS3	7	GCC	GCG,GCT	3	A	A
NS3	10	CTC	CTA	4	L	L
NS3	31	AAA	AAG	11	K	K
NS3	46	GGA	GGG	16	G	G
NS3	130	GGA	GGG	44	G	G
NS3	166	GCC,GCT	TCC	56	A	S
NS3	208	GGG	GGA	70	G	G
NS3	223	GAC	GAT	75	D	D
NS3	241	GGG	GGC,GGT	81	G	G
NS3	253	TTG	CTA	85	L	L
NS3	259	GCA	GCC	87	A	A
NS3	274	CTA,CTC,CTT	CAC,CAT	92	L	H
NS3	298	GTA	GTG	100	V	V
NS3	313	AGG	AGA	105	R	R
NS3	316	GCC,GCT	GAG,GCG	106	A	AE

NS3	319	AAA,AGA	AGG	107	KR	R
NS3	325	ATT	ATC	109	I	I
NS3	391	CCC,CCT	CCA	131	P	P
NS3	424	AAA	AAG	142	K	K
NS3	430	GGA	GGG	144	G	G
NS3	460	GTT	GTC	154	V	V
NS3	493	ATA	ATC	165	I	I
NS3	499	CAG	CAA	167	Q	Q
NS3	502	GGA	GGG	168	G	G
NS3	514	GAG	GAA	172	E	E
NS3	523	CCA,CCG	CCT	175	P	P
NS3	586	CCA	CCT	196	P	P
NS3	643	AAG	ACA,ACG,GCA	215	K	AT
NS3	655	ACA	ACC,ACG,ACT	219	T	T
NS3	691	GAG	GAA	231	E	E
NS3	706	CTA,TTG	CTT	236	L	L
NS3	712	GGA	GGG	238	G	G
NS3	745	AAC	AAT	249	N	N
NS3	778	TTG	CTA,TTA	260	L	L
NS3	886	GCT	GCA	296	A	A
NS3	916	GAA	GAG	306	E	E
NS3	970	GAT	GAC	324	D	D
NS3	1099	AGA	AGG	367	R	R
NS3	1105	GGA	GGC	369	G	G
NS3	1111	GAA	GAG	371	E	E
NS3	1141	AAG	AAA	381	K	K
NS3	1162	AGG	AGA	388	R	R
NS3	1183	TTT	TTC	395	F	F
NS3	1219	ATA	GTA,GTG	407	I	V
NS3	1294	CCA	CCG	432	P	P
NS3	1321	ATC	ATT	441	I	I
NS3	1330	GGG	GGA	444	G	G
NS3	1345	ACG	ACA	449	T	T
NS3	1354	AGT	AGC	452	S	S
NS3	1369	AGA	AGG	457	R	R
NS3	1375	CGT	CGC	459	R	R
NS3	1429	TGT	TGC	477	C	C
NS3	1498	CAG	CAA	500	Q	Q

NS3	1546	GCC,GCT	GCA	516	A	A
NS3	1570	CTG	CTT	524	L	L
NS3	1576	ACA	ACG	526	T	T
NS3	1609	AAG	AAA,AGA	537	K	KR
NS3	1624	CCC	CCG,CCT	542	P	P
NS3	1726	GTA	GTG	576	V	V
NS3	1729	CCA	CCG	577	P	P
NS3	1747	AAA,AAG	AGA	583	K	R
NS3	1750	TAT	CAC,TAC	584	Y	HY
NS3	1759	AAG	AAA,AGA	587	K	KR
NS3	1792	AGG	AGA	598	R	R
NS4A	10	TTA,TTG	CTT,TTT	4	L	FL
NS4A	16	GTA	GTG	6	V	V
NS4A	22	GAG,GAT	GAA	8	DE	E
NS4A	127	GCA	GCC	43	A	A
NS4A	142	CTG	ATG,TTA,TTG	48	L	LM
NS4A	190	GTT	GTC	64	V	V
NS4A	232	ATC	ATA,ATG	78	I	IM
NS4A	256	GTA	GTG	86	V	V
NS4A	301	GAA	GAG	101	E	E
NS4A	340	CTG,TTA,TTG	CTA	114	L	L
2k	37	GTG	GTA	13	V	V
2k	46	GGC	GGT	16	G	G
2k	55	GGT	GGC	19	G	G
2k	67	GCA	GCC	23	A	A
NS4B	16	CTG	TTG	6	L	L
NS4B	37	ATA	CTA,TTA	13	I	L
NS4B	40	GCT	AGC,GGC	14	A	GS
NS4B	61	GAA	GAG	21	E	E
NS4B	97	GAT	GAC	33	D	D
NS4B	112	TCC	TCA,TCG	38	S	S
NS4B	133	GCA	GCC,GCT	45	A	A
NS4B	214	ACA	ACG	72	T	T
NS4B	229	CTA,CTG	TTG	77	L	L
NS4B	325	CTG,TTG	CTA,TTA	109	L	L
NS4B	346	CTT	CTC	116	L	L
NS4B	415	ACA	ACG	139	T	T
NS4B	436	AAT	AAC	146	N	N

NS4B	439	CCC	CCT	147	P	P
NS4B	502	AAG	AAA	168	K	K
NS4B	520	TTA	CTA	174	L	L
NS4B	550	GTG	ATA,GTT	184	V	IV
NS4B	571	GGA	GGG	191	G	G
NS4B	601	GCA	GCC,GCG,GCT	201	A	A
NS4B	631	CCA	CCG,CCT	211	P	P
NS4B	682	AGA	AGG	228	R	R
NS4B	697	GCA	GCT	233	A	A
NS5	31	AAG	AAA	11	K	K
NS5	37	AAA	AAG	13	K	K
NS5	76	TCT	TCC	26	S	S
NS5	112	GAG	GAA	38	E	E
NS5	145	GCC	GCA	49	A	A
NS5	154	GGA	GGC	52	G	G
NS5	163	GTA	GTG	55	V	V
NS5	169	CGG	CGA	57	R	R
NS5	229	GTT	GTC	77	V	V
NS5	235	GAC	GAT	79	D	D
NS5	265	TAT	TAC	89	Y	Y
NS5	289	GTG	GTT	97	V	V
NS5	319	GGT	GGC	107	G	G
NS5	343	CTG	TTG	115	L	L
NS5	385	GGA	GGG	129	G	G
NS5	472	GAG	GAA	158	E	E
NS5	478	CGA	CGG	160	R	R
NS5	481	ACA	ACG	161	T	T
NS5	490	GTG	GTA,GTC,GTT	164	V	V
NS5	535	TTC	TTT	179	F	F
NS5	550	CTG	TTG	184	L	L
NS5	583	ATG	CTA,CTG,TTG	195	M	L
NS5	595	CAA	CAG	199	Q	Q
NS5	631	TTG	CTA,CTC,CTT	211	L	L
NS5	706	ACA	ACG	236	T	T
NS5	709	AGT	AGC	237	S	S
NS5	733	GAA,GAT	GAC	245	DE	D
NS5	772	AAC	AAT,GAT	258	N	DN
NS5	781	TCA,TCG	TCC,TCT	261	S	S

NS5	787	ACA	ACG	263	T	T
NS5	838	AAG,AGG	AAC,GAC	280	KR	DN
NS5	850	AGA	AGG	284	R	R
NS5	862	GAA	GAG	288	E	E
NS5	868	GCA	GCG	290	A	A
NS5	874	ACA	ACG	292	T	T
NS5	913	GCC	GCT	305	A	A
NS5	931	GAA	GAG,GAT,GTG	311	E	DEV
NS5	958	TCC	TCT	320	S	S
NS5	964	GTG	ATA,GTA	322	V	IV
NS5	976	GTT	GTC	326	V	V
NS5	979	AGA	AGG	327	R	R
NS5	1045	CCA	CCG	349	P	P
NS5	1054	CAA	CAG	352	Q	Q
NS5	1114	CGC	CGT	372	R	R
NS5	1156	CTG,TTG	CTA,TTA	386	L	L
NS5	1159	GGA,GGG	GGC,GGT	387	G	G
NS5	1165	CGC	CAC,CAT,TAC	389	R	HY
NS5	1168	AAG	AAA	390	K	K
NS5	1177	CGC,CGT	CGA	393	R	R
NS5	1231	GGA	GGG	411	G	G
NS5	1264	ACA,ACG	ACC,ACT	422	T	T
NS5	1267	GCC,GCT	GCA	423	A	A
NS5	1294	TTT	TTC	432	F	F
NS5	1321	GAA	GAG	441	E	E
NS5	1345	CAC,CAT	CAG	449	H	Q
NS5	1390	GGA	GGG	464	G	G
NS5	1396	TTC	TTT	466	F	F
NS5	1402	AAA	AAG	468	K	K
NS5	1405	GCA	GCC	469	A	A
NS5	1438	TTG	CTA,CTG	480	L	L
NS5	1444	GCC	GCT	482	A	A
NS5	1453	CTG,TTG	CTA,TTA	485	L	L
NS5	1486	GAC	GAT	496	D	D
NS5	1504	GAA	GAG	502	E	E
NS5	1537	CTG,TTG	CTA,TTA	513	L	L
NS5	1555	ATC,ATT	GTC,GTT	519	I	V
NS5	1570	AAC,AAT	AGC,AGT	524	N	S

NS5	1573	CGG	CGC,CGT,TGC	525	R	CR
NS5	1576	GCA,GCG	ACA,ATA	526	A	IT
NS5	1678	GAA	AAA	560	E	K
NS5	1687	AGA	AGG	563	R	R
NS5	1693	CTG	TTA,TTG	565	L	L
NS5	1705	GTG	ATA	569	V	I
NS5	1708	ATT	ATC	570	I	I
NS5	1741	CTC	CTT	581	L	L
NS5	1756	GGA	AAA,AGA	586	G	KR
NS5	1795	CAG	CAA	599	Q	Q
NS5	1822	TAT	TAC	608	Y	Y
NS5	1858	CTT	CTC	620	L	L
NS5	1885	GTG	GTC,GTT	629	V	V
NS5	1918	AAG	AGG	640	K	R
NS5	1936	AGA	AAC,AGC	646	R	NS
NS5	1963	AGA	AGG	655	R	R
NS5	2068	AAA	AAG	690	K	K
NS5	2092	TCG	TCA,TCT	698	S	S
NS5	2107	AAT	AAC	703	N	N
NS5	2119	GTC	GTT	707	V	V
NS5	2149	CTG	CTC,CTT	717	L	L
NS5	2167	AGA	AGG	723	R	R
NS5	2260	TGT	TGC	754	C	C
NS5	2311	AGA	AGG	771	R	R
NS5	2344	TCG	TCA	782	S	S
NS5	2347	GCC,GCT	TCT	783	A	S
NS5	2365	GTA,GTC	GTT	789	V	V
NS5	2431	CTC	CTT	811	L	L
NS5	2434	ATG	GCG,GTG	812	M	AV
NS5	2488	CCT	CCA,CTA	830	P	LP
NS5	2530	GAG	GAA	844	E	E
NS5	2548	TCC	TCT	850	S	S
NS5	2566	CCC,CCT	CCA,CCG	856	P	P
NS5	2680	GAG	GAA	894	E	E

Supplementary Table S2.3 Singapore group specific TAAS analysis result

Gene	Position Nuc	SG CODON	Others CODON	Position AA	SG_AA	Others_AA
E	259	GAT	GAC	87	D	D
E	325	GGA,GGG	GGC	109	G	G
E	427	GTC	GTG,GTT	143	V	V
NS1	310	CTG	TTG	104	L	L
NS1	541	CCT	CCA,TCA	181	P	PS
NS2A	400	CTG	TTG	134	L	L
NS2A	622	CTA	ATA,CTG,GTA	208	L	ILV
NS4A	40	CCG	CCA	14	P	P
NS5	643	TCA	TCC,TCT	215	S	S

Supplementary Table S2.4 Central-America group specific TAAS analysis result

Gene	Position Nuc	CA CODON	Others CODON	Position_AA	CA_AA	Others_AA
NS1	298	GCT	GGC,GGT	100	A	G
NS2A	316	CTA	CTG	106	L	L
NS3	580	TTA	CTG,TTG	194	L	L
NS3	1714	CTG	ATG	572	L	M

Supplementary Table S2.5 Caribbean group specific TAAS analysis result

Gene	Position Nuc	CB CODON	Others CODON	Position_AA	CB_AA	Others_AA
NS5	2629	GAG	GAT	877	E	D

Figure 2.1 Principal Component analysis of ZIKV

(a) To determining the number of clusters the NbClust package in R was used. The D index, result shows a significant knee (the significant peak in Dindex second differences plot) 4, number of clusters. (b) The PCA result was used 301 ZIKV dataset and the number of clusters was set to k = 4 based on the D index. And the names of clusters were based on the isolation regions. (c) The D index also shows significant knee 5, number of clusters. However, the dataset was used only the 'Mixed cluster' in (b). (d) The PCA result was used 170 ZIKV dataset making the mixed cluster in (b) and the number of clusters was set to k = 5 based on D index. The 'Caribbean', 'China', 'Central-America' group shows distinct clusters compared to the 'French_polynesia' and 'Mixed-America' group.

