



Chikungunya virus: genomic microevolution in Eastern India and its in-silico epitope prediction

Sudip Kumar Dutta¹ · Tamanash Bhattacharya^{1,2} · Anusri Tripathi¹

Received: 31 October 2017 / Accepted: 6 July 2018 / Published online: 14 July 2018
© Springer-Verlag GmbH Germany, part of Springer Nature 2018

Abstract

This is the first study reporting whole genome sequences of two CHIKV strains (KJ679577 and KJ679578) isolated from Eastern Indian patients sera during 2010–2011 outbreak, both of which were of ECSA genotype, but from different subgroups: Indian Ocean outbreak and ECSA subtypes. Furthermore, viral sequences were analyzed using different in-silico approaches to identify potential genetic variations that might have functional implications on various aspects of virus replication, viral protein functionality, immunogenicity and transmission. Epitope prediction analysis revealed 70.9% increase in number of MHC Class-II interacting epitopes of KJ679578 and 25–28% increase in Class-I interacting epitopes of KJ679577 and KJ679578 compared to that of EF027141 (CHIKV of Asian genotype circulating in India during 1973, after which CHIKV infection disappeared from India for three decades). CHIKV peptides DLAKLAFKRSSKYDLECAQIPVHMKSDA and KVVLCGDPKQCGFFNMMQMKYNYNHNI were predicted to interact with maximum number of HLA Class-I (68 and 76.5%, respectively) and Class-II (47 and 100%, respectively) alleles present within Indian population with allele frequency of >0.1 and were also recognized as predicted B-cell epitopes with BCPred score between 0.766 and 0.961 and with antigenicity ranging from 0.52 to 1.69; thus these peptides might be used to induce T- and B-cell-mediated immunity against CHIKV. Thus, the present study might help to bridge the gap between virus microevolution and its implication in host immunity by taking into account viral genetic and conformational changes. Predicted epitopes might be used as promising targets for peptide-based vaccine development and rapid diagnostics against CHIKV infection.

Keywords Chikungunya · Genome sequencing · Substitution rate · Epitope · Disease severity

Introduction

Virus dissemination is on the rise as a consequence of increased global connectivity with respect to trade and travel (Pybus and Rambaut 2009). Invariably, this exerts an effect on the genetic diversity of viruses, notably RNA viruses, owing to their elevated mutational and evolutionary rates. Given their spatial distribution, a viral

population is capable of accumulating many genetic differences over the course of an outbreak. As different viral populations adapt to distinct, sometimes novel spatial environments, they gain adaptive changes that provide advantages that ultimately result in the virus potentially adapting to new vectors or cause antigenic drifts that allows evasion of host immune system. Examples of such adaptive changes can be found in arboviruses, for example, chikungunya virus (CHIKV), an arthropod-borne positive-strand alphavirus of the Togaviridae family. Although rarely fatal, CHIKV infection leads to debilitating arthralgia that can persist for decades. After a thirty-two-year period of quiescence, massive chikungunya (CHIK) outbreaks were reported from 25 Indian states and the Indian Ocean islands, affecting approximately 14.2 million people between 2005 and 2014, thus proving to be a major socio-economic burden in a developing nation (Cecilia 2014). Previously gene genealogy-based evolutionary studies using MEGA and MrBayes software grouped

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s13205-018-1339-3>) contains supplementary material, which is available to authorized users.

✉ Anusri Tripathi
anusri.stm@gmail.com

¹ Department of Biochemistry and Medical Biotechnology, Calcutta School of Tropical Medicine, 108, C.R. Avenue, Kolkata, West Bengal 700073, India

² Present Address: Department of Biology, Indiana University Bloomington, Bloomington, IN 474057000, USA

CHIKV into three geographically associated lineages based upon their origin: Asian, East Central South African (ECSA) and West African (Waf) (Volk et al. 2010; Casal et al. 2015). ECSA clade can be further subdivided into: ECSA-I, ECSA-II and ECSA-III, all of which have been reported to share a common ancestor. Notably, majority of the CHIKV epidemic strains have been included within ECSA-III subgroup. The CHIKV epidemic strains, infecting populations of Indian Ocean islands, Indian subcontinent and South East Asian countries, originated from African enzootic strains with differential patterns and rate of evolution (Casal et al. 2015). Previous studies have indicated simultaneous circulation of two distinct lineages of CHIKV ECSA genotype among Indian patients (Dutta et al. 2014), both of which were different from CHIKV circulating within India during 1963–1973, which belonged to Asian genotype (Cecilia 2014).

CHIKV is an enveloped virus with a genome size of 11.7 kb, two-thirds of which encodes non-structural proteins (nsP1, nsP2, nsP3, nsP4) possessing protease, helicase, methyltransferase, and RNA-dependent RNA polymerase activities and the remaining one-third genome codes for structural proteins (C, E3, E2, 6K, E1) are involved in viral particle assembly. Alpha virus specific four conserved sequence elements (CSEs) present within CHIKV genome, viz., CSE 1 (found at 5' UTR of genomic RNA containing 44-nt), CSE 2 (located within the 5' end of nsP1 containing 51-nt), CSE 3 (located in between the junction region of two ORFs containing 24-nt) and CSE 4 (contains 19-nt appears just upstream of 3' UTR), are known to play important roles in viral replication (Kam et al. 2012). Alterations in specific functional regions of CHIKV, such as, the “N-wings” of E2 domain A, “acid sensitive region” and β -ribbon “connector” located within E2 domain B, fusion and ij-loop in E1 domain II could potentially affect viral lifecycle within its human host (Scholte et al. 2013). One unique mutation (CGA–UGA) was reported between nsP3 and nsP4 region resulting in generation of opal (termination) codon at the end of nsP3, which further enhances nsP34 cleavage, leading to increased nsP3 protein stability and virus infectivity (Chen et al. 2013). Aside from viral proteins, host genetic factors, such as the human leukocyte antigen (HLA) loci influence viral immunogenicity, thereby affecting disease susceptibility. Currently, no FDA approved drug or license vaccine is available against CHIKV infection, thus the antigenic aspects of viral proteins needs to be further explored for development of any specific drug. Though vaccines are considered to be a compelling method to control vector-borne infections, none have yet advanced to clinical development in case of CHIKV. Epitope-based vaccines are considered to be safer. Many, immunoinformatics tools, viz., ProPred, ProPred-1, BCPred, and NetCTLpan have

been utilized for development of stable and multi-epitope vaccines against various viruses (Bourdette et al. 2005; Knutson et al. 2001).

In the present study, microevolution and immunogenicity of recent CHIKV outbreak strains from Eastern India were compared with previously circulating strain from same region decades apart. Furthermore, in-silico approach was used to predict evolutionary conserved and highly immunogenic epitopes present within CHIKV. These epitopes might be engineered into promising targets for peptide-based vaccine development and rapid diagnostics against CHIKV infection.

