IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Genomic analysis

Rimpy Jain, Kavita, Garima Sharma, Minakshi

[1]Department of Pharmaceutical Science
P.D.M UNIVERSITY, Bahadurgarh, Haryana

Abstract : Genomics is the study of whole genomes of organisms, and incorporates elements from genetics. Genomics uses acombination of recombinant DNA, DNA sequencing methods, and bioinformatics to sequence, assemble, and analysethe structure and function of genomes

I. INTRODUCTION

Genomics divided into two basic areas: structural genomics, characterizing the physical nature of whole genomes; and functional genomics, characterizing the transcriptome (the entire range of transcripts produced by a given organism) and the proteome (the entire array of encoded proteins).
Genomics is the study of whole genomes of organisms, and incorporates elements from genetics. Genomics uses a combination of recombinant DNA, DNA sequencing methods, and bioinformatics to sequence, assemble, and analyse the structure and function of genomes.
The prime directive of structural genomic analysis is the complete and accurate elucidation of the DNA sequence of a representative haploid genome of a given species. When this sequence is known, it opens the door to numerous possibilities. By computational analysis of the sequence, using principles developed by genetic and molecular biological analysis of transcripts and proteins, we can make predictions of all of the encoded proteins. We can analyse other haploid genomes from the same species and develop a statistical picture of the genetic variation within populations of that species. We can compare the genomic sequence of different species and thereby gain an understanding of how the genome has been remodelled in the course of evolution. Studies of comparative genomics have already proceeded far enough to reveal that, in related species (for example, within all mammals), there is considerable synteny (conserved gene location within large blocks of the genome). Studies of comparative genomics also offer a powerful opportunity to identify highly conserved and therefore functionally important sequence motifs in coding and noncoding genomic DNA. This identification helps researchers confirm predictions of protein-coding regions of the genome and identify important regulatory elements within D