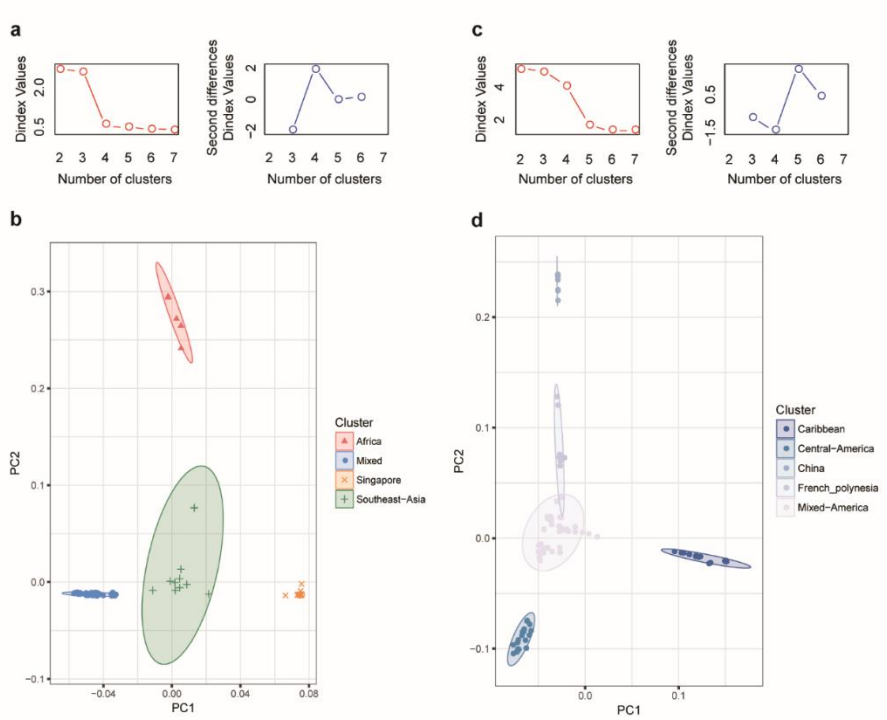


Figure 2.2 Phylogenetic analysis of ZIKV by Maximum Likelihood method

Phylogenetic tree constructed with nucleotide data from 301 viral complete coding sequences of Zika virus by the maximum likelihood method in MEGA7 base on the Tamura-Nei model. The grouping bar also was colored according to results of PCA clusters.

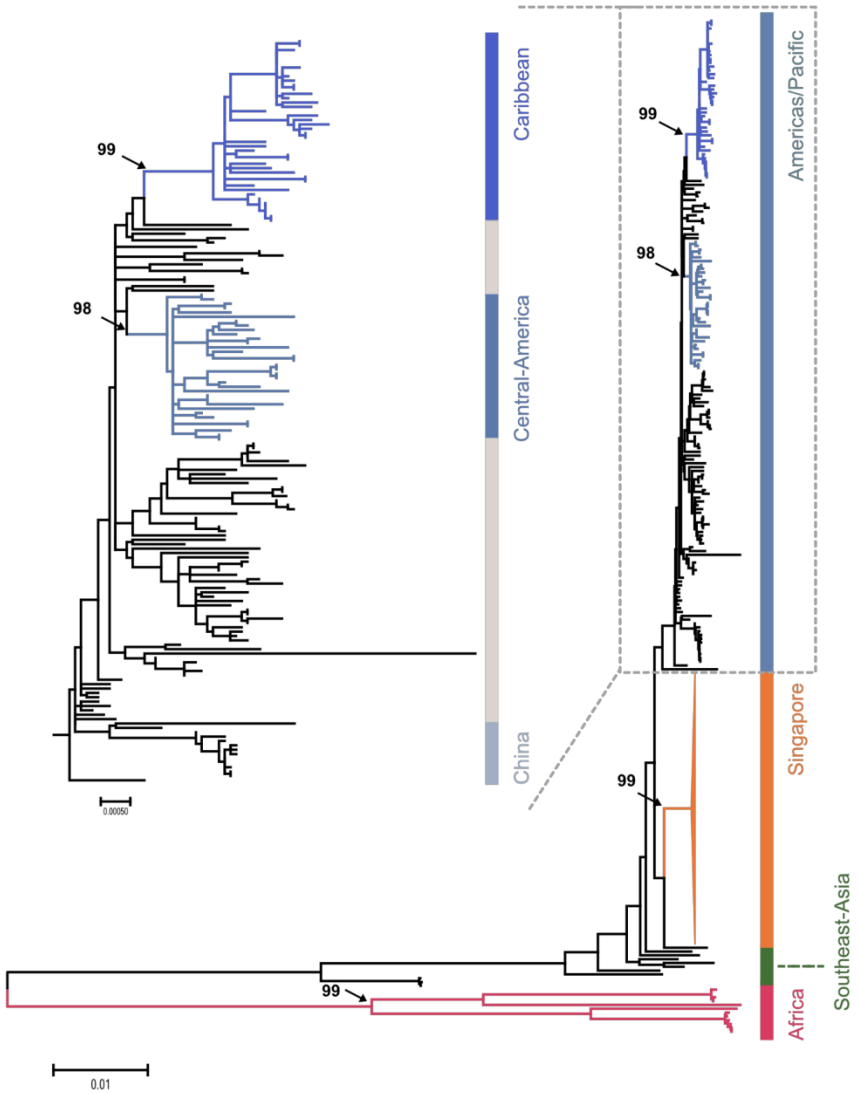


Figure 2.3 Local group-specific significant Nucleotide position on ZIKV

The bars represent statistically different (Bonferroni adjusted p-value) all nucleotide coding sequence positions between the manually selected target group and other groups based on the isolated region. The red bars are statistically most different and mutually exclusive positions in the target group.

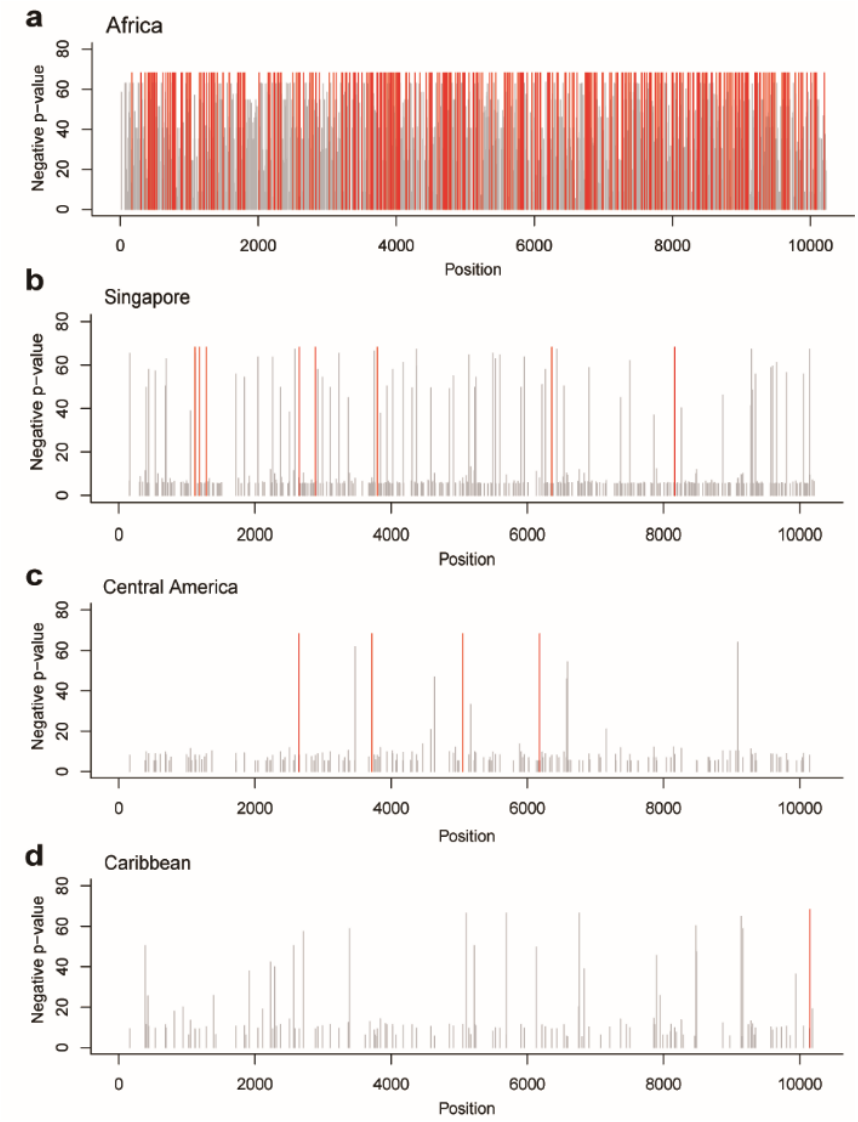
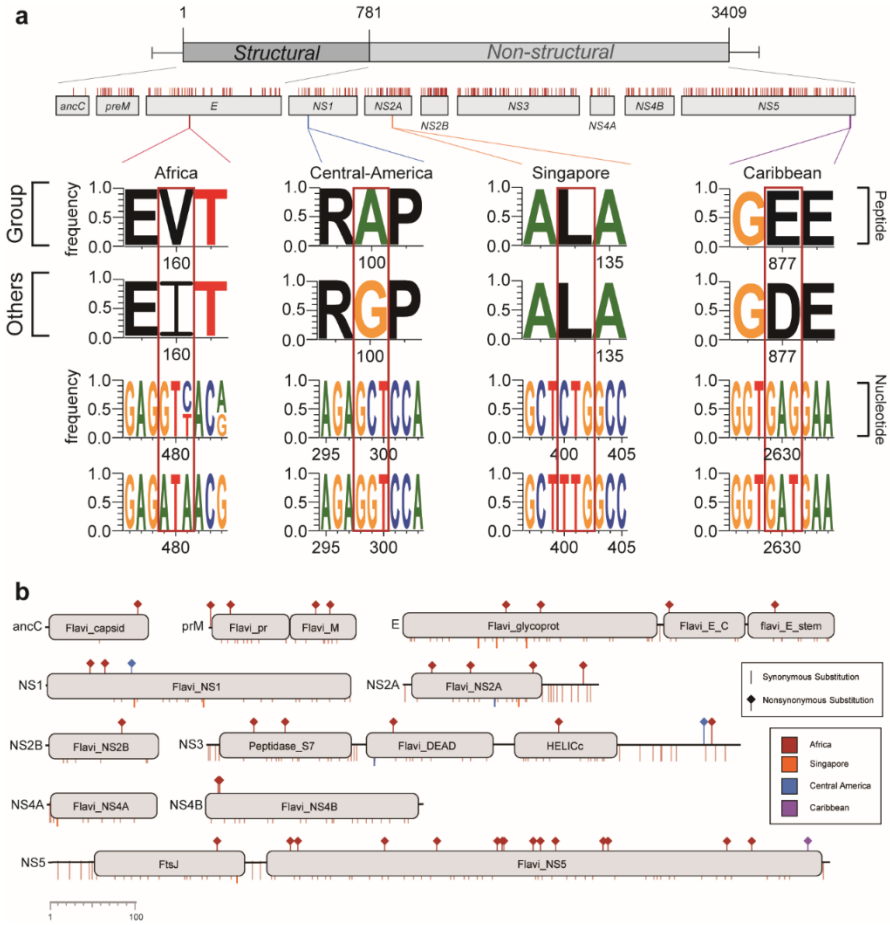