Materials and methods

Patient sample collection

As part of a previous study, approximately 3–5 ml blood was collected from each of 199 symptomatic Eastern Indian patients visiting Calcutta School of Tropical Medicine, India during October 2010 and December 2011 CHIKV epidemic and having any three of the WHO-defined symptoms of chikungunya infection, viz., fever (39–40 °C), myalgia (at forearms, arms, thighs and calves), joint pain (at large joints such as shoulders, elbows, wrists, knees, ankles), rash during blood collection, after obtaining approval from the clinical research ethical board of the institute (Dutta et al. 2014). Among them, 130 were CHIKV RT-PCR positive, out of which 15 patients infected with high copies of CHIKV (confirmed using qRT-PCR, none of them showed PCR/IgM positivity for dengue) were previously sequenced partially (571 bp of E1 region), which occupied two distinct cladistic positions in phylogenetic tree-construct. The ECSA clade consisted of 13 identical CHIKV sequences and the Indian Ocean clade consisted of remaining two identical sequences. One representative serum from each clade was randomly selected for whole genome sequencing (ECSA clade: P10: CHIKV load of 1.03×10^5 copies/ml blood and Indian Ocean clade: P19: CHIKV load of 4.25×10^5 copies/ml blood).

Mice inoculation

Since viral RNA extracted from serum of 5 ml patient blood was insufficient in quantity for whole genome sequencing of CHIKV, copy number of CHIKV was increased by intracerebrally single-passaging infected patient's sera in mice brain. After obtaining approval from the animal research ethical board of Calcutta School of Tropical Medicine, which followed the guidelines of Committee for the Purpose of Control and Supervision on Experiments on Animals (CPCSEA), Government of India (registration no.: 681/02/a/CPCSEA), Swiss albino mice (New Zealand strain) were

used in this study (reference no.: AREC-STM/53). The study strictly adhered to the CPCSEA animal use and care protocol. Zero–six-day-old suckling mice were intracerebrally inoculated in triplicate with 10 µl of the above-mentioned patient sera to propagate viral replication (Lee et al. 2005). Mice exhibiting severe hind limb swelling along with visible cramping, rash and retarded growth within 5–6 days post inoculation were euthanized and brains were surgically retrieved.

RNA extraction and cDNA conversion

One hundred microliters of total RNA were extracted from these mouse brains, using Trizol Reagent according to manufacturer's protocol (Invitrogen, USA). The cDNA was synthesized in VERITI 96-well thermal cycler (Applied Biosystems, USA) using 10 µl of DNase treated RNA and RevertAid First Strand cDNA Synthesis Kit, according to manufacturer's protocol (Thermo Scientific, USA).

CHIKV whole genome sequencing and assembly

Overlapping primers ($n = 25$) listed in Table S1 were designed based on CHIKV nucleotide sequences available in the GenBank database (Table S2) of National Institute of Health, USA, using Primer3 software (Untergasser et al. 2012). PCR reagents, primer size and annealing temperature were mentioned in Table S1. PCR products (size range: 237–778 bp) were visualized by ethidium bromide staining of agarose gel, purified using QIAquick PCR Purification Kit (Valencia, CA, USA) and sequenced using ABI-Prism Big Dye Terminator v3.1 Cycle sequencing kit (ABI-Perkin Elmer, Nordrhein-Westfalen, Germany) in an ABI-Prism 3100 Avant Genetic Analyser (ABI-Perkin Elmer, Nordrhein-Westfalen, Germany). All amplicons were sequenced on both strands. Quality of base calling was checked by PHRED score analysis using ABI sequence analysis software V.5.1.1. Sequencing was done using cDNAs obtained from each of the three sets of suckling mice brain. As an attempt to nullify presence of any viral mutation occurring due to mice passage, partial sequences of CHIKV E1 region (nt 9990–10,560), previously done with RNA directly obtained from patient sera, were matched with CHIKV whole genome sequences of P10 and P19 patients. Contig assembly of overlapping amplicons was performed using CodonCode Aligner 5.0.1 (CodonCode Corporation, MA, USA). Assembled sequences were submitted to GenBank database.

Predicted viral protein characterization

Single amino acid alterations might potentially affect viral protein stability and functionality. To identify the change in

viral protein due to amino acid substitution, Expassy translate tool available in Swiss Institute of Biotechnology (SBI) web server (<https://web.expasy.org/translate/>) was used to translate viral nucleotide sequence to corresponding amino acid sequence. Further translated amino acid sequences were analyzed using HOPE (<http://www.cmbi.ru.nl/hope/>) and PMUT (<http://mmb.pcb.ub.es/pmut2005/>) web servers available in Uniprot database. HOPE server provided knowledge about effect of mutations on protein's 3D structure that gave valuable insight into possible changes in protein function. PMUT allowed fast scanning of mutational hot spots and accurate prediction of pathological character of point mutations based on use of neural networks, trained with a large database of neutral mutations and pathological mutations. Furthermore, individual viral protein sequences were translated from the genome sequences using Expassy translate tool available in SIB web server. Physicochemical properties, viz., overall charges, theoretical isoelectric points (pI) and molar extinction coefficients of individual viral proteins were determined using EMBOSS Pepstats (Rice et al. 2000). Instability indices, aliphatic indices and grand averages of hydropathicity (GRAVY) of these proteins were computed using ProtParam Proteomics tool available at EXPASY server (Gasteiger et al. 2005). Relative flexibility of each position within the proteins was estimated using Average Flexibility Index tool available within ProtScale server. Average antigenicity of viral proteins was computed using Kolaskar and Tongaonkar prediction method (1990) (Linding et al. 2003). Protein globularity was determined by GlobPlot v2.0 server.

Phylogenetic analysis

One hundred and thirty whole genome CHIKV sequences available in GenBank database till July 2014 were obtained along with three O'nyong-nyong virus (ONNV from each genotype) genome. To avoid laboratory artifacts, sequence of cloning vector/vaccine and high-passaged strains (Ross, S27, Angola/M2022/1962, India/MH4/2000, India/ALSA-1/1986) were excluded. This led to a final dataset of 126 CHIKV (including 9 West African, 23 Asian and 94 ECSA genotypes) and three ONNV sequences (Table S2). Multiple sequence alignment was performed using MUSCLE (Edgar 2004). Only open reading frames (ORFs) were considered for phylogenetic analyses, due to inexplicit alignment of 3'UTR and 5'UTR.