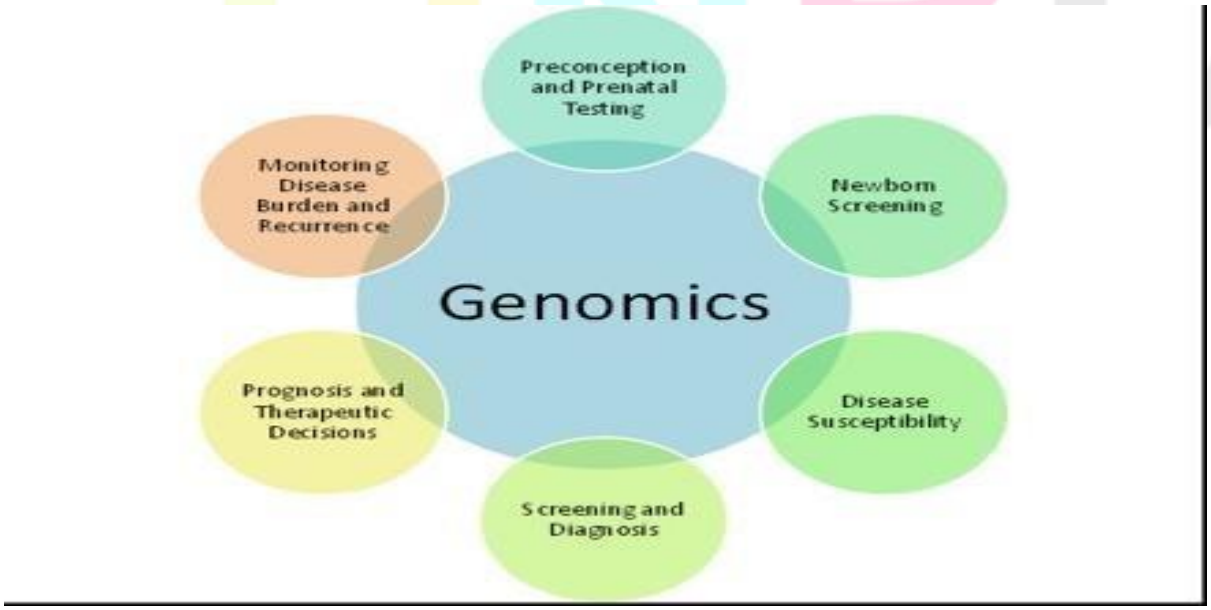

Fig-1

Structural genomics:
structural genomics is only a little more than a decade old and is already fulfilling the promise of providing complete sequences of many genomes, the leap from classical genetic maps to complete DNA sequence maps did not happen in a single bound. Rather, quite analogous to the way in which one proceeds through several increases in magnification on a light microscope, there was a step-by-step progression in genome -wide map resolution in the development of genomic technologies.only were these technologies invaluable steps on the way to the establishment of sequence-level maps, but they also proved to be extremely important tools in themselves for diseases gene identification and positional cloning.
Because protein structure is closely linked with protein function, the structural genomics has the potential to inform knowledge of protein function. In addition to elucidating protein functions, structural genomics can be used to identify novel protein folds and potential targets for drug discovery. Structural genomics involves taking a large number of approaches to structure determination, including experimental methods using genomic sequences or modelling-based approaches based on sequence or structural homology to a protein of known structure or based on chemical and physical principles for a protein with no homology to any known structure AS opposed to traditional , the determination of a through a structural genomics effort often (but not always) comes before anything is known regarding the protein function. This raises new challenges in structural bioinformatics, i.e. determining protein function from its 3D structure.
Structural genomics emphasizes high throughput determination of protein structures.