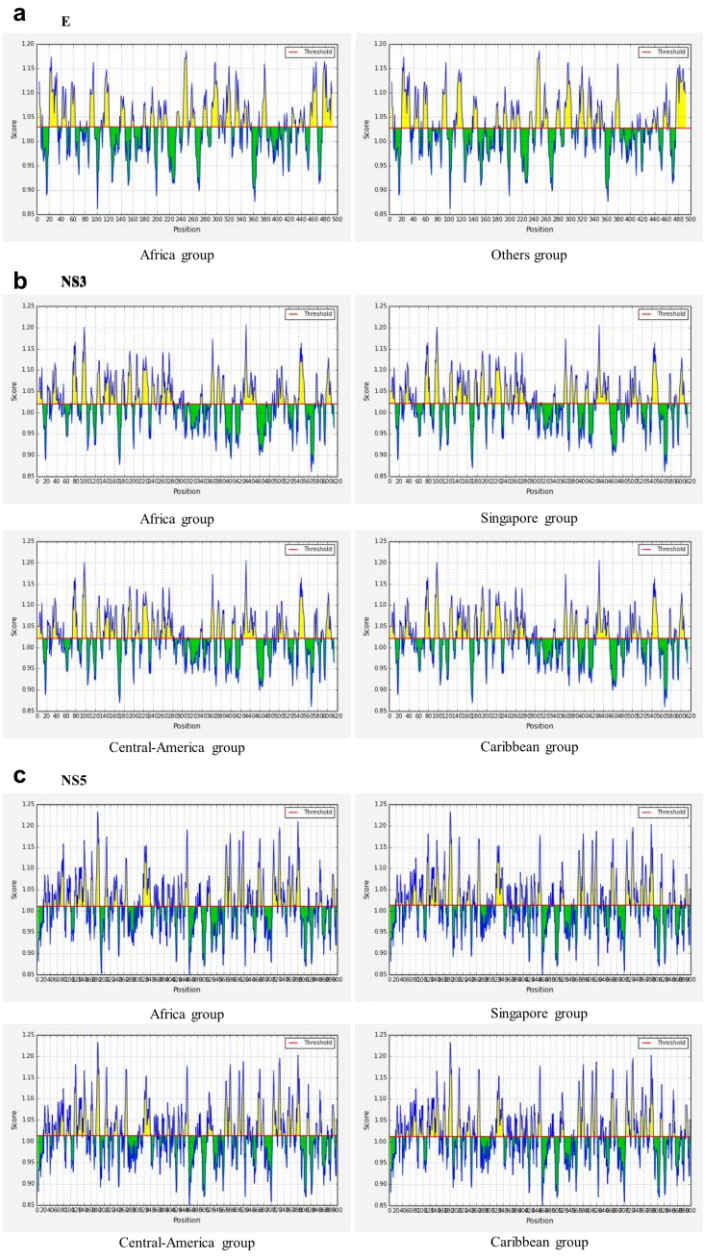
Figure 2.4 Main Site Logo of each group and Protein domains of ZIKV

(a) The statistically different positions were indicated on each protein. One of group specific significant nonsynonymous and synonymous substitution sites indicated as sequence logo. (b) Representative domains on each three structural protein and seven non-structural protein with synonymous substitution positions (no-pin and under the domains) and nonsynonymous substitutions regions (pin).



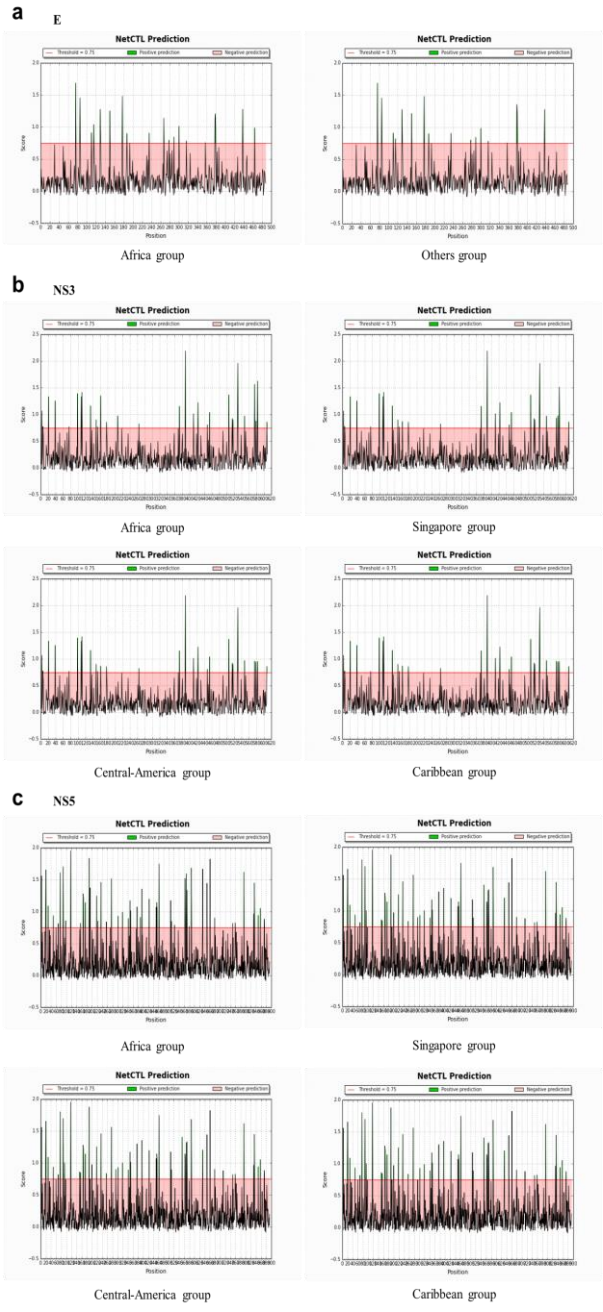
Supplementary Figure S2.1

Group specific antigenic B-cell epitope prediction results on E, NS3 and NS5 protein of ZIKV by 'Kolarskar and Tongaonkar' method.



Supplementary Figure S2.2

Group specific cytotoxic T-lymphocyte (CTL) epitope prediction on E, NS3 and NS5 protein of ZIKV using NetCTL.1.2 server.



Chapter 3. Local genetic markers clustered by coding sequences of Influenza D virus

3.1 Abstract

Influenza D virus (IDV) has recently been classified as a new genus of the *Orthomyxoviridae*. The symptoms of IDV infection, mainly found in cattle, are mild, but they play a key role in causing respiratory diseases. Moreover, IDV has the potential of infecting humans which makes the research of the IDV necessary. There have been studies identifying evolutionary relationships between IDV strains, and these studies have focused on specific gene segments, but using concatenated genome sequence data comparative genomic research was not done sufficiently. In this study, we analyzed all of the available IDV complete coding sequences from the National Center for Biotechnology Information (NCBI) database to reveal the evolutionary relationship and identify novel genetic markers. Using the results of phylogeny and principal coordinate analysis (PCoA), we compared concatenated whole coding sequence dataset and each of gene coding sequence dataset. We confirmed that concatenated dataset results were more appropriately clustered into four groups with isolated region, and we selected the main three groups (Japan, China and Italy) to process the analytical data. Focusing on the main three groups, we found single nucleotide variants (SNVs) supported by statistical significance and identified the evolutionary accelerated protein-coding regions by the ratio of nonsynonymous to synonymous substitutions (dN/dS) analysis. Our results revealed that there is a relationship between the sequence feature and isolated region. Also, through the dN/dS analysis, we identified evolutionary

accelerated protein-coding regions in each group. Out of the nonsynonymous and synonymous substitution sites on functional domains and predicted B-cell epitopes were suggested as regional genetic markers. The findings in this study may provide a deeper genetic insight of IDV.

3.2 Introduction

In the livestock industry respiratory diseases are one of the most important common causes of morbidity and mortality. The most fatal bovine respiratory viruses, known as bovine herpesvirus 1 (BHeV1) (Muylkens, Thiry et al. 2007), bovine respiratory syncytial virus (BRSV) (Valarcher and Taylor 2007) and bovine viral diarrhea virus (BVDV) (Fray, Paton et al. 2000), and swine respiratory viruses such as porcine reproductive and respiratory syndrome virus (PRRSV) (Pileri and Mateu 2016) and swine influenza virus (SIV) (Neumann, Noda et al. 2009), cause huge coastal industrial damage. Especially, the influenza viruses are one of the most deadly and commonly affecting viruses in the livestock industry. Among the various influenza types, the influenza D virus (IDV) was first isolate from pig in the Mid-west of the USA in 2011 as well as also detected in the cattle in various region of Mexico, Asia (Japan and China), and Europe (Italy and French) (Jiang, Wang et al. 2014, Ferguson, Eckard et al. 2015, Chiapponi, Faccini et al. 2016, Murakami, Endoh et al. 2016). Recently, it was classified as a new type of influenza virus with a single species by International Committee on Taxonomy of Viruses (https://talk.ictvonline.org/ictv-reports/ictv_online_report/).

IDV is a member of *Orthomyxoviridae* family which are negative sense, single-stranded and segmented RNA viruses. According to previous research, a limited serological survey detected that not only the cattle and pigs but also goats have susceptibility to IDV. However, the susceptibility of IDV to other

livestock, poultry and humans has not been known yet; through in some studies, a low percentage of human blood vessels have measurable titers of specific antibodies to IDV (Hause, Huntimer et al. 2017). Due to these potential and threats, several studies have been conducted with IDV. Phylogenetic and serological analyses were performed on each protein segments, and IDV strains isolated from some regions were classified into D/OK, D/660 and other third lineage (Collin, Sheng et al. 2015). While previous studies carried out clustering and evolutionary analysis based on the specific gene segment (Yan, Wang et al. 2018), genomic differences between local strains or comparison based on the concatenated full coding sequence have not been studied.

In this study, we tried to reveal novel gnomic markers of IDV strains clustered ty phylogenetic analysis and PCoA based on the IDV compete coding sequences isolated from various countries. Next, we examined for illuminated candidate genomic markers specific to major groups in each cluster supported by the natural episodic adaptations, the functional domain search, and the epitope prediction analysis. Through this study, we will provide a deeper insight into understanding the molecular mechanisms of the new influenza virus type D.

3.3 Materials and methods

3.3.1 Collection of complete coding sequence of IDV

We collected IDV coding genome sequences available on NCBI (https://www.ncbi.nlm.nih.gov/, release date: Sep 29th, 2018) with their isolation information, such as isolation time, region and host species. Total of 26 IDV strains was collected and only one IDV strain was filtered out because of the duplication. These 25 IDV strains were composed of three Japan (Miyazaki, Ibaraki, Yamagata each), five Italy (NA), three China (Shandong), 10 USA (four from Kansas, two from Oklahoma, one Texas, one Mississippi, one Minnesota and one Nebraska), and three Mexico (NA) (Table 3.1).