Two phylogenetic trees were inferred using maximum-likelihood (ML) method in MEGA 5.2.2 (Tamura et al. 2011) and metropolis-coupled Markov Chain Monte Carlo (MCMCMC) method in MrBayes v3.2 (Ronquist and Huelsenbeck 2003; Larget and Simon 1999) (Fig. 1). Best-fit models for ML and Bayesian analysis were determined using MEGA and jModelTest (Darrriba et al. 2012),

Evolutionary rates and times to the most recent common ancestors (tMRCA) of overall as well as individual clades were estimated using Bayesian Markov Chain Monte Carlo (MCMC) method in BEAST v1.8.0 (Drummond et al. 2012). Analyses were performed using an uncorrelated lognormal relaxed molecular clock (each branch have its own evolutionary rate) and SRD06 nucleotide substitution model that has been suggested to be superior for RNA virus and a Bayesian skyline coalescent tree prior is used to understand the influence of past population dynamics. Statistical uncertainty in parameter estimates was reflected as 95% highest probability density (HPD) values. Maximum clade credibility tree was computed using TreeAnnotator program available in BEAST package (Drummond et al. 2012). Regression analysis of root-to-tip genetic distance against sampling dates was performed using Path-O-Gen program (<http://tree.bio.ed.ac.uk/software/pathogen/>) to assess reliability of substitution rates and tMRCA estimates. Since both of the Bayesian methods yielded similar results, we chose the MCC tree generated by BEAST to construct the cladogram shown in Fig. 1.

Epitope binding prediction analyses

Chikungunya virus sequence from P10 and P19 and that of EF027141 (the CHIKV of Asian genotype detected in India during 1973) were submitted to array of immunoinformatics web servers to predict their interaction with MHC Class-I and Class-II alleles as well as towards linear and conformational B-cell epitopes. Furthermore, for consensus epitope prediction all the CHIKV structural and non-structural protein sequences were downloaded from the US National Institute of Allergy and Infectious Diseases (NIAID) Virus Pathogen Resource database (<https://www.viprbrc.org/>) and were subjected to multiple sequence alignment using default settings of ClustalW (<http://www.genome.jp/tools-bin/clustalw>) to locate different consensus regions. The consensus regions were analyzed using different in-silico epitope prediction softwares for identification of different immunogenic peptides (epitopes) and their antigenicity were determined by Vaxijen server version 2.0 (<http://www.ddg-pharmfac.net>).

MHC Class-I and -II binding prediction analysis

HLA-A, HLA-B, -DQ and -DR alleles with > 0.1 allele frequency among Indian population were retrieved from New Allele Frequency Database (<http://www.allelefrequencies.net/default.asp>). CBS NetCTLPan 1.1 (<http://www.cbs.dtu.dk/services/NetCTLPan/>) (based on artificial neural network algorithm) and NetMHCIIpan 3.0 (<http://www.cbs.dtu.dk/services/NetMHCIIpan/>) (based on quantitative matrix algorithm) web servers were used to predict nonameric viral peptide sequences binding to these HLA Class-I and Class-II

alleles, respectively. Results were filtered according to the position of amino acid changes and subsequently validated using ProPred-I (crdd.osdd.net/raghava/propred1/), NetMHC 3.4 (<http://www.cbs.dtu.dk/services/NetMHC-3.4/>), ProPred (crdd.osdd.net/raghava/propred/) and SYFPEITHI (<http://www.syfpeithi.de/>) web servers. Predicted peptide binding efficiencies towards these alleles were calculated using CBS NetCTLPan 1.1 and NetMHCIIpan 3.0, as percentile rank (on the basis of the mean value of the binding propensity scores), where values were interpreted to be either strong binders (% Rank < 0.5), weak binders (2 > % Rank > 0.5) or non-binders (% Rank > 2). Highest stringency threshold was used to minimize the number of false positives and to ensure significant binding affinity to all predicted peptides.

Linear and conformational B-cell epitope prediction

Linear B-cell epitopes were predicted using BCPREDS Server 1.0 (based on Support Vector Machine prediction algorithm) in conjunction with Bcepred web server (<http://ailab.ist.psu.edu/bcpred/predict.html>). Position of altered amino acids within the selected sets of viral protein sequences was used to filter the results.

CHIKV proteins were modeled using HHPred server (<https://toolkit.tuebingen.mpg.de/#/tools/hhpred>), while energy minimization and stereochemistry correction were performed using the CHARMM27 force field available in GROMACS package and KoBaMIN online server (Rodrigues et al. 2012). Protein loop correction was carried out using ModLoop server (<https://modbase.compbio.ucsf.edu/modloop/>). Presence of conformational epitopes on viral proteins was predicted using Discotope 2.0 server (<http://www.cbs.dtu.dk/services/Discotope/>) and validated using EPSVR server (<http://sysbio.unl.edu/EPSVR/>). Epitope propensity was calculated using Discotope 2.0 server, in terms of discotope score with cutoff value set at > -3.7 (specificity: > 0.75). Epitope mapping on modeled proteins was performed using PyMOL package (Grell et al. 2006).

Results

Whole genome sequences of CHIKV from these two patients were submitted to the GenBank database of National Institute of Health, USA (P10: KJ679577 and P19: KJ679578). Compared to EF027141 (Asian genotype detected in India during 1973), KJ679577 and KJ679578 (ECSA genotype) differed at 97 and 99 amino acids positions, respectively. Most of the alterations were observed in E2 and nsP3 protein regions of KJ679577 and KJ679578—each comprising of approximately 20% of the total changes. Overall, 23 amino acid alterations between EF027141 and KJ679577/KJ679578 were observed at functionally important sites of

Table 1 Physicochemical properties of chikungunya viral proteins

Protein	Genome	Hydropathicity ^a (GRAVY Scale)	Isoelectric point (pI)	Instability index ^b	Aliphatic index	Average antigenicity ^c
NSP1	EF027141	-0.323	7.3	36.27	80.58	1.0366
	KJ679578	-0.333	7.9	35.54	80.58	1.0363
	KJ679577	-0.334	7.0	37.14	79.50	1.0351
NSP2	EF027141	-0.271	8.9	38.13	86.79	1.0376
	KJ679578	-0.285	9.1	37.36	86.40	1.0366
	KJ679577	-0.279	9.1	37.76	86.16	1.0370
NSP3	EF027141	-0.194	4.5	48.90	87.41	1.0364
	KJ679578	-0.360	4.6	43.91	74.46	1.0304
	KJ679577	-0.355	4.6	44.37	74.87	1.0315
NSP4	EF027141	-0.194	7.4	48.90	87.41	1.0364
	KJ679578	-0.190	7.2	48.26	87.25	1.0367
	KJ679577	-0.202	7.2	48.68	86.60	1.0370
C	EF027141	-0.949	11.0	37.99	59.58	1.0057
	KJ679578	-0.973	10.9	42.10	58.70	1.0048
	KJ679577	-0.977	10.9	40.86	58.31	1.0059
E3	EF027141	-0.292	7.9	90.98	70.16	1.0615
	KJ679578	-0.586	6.8	78.49	57.24	1.0438
	KJ679577	-0.281	6.8	66.08	74.69	1.0623
E2	EF027141	-0.427	7.9	35.44	72.83	1.0356
	KJ679578	-0.406	8.0	38.63	71.76	1.0332
	KJ679577	-0.338	8.1	36.68	75.91	1.0380
6K	EF027141	0.949	8.0	43.43	131.31	1.1063
	KJ679578	1.013	6.8	42.89	130.81	1.1122
	KJ679577	1.013	6.8	43.04	130.81	1.1114
E1	EF027141	-0.010	6.5	36.04	77.82	1.0546
	KJ679578	-0.095	6.5	38.29	74.94	1.0495
	KJ679577	0.016	6.7	36.51	80.39	1.0561