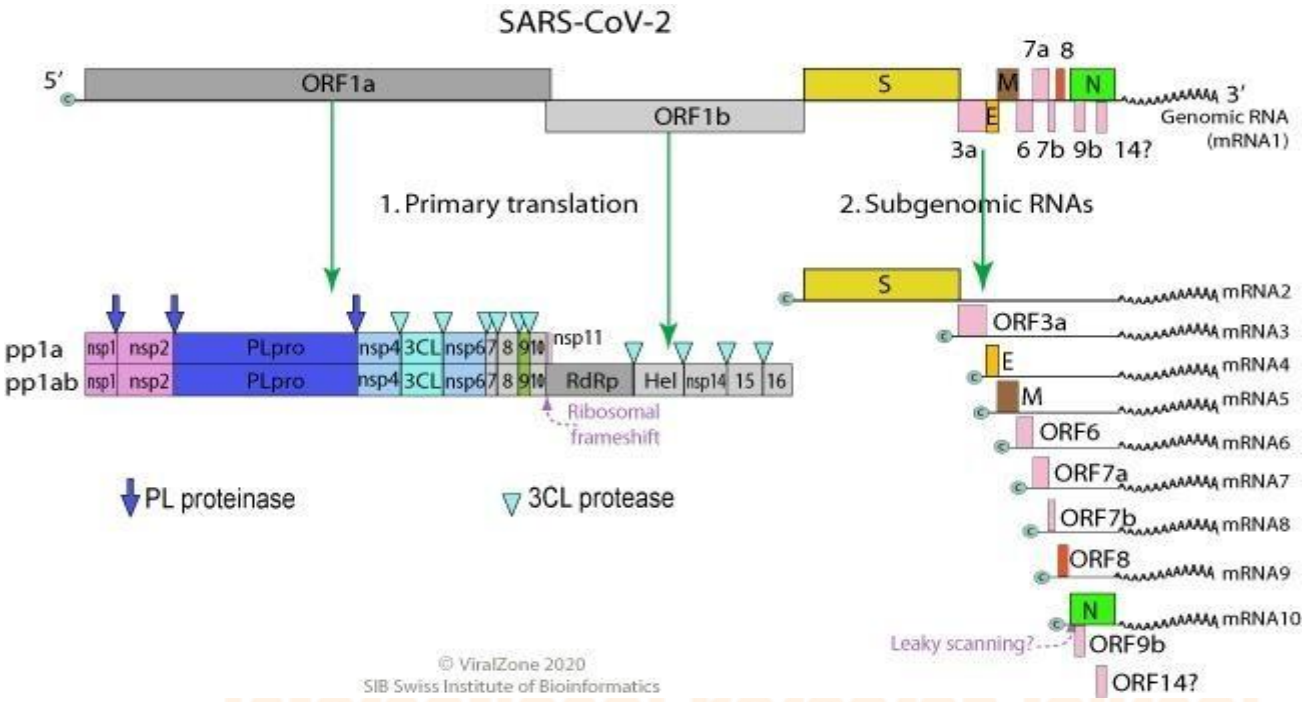

fig 2- sub genomic RNAs of SARS COV2

1

Advantages:
The protein structure initiative, is that the scientific community gets immediate access to new structures, as well as to reagents such as clones and protein.
Disadvantage:
Many of these structures are of proteins of unknown function and do not have corresponding publications. This requires new ways of communicating this structural information to the broader research community
The Bioinformatics core of the Joint centre for structural genomics (JCSG) has recently developed a wiki-based approach namely open protein structure annotations network (TOPSAN) for annotating protein structures emerging from high-throughput structural genomics centres.

Goals:
One goal of structural genomics is to identify novel protein folds. Experimental methods of protein structure determination require proteins that express and/or crystallize well, which may inherently bias the kinds of proteins folds that this experimental data elucidate.
A genomic, modelling-based approach such as ab initio modelling may be better able to identify novel protein folds than the experimental approaches because they are not limited by experimental constraints.
The high-throughput structure determination methods of structural genomics have the potential to inform our understanding of protein functions. This also has potential implications for drug discovery and protein engineering Furthermore, every protein that is added to the structural database increases the likelihood that the database will include homologous sequences of other unknown proteins. .

Examples of structural genomics:
1 Thermotogo meritima proteome
One current goal of the joint centre for structural genomics (JCSG), a part of protein structure initiative (PSI) is to solve the structures for all the proteins thermotogo meritima a thermophillic bacterium. T. meritima was selected as a structural genomics target based on its relatively small genome consisting of 1,877 genes and the hypothesis that the proteins expressed by a thermophillic bacterium would be easier to crystallize
2.. Mycobacterium tuberculosis proteome:
The fully sequenced genome of M. tuberculosis has allowed scientists to clone many of these protein targets into expression vectors for purification and structure determination by X-ray crystallography. Studies have identified a number of target proteins for structure determination, including extracellular proteins that may be involved in pathogenesis, iron-regulatory proteins, current drug targets, and proteins predicted to have novel folds. So far, structures have been determined for 708 of the proteins encoded by M. tuberculosis.

1 DPER,BPSMV

Functional genomics

Functional genomics is a field of molecular biology that attempts to describe gene (and protein) functions and interactions. Functional genomics make use of the vast data generated by genomic and transcriptomic projects (such as genome sequencing projects and RNA sequencing).
Functional genomics focuses on the dynamic aspects such as gene transcription, translation, regulation of gene expression and protein interaction, as opposed to the static aspects of the genomic information such as DNA sequence or structures. A key characteristic of functional genomics studies is their genome-wide approach to these questions, generally involving high-throughput methods rather than a more traditional “gene-by-gene” approach.

Goals:
The goal of functional genomics is to understand the function of genes or proteins, eventually all components of a genome. The term functional genomics is often used to refer to the many technical approaches to study an organism's genes and proteins, including the "biochemical, cellular, and/or physiological properties of each and every gene product"[2] while some authors include the study of nongenic elements in their definitions Functional genomics may also include studies of natural genetic variation over time (such as an organism's development) or space (such as its body regions), as well as functional disruptions such as mutations
. functional genomics is to generate and synthesize genomic and proteomic knowledge into an understanding of the dynamic properties of an organism. This could potentially provide a more complete picture of how the genome specifies function compared to studies of single genes. Integration of functional genomics data is often a part of systems biology approaches.

Applications of functional genomics:
1. Genetic interaction mapping
2. Protein interaction
3. Microarray
4. Serial analysis of gene expression (SAGE)
5. Massively parallel reporter assay(MPRA)