3.3.2 Phylogenetic tree construction and Principal Coordinate Analysis of IDV

Phylogenetic analysis using IDV coding sequences were conducted with each set of gene segment and concatenated genome sequence set. To match homologous sites of the strains, we performed multiple sequence alignment for eight gene coding sequence sets of 25 strains of IDV (PB2, PB1, P3, HEF, NP, P42, NS1 and NS2) using the PRANK program (v.140110) (Löytynoja 2014). Especially, the NS gene segment produces the two different non-structural proteins NS1 and NS2 by splicing. Finally, a total of 9 sets composed of eight gene segment sets and one concatenated set were constructed using MEGA7 program. Phylogenetic reconstruction was completed using maximum likelihood method with a bootstrap value of 1000 based on the Tamura-Nei model (Kumar, Stecher et al. 2016).

To understand the relationships between the aligned IDV strains, principal coordinate analysis (PCoA) was performed by using the two different data sets: the aligned HEF protein- coding sequence set and the aligned concatenated all of the protein-coding sequence set. The identical sequence matrix was obtained from the aligned sequences by BioEdit (v.7.0.5) (Hall 2005). From the matrix, the Past3 (v.3.22) (Hammer, Harper et al. 2001) was conducted, and then the plot results were visualized using R package pplot2 (Wickham 2016).

3.3.3 Comparative genomic analysis (MEAT-CATS) for local specific variants

In order to identify statistically different nucleotide sites in the main group strains (Japan, China and Italy) and others, the metadata-driven Comparative Analysis for Sequences (meta-CATS) tool in IRD was used (Pickett, Liu et al. 2013). The meta-CATS performs a chi-squared goodness of fit test to identify sites with significant sequence variation between the manually divided main group and others, and in parallel with Pearson's chi-square test, calculate the p-value that identifies specific pairs of groups and shows statistically significant maximum probability (default=0.05). In this study, the division of IDV strains as main groups and others is based on the results of phylogenetic analysis and PCoA.

3.3.4 Group specific Nucleotide and Amino Acid substitutions of IDV

To identify not only the statistical single nucleotide substitution but also mutually exclusive substitutions in codon and amino acid, Target-specific

Single Nucleotide Variants (TSNV) and Target-specific Amino Acid substitution (TAAS) analyses were performed (Zhang, Li et al. 2014). Codon and amino acid substitution sites were carried out with the three main group datasets (Japan, China and Italy) based on the results of the phylogenetic analysis and PCoA. Additionally, the visualization of main group-specific conserved amino acid and codon sequence substitution were used with WebLogo3.

3.3.5 dN/dS Analysis for main three group specific positive selection

Nonsynonymous/synonymous ratio (dN/dS, ω) test the most widely used method for identifying specific sites of positive selection. In order to identify an evolutionary accelerated protein in IDV with the ratio of nonsynonymous (dN) to synonymous (dS) substitutions in each protein gene sets, 'branch-site' model of the PAML package (v4.9a) was used (Yang 2007). Based on the prior population comparative studies, three main groups (Japan, China, and Italy) were selected as a foreground branch in which particular lineages influence positive selection.

The details of option for codeml control files of branch-site model set were 'model = 2, NSsites = 2, fix omega = 0, CodonFreq = 2 and cleandata = 1' in the alternative model. From the results, the p-value of each protein gene was calculated by comparing the null and alternative model maximum likelihood using the likelihood ratio test (LTR) 'D= 2 * Δ ℓ' and chi-square distribution. Hence, we identified sites affected positive selection on each lineage that was found based on the BEB inference, as well as the positive selective sites.

3.3.6 Conserved Domain search of IDV

In order to identify whether the sequence variants were located on the functional domain of each protein, the phylogenetically most ancestral and the oldest collected IDV strain was selected (strain: D/bovine/Kansas/1-35/2010) as a query. Overall, the NS coding sequence was divided into NS1 and NS2 coding sequences because the NS gene segment produces two different non-structural proteins by splicing. Total of gene sets consisted of 9 proteins (PB2, PB1, P3, HE, NP, P42, NS1 and NS2). Prior to the domain search analysis, the aligned nucleotide sequences of each protein were translated into amino acid sequences. Moreover, the details of the options were CDSEARCH/CDD (v3.16), Expected value: 0.01, Composition-corrected scoring: Applied, Low-complexity filter: not filtered, Maximum number of hits: 500 (Marchler-Bauer, Lu et al. 2010).

3.3.7 B-cell Antigenic Epitope prediction

B-cell epitope includes characteristics of hydrophilic nature and accessibility for a flexible region of an immunogen (Fieser, Tainer et al. 1987). For the B-cell epitope prediction, immunoinformatic approach analysis, a Kolaskar and Tongaonkar antigenicity scale (Kolaskar and Tongaonkar 1990) method that bas provided 75% experimental accuracy in evaluating the analyzed antigenicity on the basis of the physicochemical properties of amino acids and their abundances in experimentally known epitopes, was used at IEDB (http://tools.iedb.org/bcell/). We targeted the three P3, HEF and P42

proteins, which are identified as positively selected in dN/dS analysis. Furthermore, the ancestral strains of IDV in the main three groups, were used as queries: Japan group: D/bovine/Ibaraki/7768/2016, China group: D/bovine/Shandong/Y127, and Italy group: D/bovine/Italy/1/2014.

3.4 Results

3.4.1 Genome sequences of IDV

We collected complete coding sequences of 26 IDV strains registered in NCBI database (https://www.ncbi.nlm.nih.gov/). For the accuracy of the analysis, we excluded the redundant data, and subsequently, a total of 25 IDV coding sequence data were filtered (Table 3.1). In order to match homologous nucleotide sequence sites, we performed multiple sequence alignments on eight orthologous gene sets (PB2, PB1, P3, HEF, NP, P42, NS1, and NS2). In particular, non-structural proteins are NS protein coding sequence including NS1 (location=29~760) and NS2 (location=29~218, 484~848) protein coding sequences. Thus, we parsed the entire NS protein-coding sequences into the NS1 protein-coding sequence and NS2 protein-coding sequence, respectively. Moreover, the aligned eight genes were concatenated to 12,795 bases.

3.4.2 Comparative Phylogeny and Principal Coordinate Analysis

We constructed the maximum-likelihood (ML) phylogenetic tree based on the nucleotide coding sequences of IDV to investigate the genetic distances and historical relatedness. Total of nine phylogenetic trees based on the ML was constructed with the aligned eight orthologous gene sets and one concatenated set (Figure 3.1). In all the phylogenetic trees being constructed, the strains isolated from Japan (Miyazaki, Yamagata and Ibaraki), China (Shandong), and Italy generally formed distinct clusters in related to their respective geographic distribution. In the PB2, PB1, P3, NP, NS1, NS2 segments and concatenated sets, the Japan group was the first diverged from other strains, whereas in each

HEF and P42 segment sets, all of the IDV strains diverged into two clusters. Typically, the difference between the concatenated set and HEF segment set is the divergence pattern of the Japan and China group. In the HEF tree, IDV strains were isolated from the Americas and divided into two clades, Japan and China, respectively (Figure 3.1 A). In the concatenated tree, however, China group is a quite independently distinct from the Japan group as well as the rest of other strains. Within the larger clade, the rest strains were divided into two sub-clades including the Italy group (Figure 3.1 C).

According to two aspects of the phylogenetic analysis, we performed the Principal Coordinates Analysis (PCoA) based on the identical similarity matrix obtained from the aligned coding sequences using the HEF segment set and concatenated set. In the PCoA result of HEF set, the IDV strains were clustered into three lineages identified in the previous study (Zhai, Zhang et al. 2017): D/660, D/OK and Third lineages. Likely, three lineages were clustered in the PCoA result of concatenated set; however, there were differences between the IDV strains that make up the three lineages. In the HEF set, the IDV strains from the USA were divided into two lineages-'D/660' and 'D/OK'-, whereas the USA strains were grouped together in the concatenated set. In addition, the biggest difference is the distance of the IDV strains from the China and the rest of the IDV strains, such that comparing with the result of the HEF set, not only the China strains formed a cluster distinct from other strains but also the Italy strains formed a dense cluster (Figure 3.1).

3.4.3 Comparative nucleotide and amino acid sequence analysis (meta-CATS, TSNV, TAAS)

The chi-squared goodness of fit test was performed using the meta-CATS tool to identify specific and significant nucleotide sequence variation in the main three groups (Japan, China, and Italy). As an important step for determining the statistical significance, the p-value implies the specificity of each group pair manually divided based on the results of phylogeny and PCoA. In the nucleotide alignments of the IDV datasets, we identified 368 Japan group-specific nucleotide variations, 174 China group-specific nucleotide variations and 105 Italy group-specific variations.

Although the nucleotide sequences are the statistically significant position, it is difficult to explain that not all of the variants are sufficiently conserved and that each group contributes equally to IDV evolution. In the coding region, nucleotide sequences are classified as synonymous and nonsynonymous substitutions depending on the presence or absence of variation in amino acid variation, such that nonsynonymous can cause amino acid variation but synonymous substitutions cannot. Therefore, we performed Target-specific Single Nucleotide variants (TSNV) and Target-specific Amino Acid Substitution (TAAS) analysis to identify each group specific synonymous and nonsynonymous substitution sites. As results of TSNV analysis, we identified 265 Japan group-specific mutually exclusive TSNVs (72.0%, compared with meta-CATS results), containing both of synonymous and nonsynonymous substitutions. Likewise, we also identified 114 China group-specific TSNVs (65.5%) and 20 Italy group-specific TSNVs (19.0%). Next, we performed the TAAS analysis with translated IDV amino acid sequence dataset to identify the group-specific conserved nonsynonymous substitution. Therefore, we identified a total of 66 Japan group-specific TAASs (24.9%, synonymous and

nonsynonymous substitution ratio), China group-specific TAASs (23.6%) and Italy group-specific TAASs (20%).

3.4.4 Evolutionary positive selection analysis

To understand the evolutionary relationship between the IDV strains and estimate positive selection on the group-specific markers, we conducted the dN/dS analysis with the branch-site model of codeml in the PAML package. When the dN/dS analysis was carried out, the most recent common ancestral branches of the three main groups were used as foreground branches under positive selection, whereas the rest branches were used as background branches under neutral selection in the phylogenetic tree, as shown in Table 3.2. When each of the Japan group and Italy group were selected as the foreground branch, P42 and P3 were identified as an evolutionary accelerated protein, respectively. On the basis of posterior probability of detecting positive selected sites, the Japan group had seven markers in P42 and the Italy group had two markers in P3, each with a positive selection from their most common ancestral branch ($D > 0$, ω of foreground branches > 1, BEB > 0.5) (Table 3.2).

3.4.5 Conserved domain search

To reveal the functional domain organization on each gene segment and that the identified main group-specific markers were positioned on the functional domains, we performed NCBI conserved domain search using amino acid sequences. Total of eight gene segment coding sequences was translated into amino acid sequences and used as query. The results of the conserved

domain search are shown in Supplementary Table S3.1. As advanced analyses, the NS gene segment was spliced into the NS1 and NS2 gene segment that shares the identical sites (location: 29~218). Generally, most of results of domain searches showed high concordance with ICV domains and identified as ICV domains. These results caused by the recent IDV becoming a member of a new genus member of Orthomyxoviridae. The Japan group-specific nonsynonymous substitution sites were located in all of the domains as well as the China group-specific nonsynonymous substitution sites were located in all of the domains except in the Flu_NS2 of NS protein whose function is not accurately known. The Italy group-specific nonsynonymous substitution sites were located only in Flu_PA of P3 protein and Hema_esterase of HEF protein.