^aHydropathicity (GRAVY Scale): value > 1.00 = hydrophobic; value < 1.00 = hydrophilic

^bInstability index: value < 40 = stable; value > 40 = unstable

^cAverage antigenicity (Kolaskar and Tongaonkar Scale): value > 1.0 = antigenic

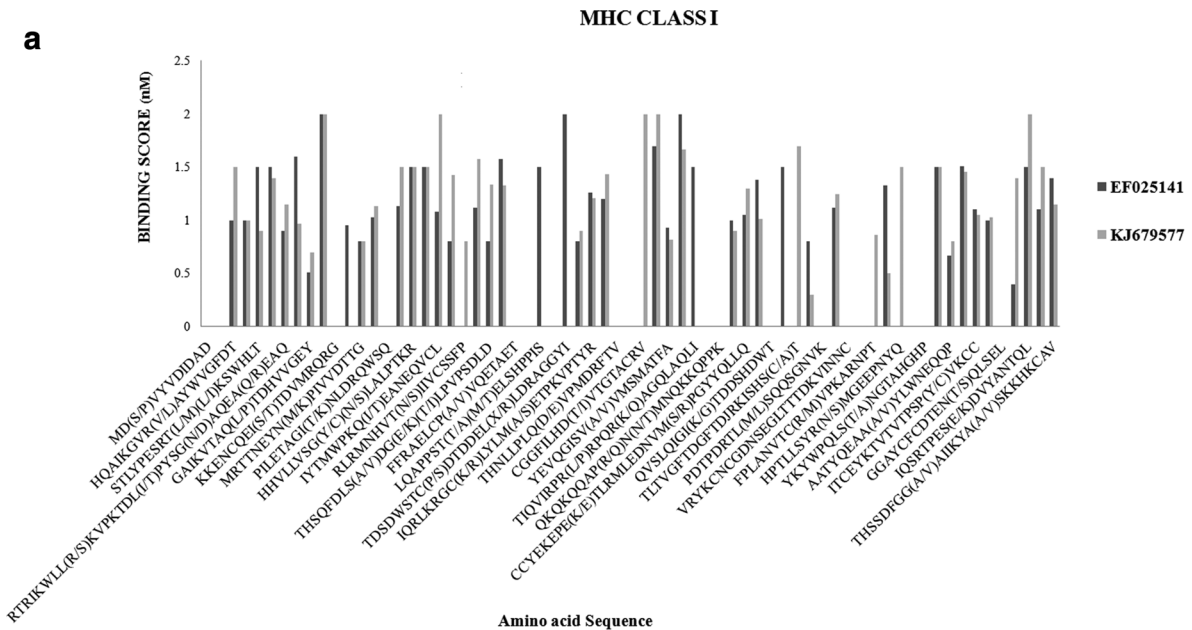
the viral proteins. Among them, Ile1Thr was located at the N-linker region (E2), Arg148Lys, Ala156Val and Val245Ala were located at the β -ribbon connector region (E2), involved in important E2–E1 interaction. Thr179Ile, Ser218Thr, Pro230Ser, Met338Lys and His374Tyr alterations were located at the helicase and Thr493Lys and Val604Ala at the peptidase domains of nsP2. Ala369Thr change was noted at the catalytic domain of nsP4. Among them, amino acid alterations at positions 1, 148 and 245 of E2; 179, 230, 338 and 493 of nsP2 and 369 of nsP4 were predicted to be located within the disordered region by HOPE server. Moreover, alterations at positions 179 and 493 of nsP2 and 369 of nsP4 were indicated to be pathological in nature by PMut software. Amino acid alterations at position 1 of E2, 604 of nsP2 and 369 of nsP4 were predicted to affect respective protein functions by SIFT software. Analysis of predicted physicochemical properties of CHIKV proteins among the three variants revealed: isoelectric points lie between 4.5

and 11.0; instability indices lie between 35.44 and 90.98; aliphatic indices ranged from 58.31 to 131.1; hydropathicity indices ranged between -0.977 and 1.013 and average antigenicity lie between 1.0048 and 1.1122 (Table 1). Interestingly, both EF027141 and KJ679578 harbored one opal codon (UGA) at the C-terminus of nsP3 which was replaced by arginine (CGA) in KJ679577.

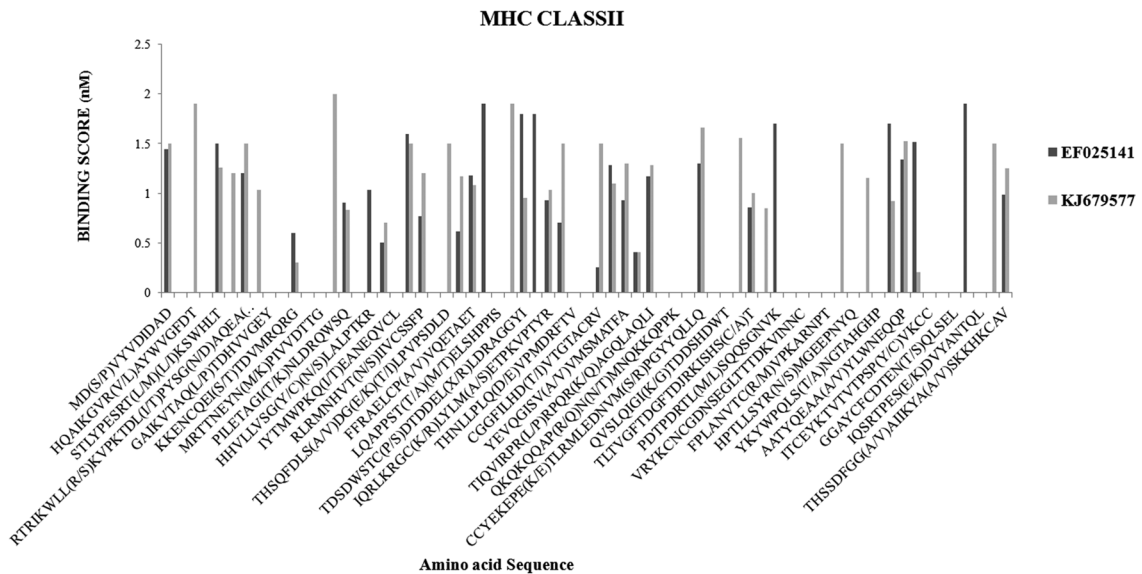
Phylogenetic cladistic positions of the two CHIKV isolates indicated their different individual lineages within ECSA genotype, where KJ679578 and KJ679577 clustered