Functional genomics includes function-related aspects of the genome itself such as mutation and polymorphism (such as single nucleotide polymorphisms (SNP) analysis), as well as the measurement of molecular activities. The latter comprise a number of "-omics" such as transcriptomic (gene expression), proteomics (protein production), and metabolomics. Functional genomics uses mostly multiple techniques to measure the abundance of many or all gene products such as mRNAs or proteins within a biological sample. A more focused functional genomics approach might test the function of all variants of one gene and quantify the effects of mutants by using sequencing as a readout of activity. Together these measurement modalities endeavor to quantitate the various biological processes and improve our understanding of gene and protein functions and interactions. There are three or four viral

proteins in the coronavirus membrane. The most abundant structural protein is the membrane (M) glycoprotein; it spans the membrane bilayer three times, leaving a short NH2-terminal domain outside the virus and a long COOH terminus (cytoplasmic domain) inside the virion.4 The spike protein (S) as a type I membrane glycoprotein constitutes the palmers. In fact, the main inducer of neutralizing antibodies is S protein. Between the envelope proteins with exist a molecular interaction that probably determines the formation and composition of the corona viral membrane. M plays a predominant role in the intracellular formation of virus particles without requiring S. In the presence of tunicamycin coronavirus grows and produces spikeless, non-infectious virions that contain M but devoid of S. There are several specific functional genomics approaches depending on what we are focused on (Figure 2):

- DNA level (genomics and epigenomics)
- RNA level (transcriptomic)
- Protein level (proteomics)
- Metabolite level (metabolomics

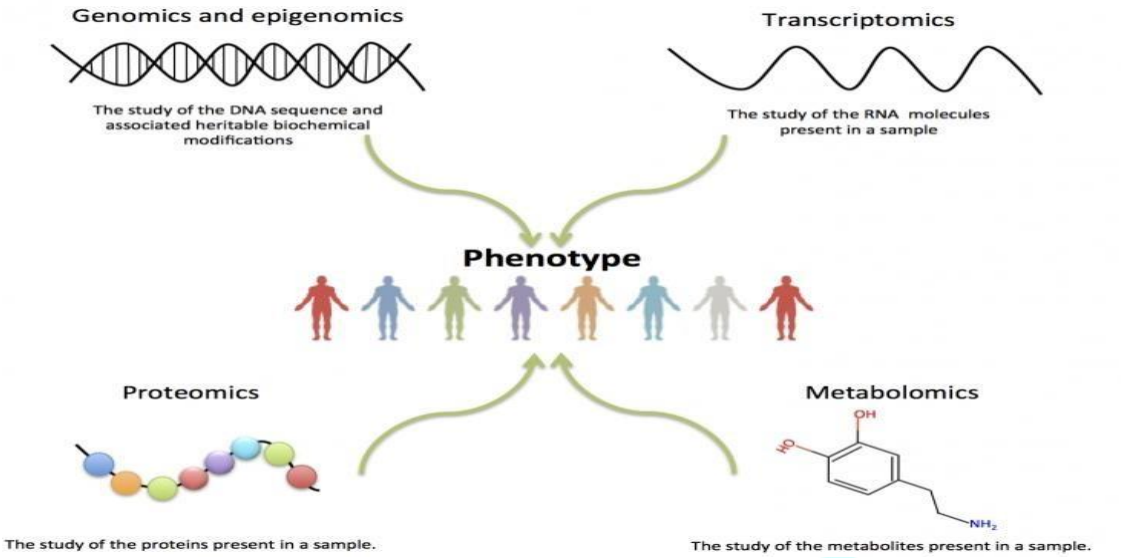

Fig-3

Figure..3 Functional genomics is the study of how the genome, transcripts (genes), proteins and metabolites work together to produce a particular phenotype.

Together, transcriptomic, proteomics and metabolomics describe the transcripts, proteins and metabolites of a biological system, and the integration of these data is expected to provide a complete model of the biological system under study.

Similar to approaches used for the transcriptome, ways to systematically and globally identify the proteome (that is, all proteins that a species can produce) are under development. Many biological decision-making processes

require protein modifications and changes in protein–protein interactions, understanding the proteome (and the transcriptome for that matter) is just as important as understanding genome.