3.4.6 B-cell epitope prediction analysis

The epitope is an antigenicity of residues in different types of antigenic determinants that are recognized by the immune systems. In this study, we clustered the main three local group (Japan, China, and Italy) and identified their specific genetic markers that evolutionally positive selected and simultaneously located in the functional domain. In addition, to determining if there are B-cell epitope candidates in the positive selected and playing a serologically important role in protein P3, P42 and protein HEF and if the genetic markers were affected antigenicity of epitopes, we conducted B-cell epitope prediction analysis using the Kolaskar & Tongaonkar Antigenicity method. In the HEF protein, containing glycosylation site and receptor-binding domain, 23 Japan group-specific B-cell epitope candidates were identified as well as 21 B-cell epitope candidates in both China and Italy group. In the P3

and P42 protein, which is distinguished as positively selected in dN/dS analysis, 28 and 11 epitope candidates were identified, respectively in all the main three groups, but the details of sequences and positions were different. Table 3.3 shows the predicted main group epitopes including the group-specific conserved variants, which is a genetic marker, and the residues in bold are indicating conserved variants.

3.5 Discussion

The goal of this study was to investigate the impact of the local-specific IDV genetic variations in global regions for a better understanding of their epidemics and to identify markers as epitopes for therapeutic vaccine development. Although there are no human cases of IDV infection yet, it infects the livestock that lives very close to humans such as swine, cattle, and horse. The symptom of IDV is mild, but due to the fact that the virus causes respiratory disease and inflammation in their respiratory tract, it threatens the infected hosts and causes minor problems in productivity when the hosts are livestock. Through the genetic similarity and evolutionary historical understanding, we clustered the IDV strains to three main groups (Japan, China and Italy), which were associated with the continental regions based on the phylogeny and PCoA. Focusing on these three main groups, we identified group-specific conserved nucleotide and amino acid variants using the genetic comparison of the complete coding sequences of 25 IDV strains. Especially in the Japan group, there were the most nucleotide and amino acid variations in the M3 and HEF proteins compared with other group of IDV strains. Out of these SNVs and TAASs, several synonymous and nonsynonymous substitutions were in the functional domains of protein-coding genes. In addition, we conducted the dN/dS analysis to identify the local-specific variations that were positively selected regarding viral survival and propagation. Only the P42 protein in the Japan group and P3 protein in the Italy group were estimated as positively selected protein. Furthermore, we conducted the B-cell epitope prediction analysis, to confirm the positively selected sites that were located in immune epitopes for escape from protective immunity in the host.

The phylogenetic analysis with IDV has been studied using the individual gene segment genome dataset, however in this study, we compared each segment trees and concatenated coding sequence dataset tree (the order of concatenated dataset: PB2, PB1, P3, HEF, NP, P42, NS1 and NS2). When constructing influenza phylogenetic tree, most of the previous studies used the most conservative gene segment called PB2 and PB1, or the glycoprotein HEF gene segment, including the biological properties. The results of the phylogenetic analysis showed a different aspect between the HEF dataset and concatenated dataset. In the previous study, IDV strains were classified as D/660, D/OK and other third lineages based on the serological and protein functional criteria. However, when using the concatenated complete coding sequences, IDV were classified according to the isolated region. Among the three main groups (Japan, China and Italy), the Italy group in Europe continent is located closer to the American strains, which are 'D/bovine/Mississippi/C00046N/2014' and 'D/swine/Oklahoma/1334/2011' than other strains, meaning that the IDV has spread from America to Europe; however, in order to accurately grasp the spreading process, subsequent epidemiological investigation is necessary.

Investigation of genetic variation provides deeper insights into the viral pathogenesis as well as protein functional changes in the virus and suggests the changes as a genetic marker. Population comparative analysis with the chi-square test between the target group and other strains using the singe nucleotide

sites of IDV coding sequence identifies target group-specific statistically meaningful as well as the most conservative mutation sites. The Japan group included the most different and mutually exclusive nucleotide and amino acid sites. In particular, the P3, HEF and NP proteins, which functionally play important roles such as replication and mRNA synthesis, contained most of the markers that are most of the conserved but mutually exclusive sites unique to the Japan group. Next, the China group also included the China group-specific markers but compared with the Japan group, the China group markers comparatively spread in most of the proteins. The Italy group included only in the P3 and HEF proteins. Interestingly, the P3 and HEF proteins were commonly contained the greatest number of markers in all of the main three groups. It can be thought of as the most variable proteins in IDV virus.

Also, to investigate which proteins were under positive selection in branch-site model within the continental group sets and functionally plays a role in IDV, we performed dN/dS and domain search analysis. As a result, only P42 protein in Japan group and P3 protein in Italy group were identified as the under positive selection. Previous studies indicated that protein P42 has a potential role in the genome packaging and uncoating process of the virus replication cycle and virion morphology. Furthermore, it functions as a protein channel during the entry of the virus into susceptible cells by allowing the acidification of the virion interior. Moreover, P3 protein in Influenza virus is RNA-dependent RNA polymerase subunit. Concerning these observed results, we

speculated that IDV evolved to local adaptation according to local specific environment and host. In addition, to comparing with other IDV strains, the evolution of the Japan IDV group was not directed to increasing the survival rate in the host, but increasing the number of particles by controlling the replication cycle. Besides, we identified the B-cell epitope candidates using the B-cell epitope prediction analysis from the positively selected and immunologically significant proteins P42, P3 and HEF. The predicted B-cell epitope candidates were listed in Table 3.3. Among the Japan group-specific markers in P42 protein, which have undergone positive selection in the branch-site model, following the amino acids at their positions were included in the epitope candidates: phenylalanine positioned at 13, asparagine positioned at 187 and isoleucine positioned at 355. Among the China group-specific markers in P3 protein which have undergone positive selection in the branch-site model as well, only the valine positioned at amino acid 194 included in the predicted epitope candidate. These markers were not only the main group specific under position selection but also included in functional domains and B-cell epitope candidates. Through these results, it would be possible to help discover and design a vaccine candidate for a group-specific vaccine and to cope with the emergence of a new type due to the evolutionary mutation of IDV.

IDV became a member of a new genus of the *Orthomkyxoviridae*, Influenza D type. It was isolated from pigs and cattle, which are the most important livestock for humans. However, there is no evidence that IDV directly affects

the livestock industry, but mutations can alter the amino acids and increase the risk of infection or infection other species. In order to prevent these threats, we conducted the genomic population comparison analysis and B-cell epitope prediction analysis between the IDV isolated from various countries. These results would be helpful to understand the evolutionary relationship between IDV strains. Also, the main three groups of Japan, China and Italy specific genetic markers might be useful to make a strategy for the arms race between mammals and IDV.

Table 3.1 Summary of IDV complete genomes

Type	Host	Isolate	Year	Strains	Accession
D	bovine	Japan (Miyazaki)	2016	D/bovine/Miyazaki/B22/2016	LC270265 - LC270271
D	bovine	Japan (Ibaraki)	2016	D/bovine/Ibaraki/7768/2016	LC128433 - LC12849
D	bovine	Japan (Yamagata)	2016	D/bovine/Yamagata/10710/2016	LC318665 - LC318671
D	swine	Italy	2016	D/swine/Italy/173287-4/2016	KX768824 - KX768830
D	swine	Italy	2015	D/swine/Italy/254578/2015	KX768831 - KX768837
D	bovine	Italy	2014	D/bovine/Italy/1/2014	KT592516 - KT592522
D	bovine	Italy	2015	D/bovine/Italy/46484/2015	KT592523 - KT592529
D	swine	Italy	2015	D/swine/Italy/199724-3/2015	KT592530 - KT592536
D	swine	Italy	2015	D/swine/Italy/354017/2015	KX768838 - KX768844
D	bovine	China (Shandong)	2014	D/bovine/Shandong/Y217/2014	KM015505 - KM015511
D	bovine	China (Shandong)	2014	D/bovine/Shandong/Y127/2014	KM015498 - KM015504
D	bovine	China (Shandong)	2014	D/bovine/Shandong/Y125/2014	KM015491 - KM015497
D	bovine	USA (Kansas)	2012	D/bovine/Kansas/14-22/2012	KM392496 - KM392502
D	bovine	USA (Kansas)	2012	D/bovine/Kansas/11-8/2012	KM392503 - KM392509
D	bovine	USA (Kansas)	2010	D/bovine/Kansas/1-35/2010	KM392496 - KM392502
D	bovine	USA (Kansas)	2012	D/bovine/Kansas/13-21/2012	KM392489 - KM392495
D	bovine	USA (Oklahoma)	2013	D/bovine/Oklahoma/660/2013	KF425659 - KF425665
D	swine	USA (Oklahoma)	2011	D/swine/Oklahoma/1334/2011	JQ922305 - JQ922311
D	bovine	USA (Texas)	2011	D/bovine/Texas/3-13/2011	KM392482 - KM392488
D	bovine	USA (Mississippi)	2014	D/bovine/Mississippi/C00046N/2014	KT581409 - KT581415
D	bovine	USA (Minnesota)	2013	D/bovine/Minnesota/628/2013	KF425652 - KF425658

D	bovine	USA (Nebraska)	2012	D/bovine/Nebraska/9-5/2012	KM392468 - KM392474
D	bovine	Mexico	2015	D/bovine/Mexico/S7/2015	KU171126, KU710418, KU710422, KU710426, KU710430, KU710434, KU710438
D	bovine	Mexico	2015	D/bovine/Mexico/S56/2015	KU1-71128, KU710420, KU710424, KU710428, KU710432, KU710436, KU710440
D	bovine	Mexico	2015	D/bovine/Mexico/S8/2015	KU171127, KU710419, KU710423, KU710427, KU710431, KU710435, KU710439

Table 3.2 Positive selection in group markers estimated by using branch-site model

Foreground branch group	Protein	H0 lnL	H1 lnL	LRT	*p* value	Proportion	W_2	Protein Position	Posterior probability
China	P42	-2311.01	-2310.76	0.502832	0.478258	0.00465	1.0	-	-
	PB1	-4699.35	-4699.35	8E-06	0.997743	0.00041	1.0	-	-
Italy	P3	-4565.03	-4565.00	0.056358	0.812348	0.00751	4.16884	193	0.788
								594	0.577
Japan	P42	-2310.49	-2308.48	4.024526	0.044843	0.00254	23.10130	13	0.607
								184	0.572
								288	0.601
								332	0.577
								359	0.576
								361	0.898
								368	0.587
	PB2	-4817.92	-4817.92	4E-05	0.994954	0.0011	1.0	-	-

Table 3.3 B-cell specific predicted Epitopes including genetic markers

		Japan		China		Italy	
Protein	**Markers**	**Start-End**	**Peptide**	**Start-End**	**Peptide**	**Start-End**	**Peptide**
HEF	212,215	212-218	**S**PQ**L**CGT	212-218	KPQVCGT	212-218	KPQVCGT
	249	232-249	IYKCNKHVVQLCYFVYE**S**	232-249	IYKCNKHVVQLCYFVY	232-249	IYKCNKHVVQLCYFVY
	288, 290, 301	288-303	**S**A**E**VKIECPSKIL**S**PG	292-299	KIECPSKI	288-299	NVGVKIECPSKI
	409, 413	408-415	VKDYL**I**PP	408-415	VKDYLSPP	408-415	V**R**DYLSPP
	473	459-475	IDDLIFGLLFIGFVAGG	459-475	IDDLIFGLLFVGFV**T**GG	459-475	IDDLIFGLLFVGFVAGG
P42	13	4-17	EQLLAELEG**F**LRGV	4-14	EQLLAELEGYL	4-17	EQLLAELEGYLRGV
	187	185-192	AN**V**VPMK	185-192	MASVVPMK	186-192	MASVVPMK
	244, 261	238-263	NLALKRSVLTLLMLVICGIPT CVNAE	238-263	NLALKR**L**VLTLLMLVICGIPTC V**D**AE	238-263	NLALKRSVLTLLMLVICGIPTC VNAE
	304,314	294-320	MTLAALILGCFSMLYILIKAI LMLLLTI	294-320	TLAALILGCF**G**MLYILIKAI**M**M LLLTI	293-320	TLAALILGCFSMLYILIKAILM LLLTI
	335	332-341	LKH**I**VKCFSE	332-341	LKHVVKCFSE	332-341	LKHVVKCFSE
P3	194	190-195	SKLFIA	190-195	SKLFIA	190-195	SKLF**V**A
	250	249-254	R**S**LKIP	248-254	LRPLEIP	-	-
	307	-	-	304-310	PKPIFGK	304-310	PKPIFGK
	345	340-348	DFLCG**V**GRA	340-346	DFLCGIG	340-346	DFLCGIG
	528	517-524	KYTVFEAG	517-535	KYTVFEAGTVPVEAVVLTP	517-535	KYTVFEAGTVPVEAVVLTP
		528-535	**M**EAVVLTP				
	541	541-551	IKE**K**KLFLYCR	538-551	ERVLKEKKLFLYCR	538-551	ERVLKEKKLFLYCR