Fig. 2 a Graphical representation of the comparative binding affinity of EF025141 vs KJ679577 towards MHC Class-I and -II molecules. Binding score ranges from 0 to 0.5 for strong binders and > 2 for non-binders. **b** Graphical representation of the comparative binding affinity of EF025141 vs KJ679578 towards MHC Class-I and -II molecules. Binding score ranges from 0 to 0.5 for strong binders and > 2 for non-binders

a

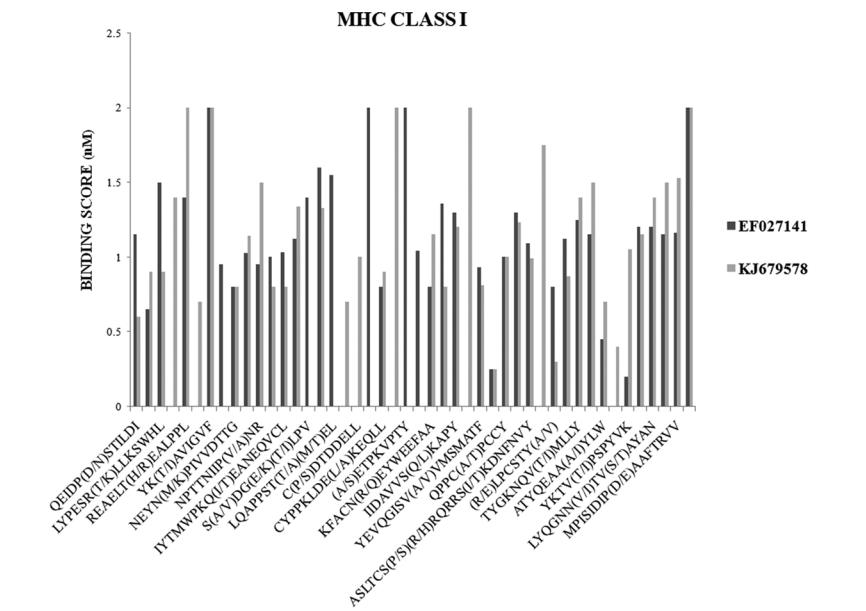


Change in MHC Class I Epitopes in KJ679577 with respect to EF027141	Binding Propensity Increased	Binding Propensity Decreased	Binding Lost	Binding Gained	Binding Propensity Unchanged
%	25.37 (17/67)	22.83 (15/67)	4.4 (3/67)	6.9 (6/67)	35.82 (24/67)



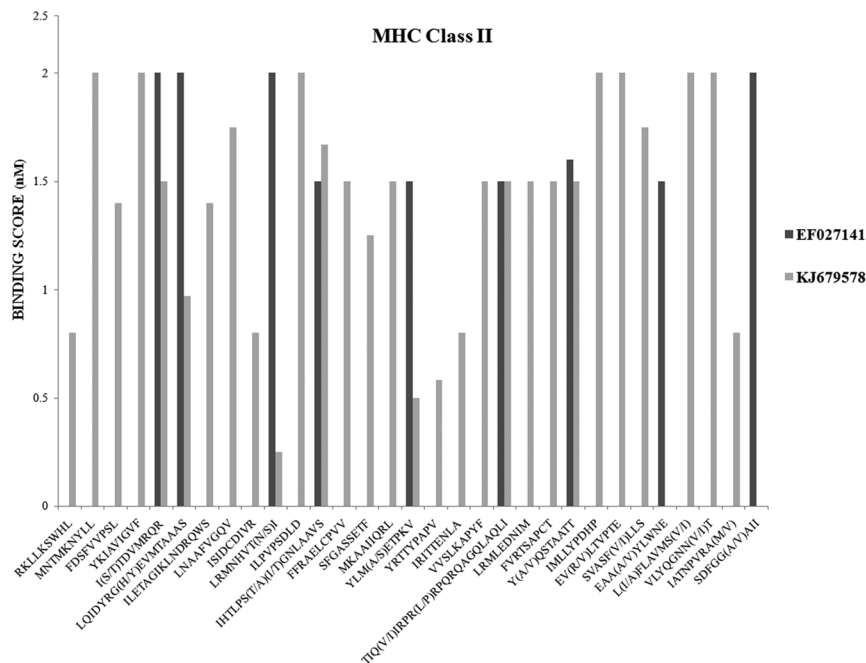
Change in MHC Class II Epitopes in KJ679577 with respect to EF027141	Binding Propensity Increased	Binding Propensity Decreased	Binding Lost	Binding Gained	Binding Propensity Unchanged
%	16.1 (11/67)	19.4 (13/67)	5.9 (4/67)	14.9 (10/67)	41.7 (28/67)

b



Amino acid Sequence

Change in MHC Class I Epitopes in KJ679578 with respect to EF027141	Binding Propensity Increased	Binding Propensity Decreased	Binding Lost	Binding Gained	Binding Propensity Unchanged
%	28.2 (13/46)	30.4 (14/46)	13 (6/46)	17.4 (8/46)	10.8 (5/46)



Amino acid Sequence

Change in MHC Class II Epitopes in KJ679578 with respect to EF027141	Binding Propensity Increased	Binding Propensity Decreased	Binding Lost	Binding Gained	Binding Propensity Unchanged
%	16.1 (5/31)	3.2 (1/31)	6.45 (2/31)	70.9 (22/31)	3.2 (1/31)