SARS COV 2-

Coronaviruses (Cov) are a group of enveloped viruses, having a positive single-stranded RNA genome and pathogenic. COVID-19 is caused by the SARS-CoV-2 is a more pathogenic form in comparison to previously identified SARS-CoV (2002) and Middle East respiratory syndrome coronavirus (MERS-CoV, 2013). There is an urgent need to study the virus more holistically to understand the mechanism of pathogenesis, its virulence, and to develop effective therapeutic strategies

A novel severe acute respiratory syndrome-related coronavirus-2 (SARS-CoV-2) causinCOVID- 19 pandemic in humans, recently emerged and has exported in more than 200 countries as a result of rapid spread. In this study, we have made an attempt to investigate the SARS-CoV-2 genome reported from 13 different countries, identification of mutations in major coronavirus proteins of these different SARS-CoV-2 genomes and compared with SARS-Cov. These thirteen complete genome sequences of SARS-CoV-2 showed high identity (>99%) to each other, while they shared

82% identity with SARS-Cov. Here, we performed a very systematic mutational analysis of SARS-CoV-2 genomes from different geographical locations, which enabled us to identify numerous unique features of this viral genome.

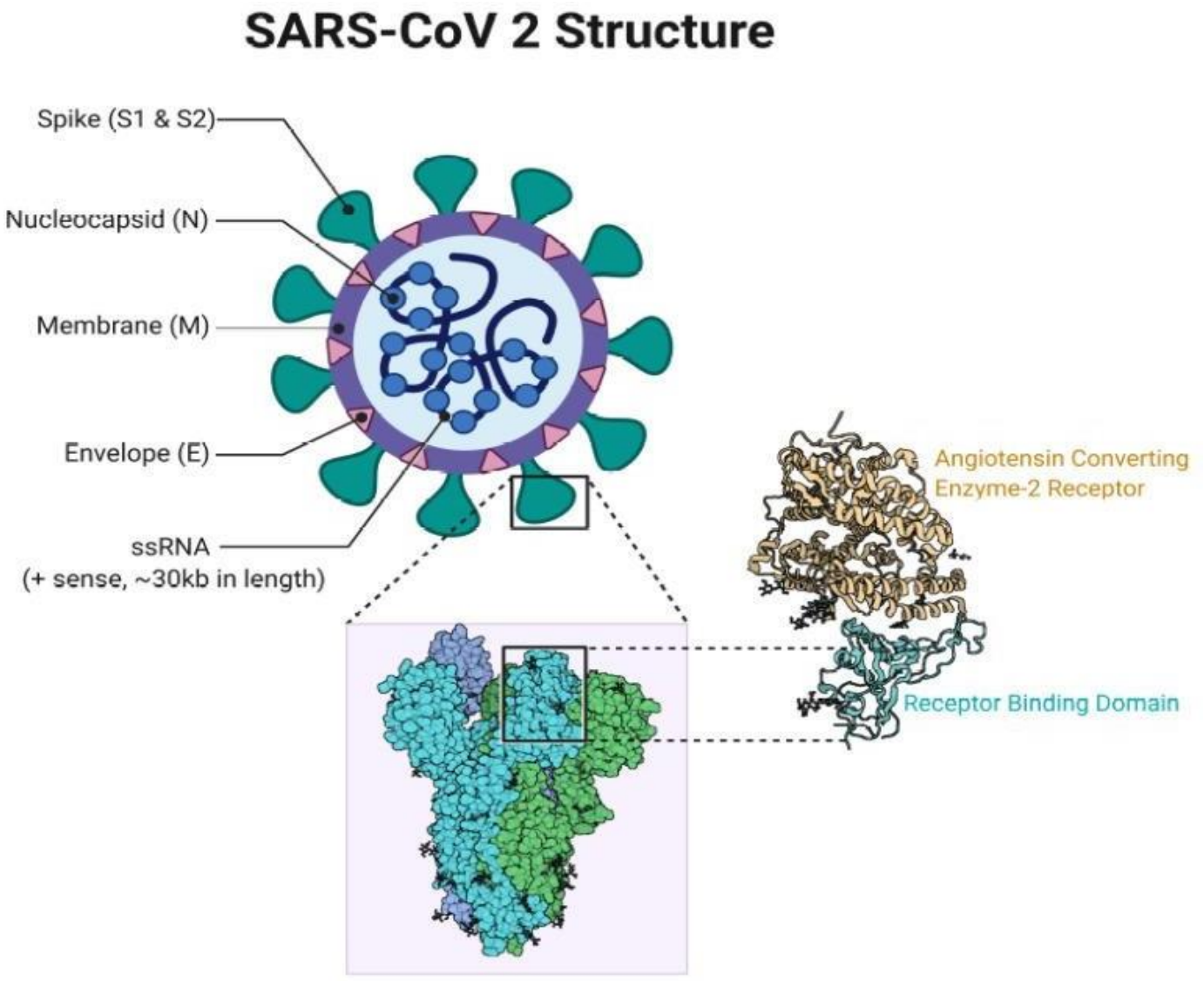

Fig-4

SARS-CoV-2 is a Baltimore class IV] positive-sense single-stranded RNA virus that is contagious in humans. s described by the U.S. National Institutes of Health, it is the successor to SARS-CoV-1, the strain that caused the 2002–2004 SARS outbreak.