Supplementary Table S3.1 List of predicted functional protein domains

Protein	PSSM-ID	From (AA)	To (AA)	E-Value	Bitscore	Accession	Short name	Function & Characteristic
PB2	278999	5	770	0	679.325	cl20020	Flu_PB2 superfamily	**Influenza RNA-dependent RNA polymerase subunit PB2** PB2 can bind 5' end cap structure of RNA.
PB1	278997	1	736	0	982.054	cl15495	Flu_PB1 superfamily	**Influenza RNA-dependent RNA polymerase subunit PB1** Two GTP binding sites exist in this protein.
P3	278998	23	702	0	617.341	cl02905	Flu_PA superfamily	**Influenza RNA-dependent RNA polymerase subunit PA**
HEF	332837	52	422	2.52E-115	350.492	cl28016	Hema_esterase superfamily	**Hemagglutinin esterase**
	72144	456	628	7.03E-54	182.156	cl07368	Hema_stalk superfamily	**Influenza C hemagglutinin stalk** This domain corresponds to the stalk segment of hemagglutinin in influenza C virus. It forms a coiled coil structure.
NP	278907	3	501	3.08E-25	108.986	cl27387	Flu_NP superfamily	**Influenza virus nucleoprotein**
P42	281077	1	233	1.62E-59	192.669	cl03846	CM1 superfamily	**Influenza C virus M1 protein** This family represents the matrix 1 protein of influenza C virus. The protein is the product of a spliced mRNA. Small quantities of the unspliced mRNA are found in the cell additionally encoding the M2 protein (see pfam03021).
	281073	236	383	7.28E-10	56.5925	cl03843	CM2 superfamily	**Influenza C virus M2 protein** Influenza C virus M1 protein is encoded by a spliced mRNA. The unspliced mRNA is also found in small quantities and can encode the protein represented by this family.

NS1	281544	7	63	1.09E-06	44.4998	cl04153	Flu_C_NS2 superfamily	**Influenza C non-structural protein (NS2)** The influenza C virus genome consists of seven single-stranded RNA segments. The shortest RNA segment encodes a 286 amino acid non-structural protein NS1 pfam03506 as well as the NS2 protein. The NS2 protein is only about 60 amino acids in length and of unknown function.
	281501	72	210	2.05E-13	65.8931	cl04118	Flu_C_NS1 superfamily	**Influenza C non-structural protein (NS1)** The influenza C virus genome consists of seven single-stranded RNA segments. The shortest RNA segment encodes a 286 amino acid non-structural protein NS1. This protein contains 6 conserved cysteines that may be functionally important, perhaps binding to a metal ion.
NS2	281544	7	63	9.42E-05	38.7218	cl04153	Flu_C_NS2 superfamily	**Influenza C non-structural protein (NS2)** The influenza C virus genome consists of seven single-stranded RNA segments. The shortest RNA segment encodes a 286 amino acid non-structural protein NS1 pfam03506 as well as the NS2 protein. The NS2 protein is only about 60 amino acids in length and of unknown function.

Supplementary Table S3.2 Japan group specific TAAS analysis result

Gene	Position	TargetAA	OthersAA
PB2	139	S	P
PB2	204	T	A
PB2	439	R	K
PB2	494	I	V
PB2	518	M	I
PB2	556	Q	K
PB2	580	I	L
PB2	722	V	I
PB2	730	V	I
PB1	835	M	I
PB1	965	Y	H
PB1	995	M	I
PB1	1068	R	Q
P3	1670	S	N
P3	1706	R	G
P3	1775	S	P
P3	1777	K	E
P3	1780	E	K
P3	1832	T	I
P3	1870	V	I
P3	1920	E	D
P3	1927	N	D
P3	2001	P	S
P3	2010	V	I
P3	2053	M	V
P3	2066	I	L
P3	2121	I	MV
HEF	2246	M	I

HEF	2253	K	R
HEF	2447	S	KR
HEF	2450	L	AV
HEF	2484	S	N
HEF	2490	A	T
HEF	2523	S	GN
HEF	2525	E	GX
HEF	2536	S	N
HEF	2648	I	S
HEF	2722	E	G
HEF	2770	S	N
HEF	2839	G	S
HEF	2876	T	IM
NP	2905	V	A
NP	2979	T	A
NP	3017	D	E
NP	3101	K	R
NP	3111	D	N
NP	3116	T	S
NP	3192	S	P
NP	3306	S	F
NP	3396	N	D
NP	3418	E	G
NP	3422	A	V
NP	3449	D	G
P42	3464	F	Y
P42	3638	N	S
P42	3742	Q	R
P42	3786	I	V
P42	3813	G	E
P42	3815	D	S

P42	3822	S	G
NS1	3881	K	E
NS1	3911	V	I
NS1	3946	E	D
NS1	3978	L	V
NS2	4124	K	E
NS2	4194	S	N

Supplementary Table S3.3 China group specific TAAS analysis result

Gene	Position	TargetAA	OthersAA
PB2	333	R	K
PB2	400	A	S
PB2	521	T	IMV
PB1	937	L	M
PB1	1140	V	I
PB1	1142	V	I
PB1	1354	D	E
P3	1771	S	N
P3	1894	D	E
P3	2136	T	N
HEF	2392	M	KR
HEF	2398	IK	R
HEF	2662	R	EG
HEF	2708	T	A
NP	2995	N	S
NP	3020	K	R
NP	3187	I	M
NP	3288	T	A
NP	3390	V	I
P42	3671	R	G
P42	3695	L	S
P42	3712	D	N
P42	3755	G	S
P42	3765	M	L
NS1	3912	T	A
NS1	3921	L	HY
NS1	4068	I	V

Supplementary Table S3.4 Italy group specific TAAS analysis result

Gene	Position	TargetAA	OthersAA
P3	1719	V	I
P3	2121	V	IM
HE	2644	R	K
HE	2887	V	A

Figure 3.1 Phylogenetic analysis showing evolutionary history and PCoA of IDV

(A) and (C) are the phylogenetic trees of 25 IDV strains using the maximum likelihood analysis in the MEGA7 showing the evolutionary history and distance. (B) and (D) are the PCoA plots showing the distribution between the 25 strains of IDV based on the identical matrix. Figure (A) and (B) used the HEF protein coding sequence, whereas, figure (C) and (D) used concatenated all of the protein-coding sequence of IDV (PB2, PB1, P3, HEF, NP, P42, NS1 and NS2). The bar in figure (A) indicates the representative IDV lineages D/OK, D/660 and other third lineages consisted of Japan strains. Also, clusters in figure (B) indicates the three lineages.

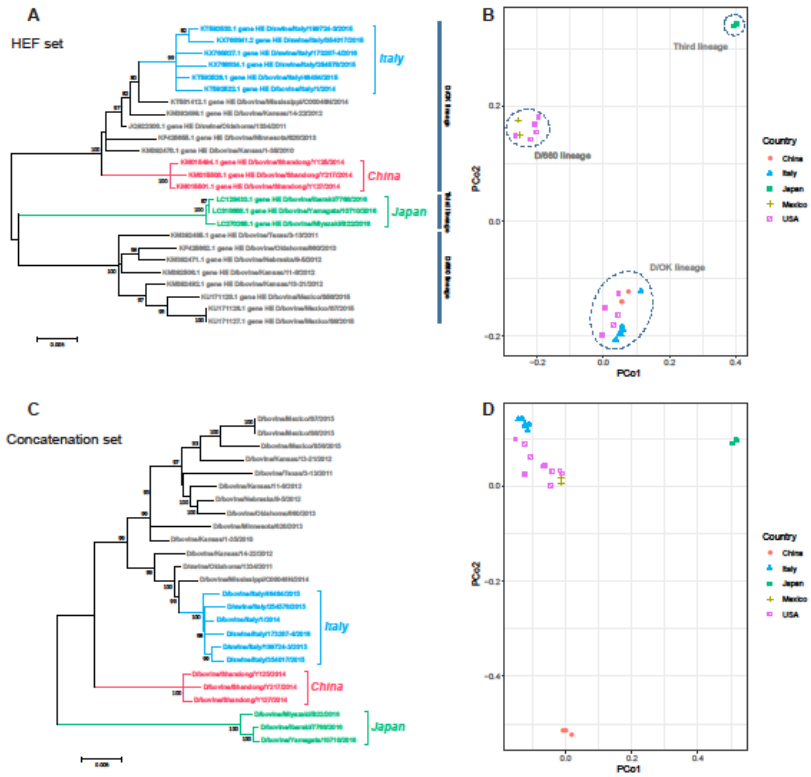


Figure 3.2 Local group-specific significant Nucleotide positions on IDV

Bar plots are showing the represent statistically different nucleotide positions between the manually divided main group and others. The colored bars indicate the group-specific mutually exclusive nucleotide positions in each group.

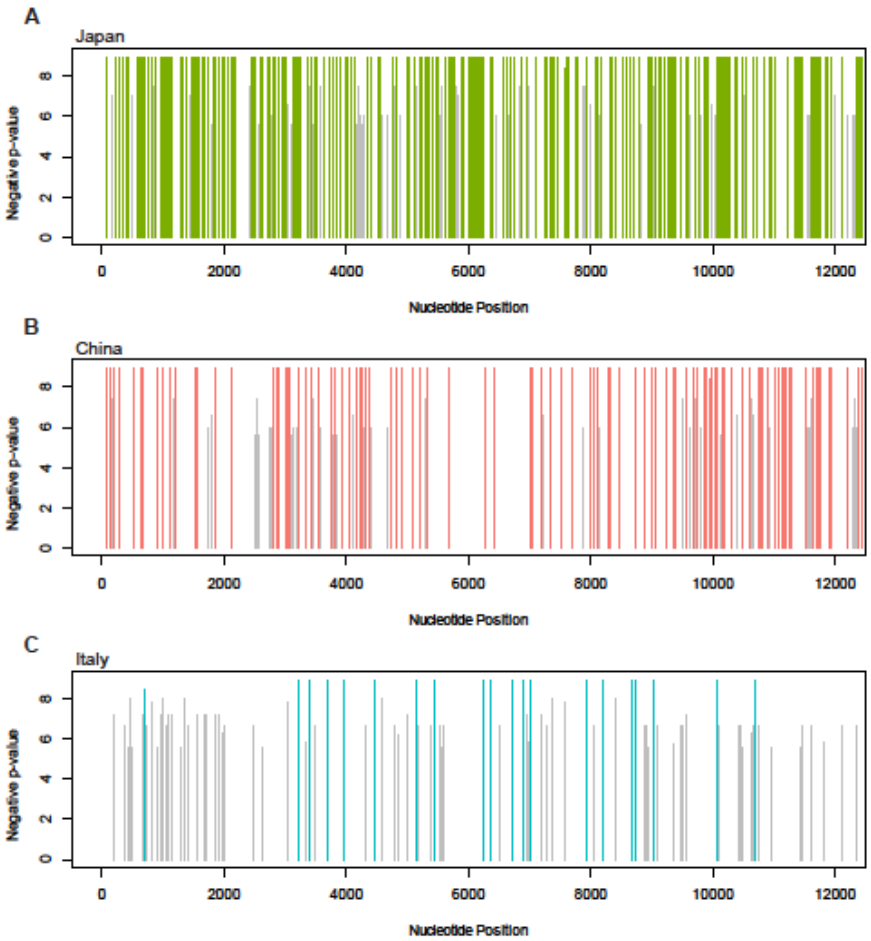


Figure 3.2 Conserved protein domains with Group-specific nonsynonymous substitution sites.

