4. OTHER SCIENTIFIC STUDIES/CENTRES

4.1 Biomedical Informatics Centre

Project Coordinator: *Dr. G. S. Toteja, Scientist 'G' & Director*
Principal Investigator: *Dr. Vinod Joshi, Scientist 'G'*
Principal co- Investigators: *Dr. Manju Singhi, Scientist 'D', Dr. P. K. Anand, Scientist 'D', Dr. S. S. Mohanty, Scientist, 'D'*

Commencement: April, 2013 **Duration**: Five Years **Status:** Ongoing
Funding agency: Indian Council of Medical Research

OBJECTIVES

1. To identify genetic loci associated with diseases of National interest such as Diabetes, Cancer, Stress, Mental illness etc. in Indian population.
2. To develop solutions for controlling pathogens causing diseases of National interest
3. Such as Tuberculosis, Malaria, and AIDS etc.
4. To develop a National Repository of clinical information/data, high-throughput data, Genotype and phenotype.
5. To promote applications of cutting-edge technologies in medical research.

PROGRESS

Development of highly specific and sensitive primers for Sanger technique based protocol for whole genome sequencing of Dengue virus

Dengue virus is a positive sense single stranded RNA virus belongs to class flaviviridae. Dengue virus genome consists of nearly 10,700 nucleotides. The genome is linear and non-segmented and encodes a single long open reading frame (ORF), flanked by highly structured 5′ and 3′ untranslated regions (UTRs). The N-terminal of the polyprotein encodes the three structural proteins Capsid protein (C), Membrane protein (prM/M), Envelope protein (E), followed by at least seven non-structural (NS) proteins (NS1-NS2A-NS2B-NS3-NS4A-NS4B-NS5) totaling to 10 proteins. There are also reports which describe functionality arising due to secondary RNA structure. There are four closely related serotypes of DENV (DENV-1 to DENV-4). For successful whole genome sequencing of dengue virus serotype using sanger based method, availability of highly specific and sensitive primers covering whole genome is a critical determinant.There exists spatiotemporal genomic diversity in individual serotypes leading to auxiliary categorization in lineages and genotypes. Also, evidences of concurrent infections of multiple dengue virus serotypes in a single host further adds to complexity as there

is observed 60 to 75% similarity in dengue virus serotypes at amino acid level. During design and selection of primers these complexities needs to be prevailed over. We have developed indigenous primers for whole genome sequencing of dengue virus subtypes based upon novel methodology of selection and design of primers utilizing various databases and bioinformatics tools. We used available genomic data from public databases, bioinformatics & primer design tools and programs written in perl language to achieve the objective.

Whole Genome sequence assembly and analysis of Dengue virus type 3 from Rajasthan, India

Sequencing was performed for dengue virus type 3 isolated directly from serum and passage culture of same sample using indigenously designed primers. To obtain whole genome, reference based contig assembly of sequenced fragments was performed using SeqScape v2.6 (Applied Biosystem) software and also analyzed for nucleotide and amino acid variations compared to NCBI reference genome for Dengue virus type 3(Accession: NC_001475.2). Assembly was reviewed manually and adjustments were performed for observed discrepancies. Comparative genome analysis was also performed with nearest genome in phylogenetic analysis. Phylogenetic analysis and distance calculations were performed using the MEGA v.6 software with the Neighbor-Joining method of the Maximum Composite Likelihood model, uniform rates among sites with 1,000 bootstrap replicates. After assembly and annotation, Whole genomes were than submitted to NCBI Genbank using BankIt web based sequence submission platform. GenBank accession number for whole genome directly from sample was **KU216209** and from passaged culture of same sample as **KU216208.**

Structure analysis of envelope protein of dengue virus type 3 (DEN-3) from Rajasthan

P-Blast analysis:

Amino acid sequence of envelope (target) protein compared with the template sequence (1UZG-PDB ID). Alignment results showed there are few amino acid variations in target sequence. Variations are T95A, S124P, S164P, A169T, T270N and K383N. Target sequence contains 100% query coverage with 96% identity. Structure prepared using template PDB (1UZG). Structure predicted using Phyre2 webServer. Envelope protein of dengue virus type 3 and template PDB (1UZG) visualized by UCSF Chimera.

Pocket analysis:

Large pockets are frequently found to be the location of active sites. The largest pocket as detected by the fpocket2 program is shown in wireframe mode, coloured red. Pocket detection site in target sequence is: PRO39, THR40, VAL143, HIS144, THR145, GLY1 46, ASP147, GLN148, TYR176, LEU292, CYS293, TYR297, THR351, ALA352, ASN353, PRO 354, VAL355, THR357, GLU361, PRO362, VAL363, ASN364 and ILE365.

Prediction of N-Glycosylation site:

The four different types of dengue virus have slightly different genomes, which mean that they may have different protein structures because the genomic sequence for an envelope (E) protein will alter the amino acid sequence and the folding of the protein. Antibodies that neutralize one type of dengue virus may be ineffective against another dengue strain due to these structural differences so understanding structural differences is key to treating the virus. Herein, we predicted the two N-linked glycosylation sites on the glycoprotein's using NetNgly 1.0 Server. The two resides where this occurs are Asn-67 and Asn-153 using threshold level (0.5).These glycosylation sites help the viruses successfully to attack on cells. Possible antiviral inhibit proper formation of the viral proteins during replication in the host cell by stopping the N-linked glycosylation sites from forming on the new virus E proteins. Altering

the protein formation could affect the ability of the virus to enter a new host cell. These sites can be used as antiviral drug therapy.

Bioinformatics analysis of whole genome sequencing data of H1N1 virus samples of pandemic 2009 and in subsequent years from Rajasthan

H1N1 virus belongs to Influenza A viruses (IAV). The IAV H1N1 genome consists of 8 single-stranded RNA (ssRNA) segments and encodes 12 proteins: hemagglutinin (HA), M proteins (M1 and M2), neuraminidase (NA), nucleocapsid protein (NP), nonstructural proteins (NS1 and NS2), and polymerase subunits (PA, PA-X, PB1-F1, PB1-F2, and PB2). Whole genome sequencing was performed for four samples. 46 pairs of primer pairs as described in were used for sanger based whole genome sequencing of H1N1 virus. Around 400 individual sequences were analyzed for quality filtering, reference based assembly (pdm/H1N1/California/2009/07 strain genome) followed by genomic variation and amino acid mutation analysis was performed using SeqScape v 2.6 software (Applied Biosystems). Subsequently comparative genomics analysis and phylogenetic analysis were performed using SeqScape v 2.6 and MEGA6 software and other tools.