Fig. 2 (continued)

with CHIKV isolates of 2006 Indian Ocean outbreak and ECSA subtypes respectively (Fig. 1). KJ679578 exhibited higher substitution rate compared to its nearest neighbors. Average substitution rates of Indian Ocean epidemic (4.5×10^{-4} subs/nt/year; 95% HPD: 3.3×10^{-4} – 5.6×10^{-4} subs/nt/year) and Asian endemic (4.36×10^{-4} subs/nt/year; 95% HPD: 3.69×10^{-4} – 5.1×10^{-4} subs/nt/year) CHIKV lineages were significantly higher than that of two enzootic CHIKV lineages, West African (2.0×10^{-4} subs/nt/year; 95% HPD: 4.6×10^{-5} – 3.5×10^{-4} subs/nt/year) and ECSA (2.75×10^{-4} subs/nt/year; 95% HPD: 2.0×10^{-4} – 3.4×10^{-4} subs/nt/year). Substitution rate of KJ679578 (8.74×10^{-4} subs/nt/year) was twenty-two fold higher than that of KJ679577 (3.97×10^{-5} subs/nt/year). Our estimated time of origin of CHIKV lie within the last 500 years, with tMRCA of West African, ECSA, Asian and Indian Ocean epidemic lineages predicted to occur approximately 85 (95% HPD: 41–178 years), 63 (95% HPD: 59–72 years), 57 (95% HPD: 54–61 years) and 44 (95% HPD: 35–62 years) years earlier, respectively. Indian Ocean and Asian lineages had positive selection pressure at ten (nsP1: 171, 314, 393, 488, nsP3: 334, 452, E2: 382, 430, E1: 211, 304) and six (nsP3: 437, 520, nsP4: 84, E3: 62, E1: 304, 397) amino acid positions, respectively, whereas none was found in case of West African and ECSA lineages.

Among the Indian population, the following MHC alleles had allele frequency of > 0.1: ten MHC Class-I alleles within eight super types, viz., A*01:01, A*01:02, A*02:01, A*11:01, A*24:02, B*08:01, B*15:01, B*35:01, B*51:01, B*40:01 and 13 MHC Class-II HLA-DR and -DQ alleles, viz., DRB1*01:01, DRB1*03:01, DRB1*04:01, DRB1*07:01, DRB1*11:01, DRB1*13:01, DRB1*15:01, DRB1*15:02, DQA1*02:01-DQB1*02:01, DQA1*03:01-DQB1*03:01, DQA1*04:01-DQB1*04:01, DQA1*05:01-DQB1*05:01 and DQA1*06:01-DQB1*06:01. In-silico immunogenetic approach has been used to study the effect

of viral amino acid substitutions on immunogenicity of recent strains KJ679577 and KJ679578 (of ECSA genotype), circulating in Eastern India during 2011–2012 outbreaks with that of EF027141 (of Asian genotype) and previously circulating strain in Eastern India during 1973. Comparison of these three sequences using different T- and B-cell epitope prediction servers depicted an increase in number of MHC Class-II interacting epitopes (70.9%) and their binding propensity (16%) of KJ679578 than that of EF027141 (Fig. 2a, b). However, only 15–16% increase in interacting epitope number and binding propensity has been observed in comparison of KJ679577 with EF02714. Interestingly, KJ679577 and KJ679578 exhibited only 25–28% increase in interacting epitope number and binding propensity against MHC Class-I allele when compared to EF027141.

Five promiscuous epitopes within non-structural regions of KJ679578, KJ679577 and EF027141 were found against Class-II alleles of Indian population. In addition, KJ679578 and KJ679577 harbored one such epitope each at nsP4 and nsP2 regions respectively. E2 and 6K regions of all three CHIKV strains harbored three such epitopes which could bind with all Class-II alleles promiscuously. No such epitope against Class-I alleles was found within all three viruses.

KJ679578 differed from EF027141 at seventy-nine (79) linear B-cell epitopes, approximately 60% of which clustered at non-structural regions of CHIKV (Fig. 3). Among them, two epitopes at nsP1 region were of high propensity. Frequency of B-cell epitopes was more than two fold higher in KJ679578 ($n = 53$) compared to EF027141 ($n = 23$), most of which clustered at nsP1, nsP2, nsP4 and E1 regions. Similar directionality was observed on comparing linear B-cell epitopes of KJ679577 with respect to EF027141. Seventy-one (71) linear B-cell epitopes, predominantly clustering at non-structural region were different between them. Among them, two epitopes at nsP1 and one at C regions were of high propensity. Forty-nine linear

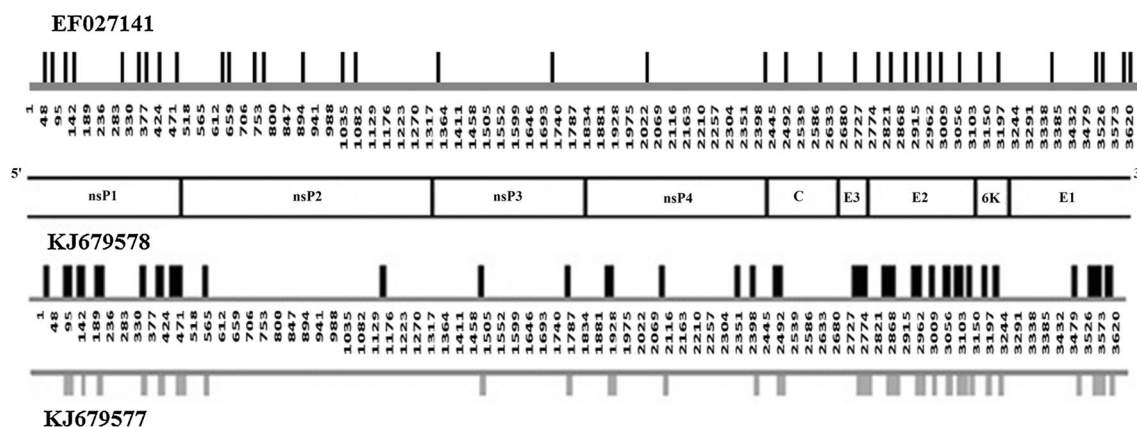


Fig. 3 Predicted distribution map of continuous B-cell epitopes of chikungunya virus. Location of the epitope has been mentioned between EF027141, KJ679578 and KJ679577