The Indicates domain structure and architecture of each IDV protein with the conserved domains, which were identified using the NCBI Conserved Domains Search software. The amino acid sequences of each segment were used as query. Additionally, the main three group specific nonsynonymous substitution positions were marked as pins and were colored by the group.

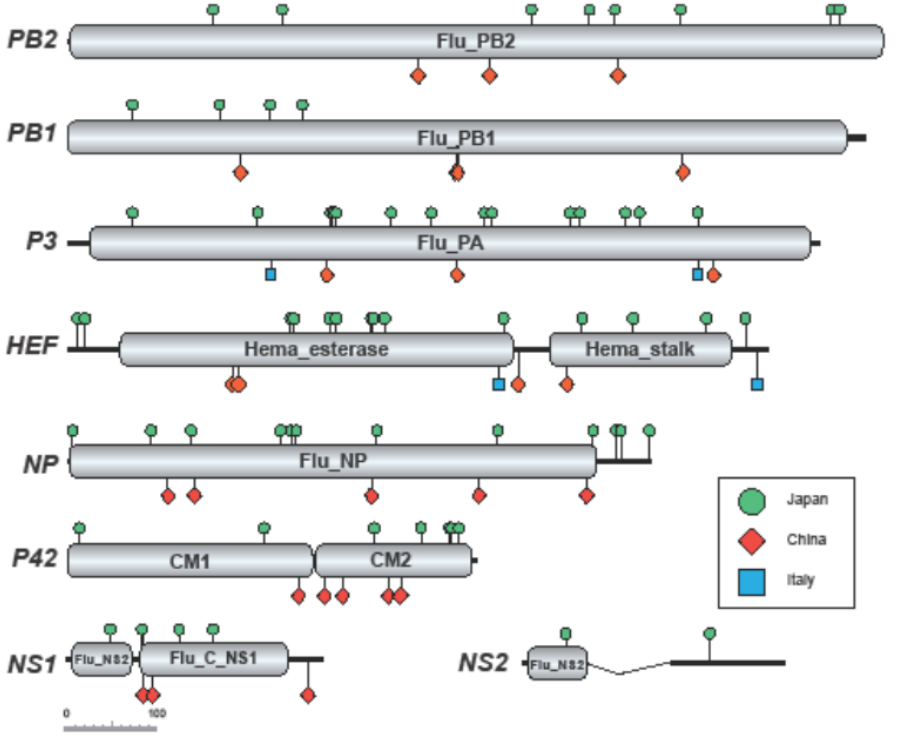
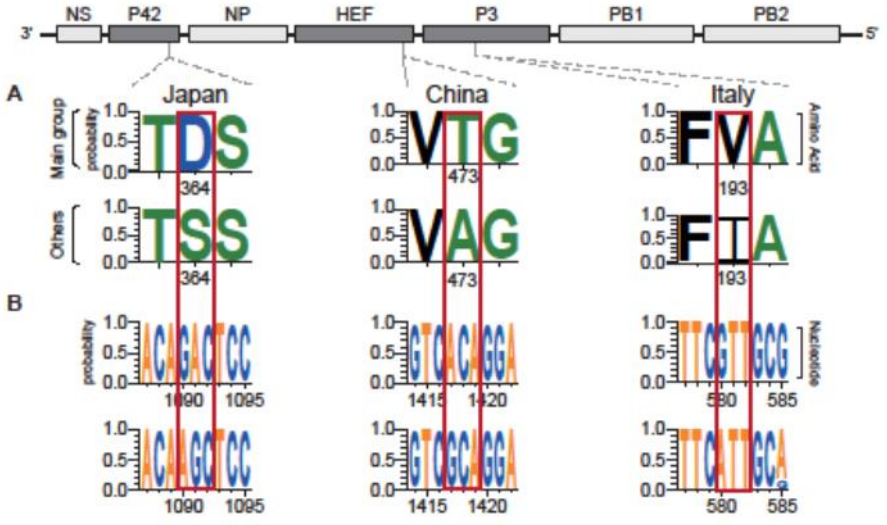


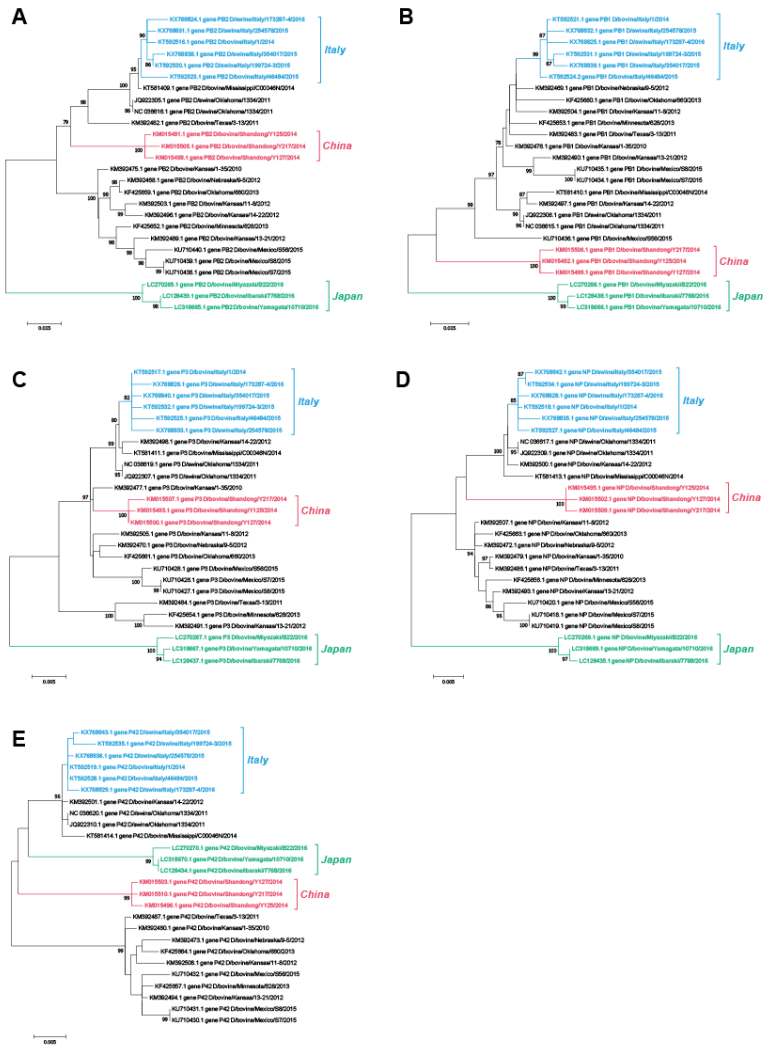
Figure 3.3 Representative Nonsynonymous substitution sites Logo of each main three group.

Seven single-stranded negative-sense, viral RNA segments were illustrated. The main three group-specific representative nonsynonymous substitution sites were also indicated as amino acid and nucleotide sequence logo.



Supplementary Figure S3.1

Phylogenetic trees of the five segments of Influenza D virus. Using maximum likelihood analysis in MEGA7 with 1000 bootstrap replicates the five phylogenetic trees were constructed. A) PB2, B) PB1, C) P3, D) NP, E) P42.



References

Alera, M. T., et al. (2015). "Zika virus infection, Philippines, 2012." Emerging infectious diseases **21**(4): 722.

Bahir, I., et al. (2009). "Viral adaptation to host: a proteome-based analysis of codon usage and amino acid preferences." Molecular systems biology **5**(1): 311.

Buathong, R., et al. (2015). "Detection of Zika virus infection in Thailand, 2012–2014." The American journal of tropical medicine and hygiene **93**(2): 380-383.

Chapman, B. and J. Chang (2000). "Biopython: Python tools for computational biology." ACM Sigbio Newsletter **20**(2): 15-19.

Charrad, M., et al. (2012). "NbClust Package: finding the relevant number of clusters in a dataset." UseR! 2012.

Chiapponi, C., et al. (2016). "Detection of influenza D virus among swine and cattle, Italy." Emerging infectious diseases **22**(2): 352.

Clamp, M., et al. (2004). "The jalview java alignment editor." Bioinformatics **20**(3): 426-427.

Collin, E. A., et al. (2015). "Cocirculation of two distinct genetic and antigenic lineages of proposed influenza D virus in cattle." Journal of virology **89**(2): 1036-1042.

Crooks, G. E., et al. (2004). "WebLogo: a sequence logo generator." Genome research **14**(6): 1188-1190.

Deng, Y.-Q., et al. (2016). "Isolation, identification and genomic characterization of the Asian lineage Zika virus imported to China." Science China Life Sciences **59**(4): 428-430.

Dick, G., et al. (1952). "Zika virus (I). Isolations and serological specificity." Transactions of the Royal Society of Tropical Medicine and Hygiene **46**(5): 509-520.

Dohnal, V., et al. (2003). "D-index: Distance searching index for metric data sets." Multimedia Tools and Applications **21**(1): 9-33.

Enfissi, A., et al. (2016). "Zika virus genome from the Americas." The Lancet **387**(10015): 227-228.

Ferguson, L., et al. (2015). "Influenza D virus infection in Mississippi beef cattle." Virology **486**: 28-34.

Fieser, T. M., et al. (1987). "Influence of protein flexibility and peptide conformation on reactivity of monoclonal anti-peptide antibodies with a protein alpha-helix." Proceedings of the National Academy of Sciences **84**(23): 8568-8572.

Foy, B. D., et al. (2011). "Probable non–vector-borne transmission of Zika virus, Colorado, USA." Emerging infectious diseases **17**(5): 880.

Fray, M., et al. (2000). "The effects of bovine viral diarrhoea virus on cattle reproduction in relation to disease control." Animal Reproduction Science **60**: 615-627.

Grubaugh, N. D., et al. (2018). "Genomic Insights into Zika Virus Emergence and Spread." Cell **172**(6): 1160-1162.

Haddow, A. D., et al. (2012). "Genetic characterization of Zika virus strains: geographic expansion of the Asian lineage." PLoS neglected tropical diseases **6**(2): e1477.

Hall, T. (2005). "Bioedit v 7.0. 5." Ibis Therapeutics, a division of Isis Pharmaceuticals, Carlsbad.

Hammer, Ø., et al. (2001). "PAST: paleontological statistics software package for education and data analysis." Palaeontologia electronica **4**(1): 9.

Hartigan, J. A. and M. A. Wong (1979). "Algorithm AS 136: A k-means clustering algorithm." Journal of the Royal Statistical Society. Series C (Applied Statistics) **28**(1): 100-108.

Hause, B. M., et al. (2017). "An inactivated influenza D virus vaccine partially protects cattle from respiratory disease caused by homologous challenge." Veterinary microbiology **199**: 47-53.

Hayes, E. B. (2009). "Zika virus outside Africa." Emerging infectious diseases **15**(9): 1347.

Ho, Z. J. M., et al. (2017). "Outbreak of Zika virus infection in Singapore: an epidemiological, entomological, virological, and clinical analysis." The Lancet Infectious Diseases **17**(8): 813-821.

Huang, J. and W. J. B. i. Honda (2006). "CED: a conformational epitope database." **7**(1): 7.

Israr-ul, H. A., et al. (2013). "Phenotypic analysis of NS5A variant from liver transplant patient with increased cyclosporine susceptibility." Virology **436**(2): 268-273.