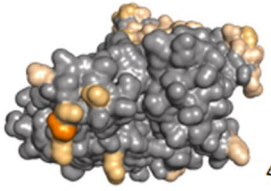
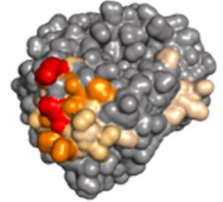
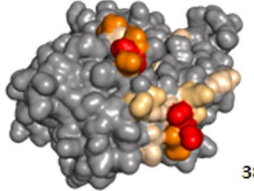
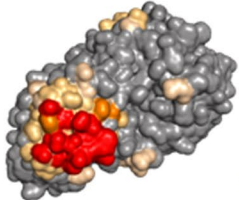
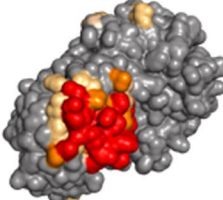
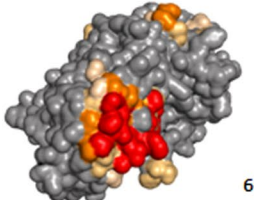
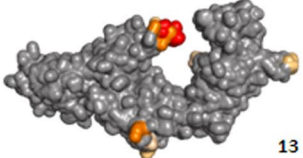
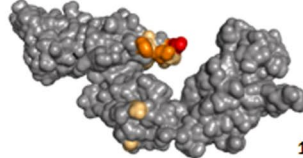
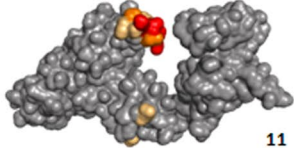
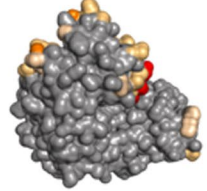
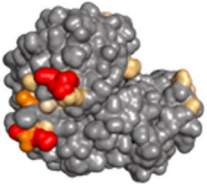
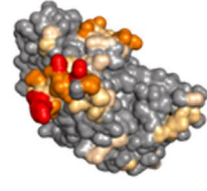
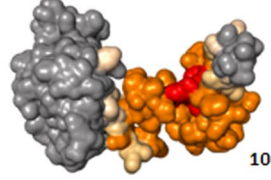
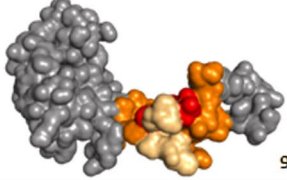
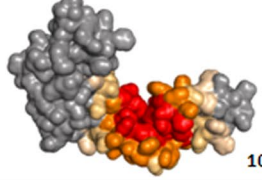
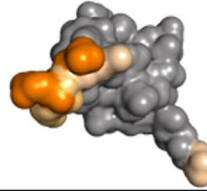
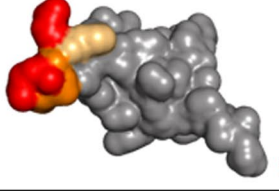
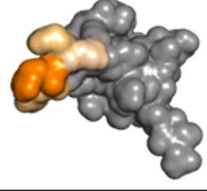
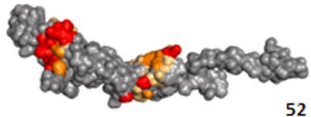
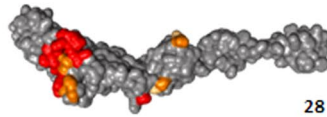
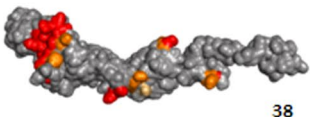
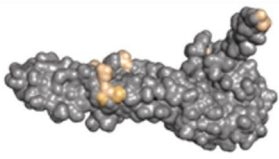
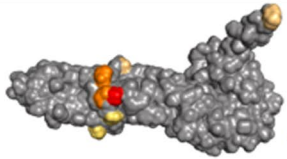
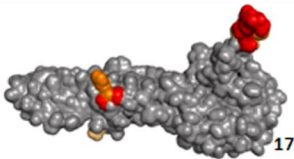
PROTEIN	EF027141	KJ679577	KJ679578
NSP1	 46	 35	 38
NSP2	 62	 62	 66
NSP3	 13	 12	 11
NSP4	 46	 30	 63
CAPSID	 100	 93	 104
E3	 6	 5	 5
E2	 52	 28	 38
E1	 8	 7	 17

Fig. 4 Predicted distribution map of conformational B-cell epitopes within individual CHIKV proteins of EF027141, KJ679577 and KJ679578. Each viral protein has been represented as space fill model. Red epitopes are with discotope score >0.5 , orange epitopes are with discotope score >-1.0 , yellow epitopes are with discotope score >-2.5 and wheat epitopes are with discotope score >-3.7 . Total number of these epitopes in each protein mentioned in bottom right corner of each cell

epitopes were exclusively present in KJ679577, while only seventeen were present in EF027141. Generation of novel epitopes in both non-structural (nsP1, nsP2 and nsP4) and structural (C, E2 and E1) regions was predicted to be more than two fold higher in KJ679577 compared to EF027141.

Number of B-cell conformational epitopes was highest at C and lowest at E3 regions of KJ679578, KJ679577 and EF027141 (Fig. 4 and Table S3). The nsP1 region of each of KJ679577 and KJ679578 harbored three highly antigenic conformational epitopes, which were completely absent in EF027141 (discotope score >0.5). Similarly, number of such epitopes in the C and E1 region of KJ679578 was significantly greater than that of KJ679577 and EF027141. Interestingly, highly antigenic epitopes were completely absent at nsP1, E3 and E1 regions of EF027141.

To identify peptide for construction of good vaccine candidature, the consensus amino acid sequence of all structural and non-structural proteins of CHIKV were analyzed using ProPred MHC Class-II binding prediction server for determining conserved immunogenic peptide across all CHIKV strains. A total of 32 epitopes were predicted using ProPred server, all of which interacted with minimum six HLA Class-I alleles, whereas 24 epitopes interacted with at least one HLA Class-II allele (Tables 2, 3). Approximately, 37.5% (12/32) epitopes were clustered in the structural region, among which four epitopes were predicted in E1 and E2, three in capsid and only one in 6K regions. Interestingly, 62.5% (20/32) of predicted epitopes were located within the non-structural region, viz., nsP1: 8 epitopes, nsP2: 5 epitopes, nsP3: 3 epitopes and nsP4: 4 epitopes. Notably, peptide DLAKLAFKRSSKYDLECAQIPVHMKSDA, located within capsid region of structural polypeptide was found to interact with the maximum number of HLA Class-I (68%; 32/47) and HLA Class-II (47%; 24/51) alleles with an antigenicity score 0.897. Peptide sequence PEDAQKLLVGLNQRIVVNGRTQRN, located in the nsP1 region of non-structural polyprotein, showed interaction with the highest number of HLA Class-I (85.1%; 40/47) and HLA Class-II (88.2%; 45/51) alleles with an antigenic score 0.553. Peptide KPGRRRERMCMKIEN, located within the capsid region showed highest antigenic score (1.885) among all predicted epitopes, but it only interacts with 19.14% (9/47) HLA Class-I alleles.

A good vaccine candidate should contain epitopes which are capable to induce both T-cell and B-cell mediated immunity. All the previously predicted T-cell epitopes were analyzed for B-cell epitope identification using BCPred server and were shortlisted according to BCPred score. A total of 11 epitopes were predicted for all proteins, among which four were located in the structural region (three epitopes were predicted for capsid region and one epitope for E2 region) and seven in non-structural region (three epitopes were predicted for nsP1 and two epitopes for each of nsP2 and nsP4) (Table 4). Peptide KPGRRRERMCMKI located within capsid region showed highest BCPred score (0.92) as well as antigenic score (1.84). Interestingly, according to overall score, binding propensity and antigenicity predicted peptide with amino acid sequence DLAKLAFKRSSKYDLECAQIPVHMKSDA (located in the capsid region) and KVVLCGDPKQCGFFNMMQMKYNYNHNI (located in the nsP2 region) showed best interaction with both T- and B-cells and hence might be considered for development of universal vaccine and diagnostic kits against CHIKV infection.