Jiang, W.-M., et al. (2014). "Identification of a potential novel type of influenza virus in Bovine in China." Virus genes **49**(3): 493-496.

Jun, S.-R., et al. (2017). "Suggested mechanisms for Zika virus causing microcephaly: what do the genomes tell us?" BMC bioinformatics **18**(14): 471.

Kaiser, C. A., et al. (2007). Molecular cell biology, WH Freeman.

Kaminski, M. M., et al. (2013). "Pandemic 2009 H1N1 influenza A virus carrying a Q136K mutation in the neuraminidase gene is resistant to zanamivir but exhibits reduced fitness in the guinea pig transmission model." Journal of virology **87**(3): 1912-1915.

Kiepiela, P., et al. (2007). "CD8+ T-cell responses to different HIV proteins have discordant associations with viral load." Nature medicine **13**(1): 46.

Kindhauser, M. K., et al. (2016). "Zika: the origin and spread of a mosquito-borne virus." Bulletin of the World Health Organization **94**(9): 675.

Kolaskar, A. and P. C. Tongaonkar (1990). "A semi-empirical method for prediction of antigenic determinants on protein antigens." FEBS letters **276**(1-2): 172-174.

Kolaskar, A. and P. C. Tongaonkar (1990). "A semi-empirical method for prediction of antigenic determinants on protein antigens." FEBS letters **276**(1-2): 172-174.

Kumar, S., et al. (2016). "MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets." Molecular biology and evolution **33**(7): 1870-1874.

Larsen, M. V., et al. (2005). "An integrative approach to CTL epitope prediction: a combined algorithm integrating MHC class I binding, TAP transport efficiency, and proteasomal cleavage predictions." European journal of immunology **35**(8): 2295-2303.

Longdon, B., et al. (2014). "The evolution and genetics of virus host shifts." PLoS pathogens **10**(11): e1004395.

Löytynoja, A. (2014). Phylogeny-aware alignment with PRANK. Multiple sequence alignment methods, Springer**:** 155-170.

Löytynoja, A. and N. Goldman (2008). "Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis." Science **320**(5883): 1632-1635.

Malet, H., et al. (2008). "The flavivirus polymerase as a target for drug discovery." Antiviral research **80**(1): 23-35.

Malone, R. W., et al. (2016). "Zika virus: medical countermeasure development challenges." PLoS neglected tropical diseases **10**(3): e0004530.

Marchler-Bauer, A., et al. (2010). "CDD: a Conserved Domain Database for the functional annotation of proteins." Nucleic acids research **39**(suppl_1): D225-D229.

Mehrjardi, M. Z. (2017). "Is Zika virus an emerging TORCH Agent? An invited commentary." Virology: research and treatment **8**: 1178122X17708993.

Murakami, S., et al. (2016). "Influenza D virus infection in herd of cattle, Japan." Emerging infectious diseases **22**(8): 1517.

Musso, D. and D. J. Gubler (2016). "Zika virus." Clinical microbiology reviews **29**(3): 487-524.

Muylkens, B., et al. (2007). "Bovine herpesvirus 1 infection and infectious bovine rhinotracheitis." Veterinary research **38**(2): 181-209.

Neumann, G., et al. (2009). "Emergence and pandemic potential of swine-origin H1N1 influenza virus." Nature **459**(7249): 931.

Organization, W. H. (2016). "Zika virus, microcephaly and Guillain-Barré syndrome situation report."

Pepin, K. M., et al. (2010). "Identifying genetic markers of adaptation for surveillance of viral host jumps." Nature Reviews Microbiology **8**(11): 802.

Perera, R., et al. (2008). "Closing the door on flaviviruses: entry as a target for antiviral drug design." Antiviral research **80**(1): 11-22.

Pickett, B., et al. (2013). "Metadata-driven comparative analysis tool for sequences (meta-CATS): an automated process for identifying significant sequence variations that correlate with virus attributes." Virology **447**(1-2): 45-51.

Pickett, B. E., et al. (2011). "ViPR: an open bioinformatics database and analysis resource for virology research." Nucleic acids research **40**(D1): D593-D598.

Pileri, E. and E. Mateu (2016). "Review on the transmission porcine reproductive and respiratory syndrome virus between pigs and farms and impact on vaccination." Veterinary research **47**(1): 108.

Plotkin, J. B. and G. Kudla (2011). "Synonymous but not the same: the causes and consequences of codon bias." Nature Reviews Genetics **12**(1): 32.

Ragoza, M., et al. (2017). "Protein–Ligand scoring with Convolutional neural networks." Journal of chemical information and modeling **57**(4): 942-957.

Rasmussen, S. A., et al. (2016). "Zika virus and birth defects—reviewing the evidence for causality." New England Journal of Medicine **374**(20): 1981-1987.

Schvoerer, E., et al. (2013). "Hepatitis C virus envelope glycoprotein signatures are associated with treatment failure and modulation of viral entry and neutralization." The Journal of infectious diseases **207**(8): 1306-1315.

Suyama, M., et al. (2006). "PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments." Nucleic acids research **34**(suppl_2): W609-W612.

Talavera, G. and J. Castresana (2007). "Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments." Systematic biology **56**(4): 564-577.

Tappe, D., et al. (2014). "First case of laboratory-confirmed Zika virus infection imported into Europe, November 2013." Eurosurveillance **19**(4): 20685.

Valarcher, J.-F. and G. Taylor (2007). "Bovine respiratory syncytial virus infection." Veterinary research **38**(2): 153-180.

Wang, L., et al. (2016). "From mosquitos to humans: genetic evolution of Zika virus." Cell host & microbe **19**(5): 561-565.

Weaver, S. C. (2017). "Emergence of epidemic Zika virus transmission and congenital Zika syndrome: are recently evolved traits to blame?" MBio **8**(1): e02063-02016.

White, M. K., et al. (2016). "Zika virus: an emergent neuropathological agent." Annals of neurology **80**(4): 479-489.

Wichman, H., et al. (1999). "Different trajectories of parallel evolution during viral adaptation." Science **285**(5426): 422-424.

Wickham, H. (2016). ggplot2: elegant graphics for data analysis, Springer.

Yan, Z., et al. (2018). "Evolutionary changes of the novel Influenza D virus hemagglutinin-esterase fusion gene revealed by the codon usage pattern." Virulence(just-accepted).

Yang, Z. (2007). "PAML 4: phylogenetic analysis by maximum likelihood." Molecular biology and evolution **24**(8): 1586-1591.

Ye, Q., et al. (2016). "Genomic characterization and phylogenetic analysis of Zika virus circulating in the Americas." Infection, Genetics and Evolution **43**: 43-49.

Yun, S.-I., et al. (2016). "Complete genome sequences of three historically important, spatiotemporally distinct, and genetically divergent strains of Zika virus: MR-766, P6-740, and PRVABC-59." Genome announcements **4**(4): e00800-00816.

Zanluca, C., et al. (2015). "First report of autochthonous transmission of Zika virus in Brazil." Memórias do Instituto Oswaldo Cruz **110**(4): 569-572.

Zhai, S.-L., et al. (2017). "Influenza D virus in animal species in Guangdong Province, Southern China." Emerging infectious diseases **23**(8): 1392.

Zhang, G., et al. (2014). "Comparative genomics reveals insights into avian genome evolution and adaptation." Science **346**(6215): 1311-1320.

요약(국문초록)

생물정보학적 접근법을 통한 포유류 감염성 바이러스 유전자 마커의 동정

김정웅

농생명공학부 농생명공학전공

서울대학교 대학원 농업생명과학대학

감염성 바이러스는 인간을 비롯한 많은 종의 동물을 감염시켜 돌이킬 수 없는 결과를 초래하기도 합니다. 수많은 사람을 죽음에 이르게 하는 것은 물론, 매 해마다 대규모 가축 감염사례로 인하여 축산업에 커다란 경제적 피해를 끼치고 있습니다. 그렇기 때문에 감염성 바이러스에 대한 충분한 연구가 필요합니다. 바이러스는 다른 미생물이나 생명체에 비하여 유전자 변형이 보다 빠르고 무작위로 이루어지는 특징이 있습니다. 대부분의 바이러스는 숙주의 종에 따라 감염 여부가 달라지지만, 뉴클레오타이드와 아미노산 서열 하나의 변형으로도 새로운 종의 숙주를 감염시키거나 그 독성이 달라지기도 하기 때문에 그들의 유전체 차원에서의 특징을 발견하고 분석하는 것은 상업적 및 과학적 주요한 가치를 제공합니다. 이러한 유전체 특징 중에서 단일 유전자 변이체(Single

Nucleotide and Amino acid variant)는 많은 연구에서 연구 대상으로 사용되고 있습니다. 실제적으로 바이러스 연구에서 바이러스의 종을 동정하거나 백신 개발 등 다양한 분야에 사용되고 있습니다.

챕터 2 | 지카바이러스는 일반적인 성인이 감염되었을 시에는 지카열, 두통 및 관절통 등의 증상을 유발하지만 임산부가 감염되었을 시에는 태아의 소두증을 일으키는 것과 연관이 있다고 알려져 있습니다. 지난 10 년간 전 세계에 폭발적으로 퍼져 나갔으며 많은 학자들이 지카바이러스의 분자 메커니즘에 대한 연구를 수행했습니다. 그러나 치료와 예방을 위한 의약품 및 백신 개발은 아직까지 진행 중이며 보다 많은 유전체 수준에서의 연구가 필요합니다.

이 연구에서 공개데이터베이스로부터 이용 가능한 지카바이러스의 NGS 유전체 데이터를 수집하고 분석을 통하여 지리적, 시기적 관점을 고려한 지역 특이적 유전체 변이(Single Nucleotide and Amino Acid variants)를 유전자 마커로써 제시하였습니다. 진화적 연관분석과 자율학습 k-means 클러스터링 알고리즘을 이용하여 4 개의 대표그룹을 선정하였습니다. 대표 4 그룹에 초점을 맞추어 통계적으로 유의미한 유전체 변이들을 찾아내고 dN/dS 진화 분석으로 진화적으로 가속화된 단백질 암호화 영역을 확인했습니다. 이후 그룹 기능성 단백질 영역과 B-cell, T-cell 특이적 항원결정기 후보를 예측하여 찾아낸 유전체

변이들이 단백질 및 항원결정기 형성의 결정적인 역할을 확인하여 그룹별 주요 유전자 마커로써 제안하였습니다.

챕터 3 | 인플루엔자의 새로운 타입으로 분류된 인플루엔자 D 바이러스는 소를 비롯한 반추동물을 감염시키는 호흡기성 바이러스입니다. 감염 증상은 경미하지만 다른 치명적인 호흡기성 바이러스 감염을 유발하고 인간에게도 감염될 수 있는 잠재성이 있기 때문에 유전체 차원에서의 연구를 수행하였습니다. 인플루엔자 D 바이러스의 모든 유전자 단편 NGS 데이터를 이용한 유전체 특성 및 진화적 상관관계 분석으로 하나의 유전자 단편을 통한 분석의 결과와의 차이점을 밝혀냈습니다. 그 결과를 토대로 선정한 대표 그룹을 초점으로, 통계적으로 유의미한 특이적 유전체 변이를 찾아냈습니다. 이후 dN/dS 진화 분석과 단백질 코딩영역, B-cell 특이적 항원결정기 예측 분석 결과와 비교하여 그룹 특이적 유전자 마커로써 제안하였습니다.

이 연구를 통하여 감염성 바이러스의 그룹별 특이적 유전자 마커를 제시하고 이 마커가 새로운 바이러스 종의 동정과 병독성 진화에 대한 통찰, 그리고 백신 개발에 도움을 줄 수 있을 것입니다.

주요어: 감염성 바이러스, 유전자 마커, 유전체 변이, 진화 분석, 기능성 단백질 영역, 항원결정기

학번: 2017-29310