Discussion

Along with other reports, our previous work indicated simultaneous circulation of two different CHIKV strains of ECSA genotype among Indian patients. This is the first study reporting whole genome sequences of two recent CHIKV strains isolated from Eastern India and their genome comparison with the CHIKV of Asian genotype, circulated in India between 1963 and 1973 (Cecilia 2014). Similar to previous studies, this study also demonstrated clear division of CHIKV ECSA genotypes, with most of the recent CHIKV epidemic strains clustering within Indian Ocean lineages (Volk et al. 2010; Casal et al. 2015).

The amino acid alterations (Val264Ala and Met267Arg) of KJ679577 and KJ679578 (ECSA genotype), located within β -ribbon connector (arch 2) of E2 might affect E2–E1 interactions, thereby potentially influencing viral spike formation (Voss et al. 2010). Earlier study demonstrated association of E1-A226V and E2-K252Q mutations with increased virus fitness towards *Ae. albopictus*; however, these alterations were absent in KJ679577 and KJ679578 suggesting their transmission via mainly *Ae. aegypti*. Interestingly, combination of amino acid substitution, E2: Val264Ala, E1: Lys211Glu and E1: Val226Ala has been reported to be associated with increased CHIKV fitness, infectivity, dissemination and transmission towards *Ae. aegypti*. Presence of all the three substitution within KJ679578 (E2: Val264Ala, E1: Lys211Glu and E1: Val226Ala) and E1: Val226Ala in KJ679577 might support transmission of both these strains mainly by *Ae. aegypti* (Agarwal et al. 2016). Changes within

- Casal PE, Chouhy D, Bolatti EM, Perez GR, Stella EJ, Giri AA (2015) Evidence for homologous recombination in chikungunya virus. *Mol Phylogenet Evol* 85:68–75
- Cecilia D (2014) Current status of dengue and chikungunya in India. WHO South-East Asia. *J Public Health*. <https://doi.org/10.4103/2224-3151.115828>
- Chen KC, Kam YW, Lin RT, Ng MM, Ng LF, Chu JJ (2013) Comparative analysis of the genome sequences and replication profiles of chikungunya virus isolates within the East, Central and South African (ECSA) lineage. *Virology* 10:169
- Darriba D, Taboada GL, Doallo R, Posada D (2012) jModelTest 2: more models, new heuristics and parallel computing. *Nat Methods* 9:772
- Drummond AJ, Suchard MA, Xie D, Rambaut A (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol* 29:1969–1973
- Dutta SK, Pal T, Saha B, Mandal S, Tripathi A (2014) Copy number variation of chikungunya ECSA virus with disease symptoms among Indian patients. *J Med Virol* 86:1386–1392
- Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792–1797
- Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, Bairoch A (2005) Protein identification and analysis tools on the ExPASy server. In: Walker JM (ed) *The proteomics protocols handbook*. Humana Press, New York, pp 571–607
- Grell L, Parkin C, Slate L, Craig PA (2006) EZ-Viz, a tool for simplifying molecular viewing in PyMOL. *Biochem Mol Biol Educ* 34:402–407
- Kam YW, Lum FM, Teo TH, Lee WW, Simarmata D, Harjanto S, Chua CL, Chan YF, Wee JK, Chow A, Lin RT, Leo YS, Le Grand R, Sam IC, Tong JC, Roques P, Wiesmüller KH, Réna L, Röttschke O, Ng LF (2012) Early neutralizing IgG response to chikungunya virus in infected patients targets a dominant linear epitope on the E2 glycoprotein. *EMBO Mol Med* 4:330–343
- Knutson KL, Schiffman K, Disis ML (2001) Immunization with a HER-2/neu helper peptide vaccine generates HER-2/neu CD8 T-cell immunity in cancer patients. *J Clin Invest* 107:477–484
- Larget B, Simon D (1999) Markov chain monte carlo algorithms for the bayesian analysis of phylogenetic trees. *Mol Biol Evol* 16:750–759
- Lee YR, Huang KJ, Lei HY, Chen SH, Lin YS, Yeh TM, Liu HS (2005) Suckling mice were used to detect infectious dengue-2 viruses by intracerebral injection of the full-length RNA transcript. *Intervirology* 48:161–166
- Linding R, Russell RB, Neduva V, Gibson TJ (2003) Gibson GlobPlot: exploring protein sequences for globularity and disorder. *Nucleic Acids Res* 31:3701–3708
- Pybus OG, Rambaut A (2009) Evolutionary analysis of the dynamics of viral infectious disease. *Nat Rev Genet* 10(8):540–550
- Rice P, Longden I, Bleasby A (2000) The European molecular biology open software suite. *Trends Genet* 16:276–277
- Rodrigues J, Levitt M, Chopra G (2012) KoBaMIN: a knowledge-based minimization web server for protein structure refinement. *Nucleic Acids Res* 40:323–328
- Ronquist F, Huelsenbeck JP (2003) MRBAYES 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19:1572–1574
- Scholte FE, Tas A, Martina BE, Cordioli P, Narayanan K, Makino S, Snijder EJ, van Hemert MJ (2013) Characterization of synthetic chikungunya viruses based on the consensus sequence of recent E1-226V isolates. *PLoS One* 8:e71047. <https://doi.org/10.1371/journal.pone.0071047>
- Singh RP, Singh RN, Srivastava MK, Srivastava AK, Kumar S, Dubey RC, Sharma AK (2012) Structure prediction and analysis of MxA^F from obligate, facultative and restricted facultative methylobacterium. *Bioinformation* 8:1042–1046
- Strauss JH, Strauss EG (1994) The alphaviruses: gene expression, replication, and evolution. *Microbiol Rev* 58(3):491–562
- Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* 28:2731–2739
- Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG (2012) Primer3—new capabilities and interfaces. *Nucleic Acids Res* 40:e115
- Volk SM, Chen R, Tsetsarkin KA, Paige Adams A, Garcia TI, Sall AA, Nasar F, Schuh AJ, Holmes EC, Higgs S (2010) Genome-scale phylogenetic analyses of chikungunya virus reveal independent emergences of recent epidemics and various evolutionary rates. *J Virol* 84:6497–6504
- Voss JE, Vaney MC, Duquerroy S, Vonrhein C, Girard-Blanc C, Crublet E, Thompson A, Bricogne G, Rey FA (2010) Glycoprotein organization of chikungunya virus particles revealed by X-ray crystallography. *Nature* 468:709–712