Taxonomically, SARS-CoV-2 is a strain of severe acute respiratory syndrome-related coronavirus (SARSr- CoV). It is believed to have zoonotic origins and has close genetic similarity to bat coronaviruses, suggesting it emerged from a bat-borne virus. There is no evidence yet to link an intermediate host, such as a pangolin, to its introduction to humans. The virus shows little genetic diversity, indicating

the spill over event introducing SARS-CoV-2 to humans is likely to have occurred in late 2019 In September 2020,

Mechanism of cov entry and infect host

CoV rely on their spike (S) proteins for binding to the host cell-surface receptor during host cell entry . The spike protein bind the host to reception through the receptor-binding domain (RBD) in the S1 subunit, followed by the fusion of the S2 subunit to the cell membrane. Different cell surface receptors recognize RBD of S proteins of SARS-CoV and MERS-CoV. MERS-CoV recognizes the dipeptidyl peptidase 4 receptor. Whereas, SARS-CoV and SARS-CoV-2 –

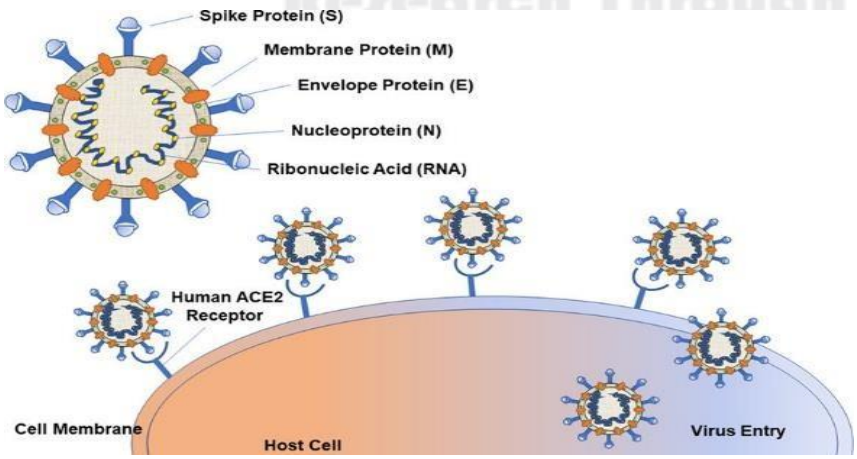

Fig-5

recognize the ACE2 receptor to bind with the viral S protein.
These CoV mainly differ in their mechanism of host entry, suggesting possible changes in the residual composition of S protein that may dictate host entry.
After the virus enters the host cell and uncoats, the genome is transcribed and then translated. Coronavirus genome replication and transcription takes place at cytoplasmic membranes and involve coordinated processes of both continuous and discontinuous RNA synthesis that are mediated by the viral replicate, a huge protein complex encoded by the 20-kb replicase gene.12 The replicase complex is believed to be comprised of up to 16 viral subunits and a number of cellular proteins. Besides RNA-dependent RNA polymerase, RNA helicase, and protease activities, which are common to RNA viruses, the coronavirus replicase was recently predicted to employ a variety of RNA processing enzymes that are not (or extremely rarely)

found in other RNA viruses and include putative sequence-specific endoribonuclease, 3′ -to-5′ exoribonuclease, 2′ -O-ribose methyltransferase, ADP ribose 1′ -phosphatase and, in a subset of group 2 coronaviruses, cyclic phosphodiesterase activities.13 , 14 The proteins are assembled at the cell membrane and genomic RNA is incorporated as the mature particle forms by budding from the internal cell membranes.15

Epidemiology

Epidemiological studies estimate each infection results in 5.7 new ones when no members of the community are immune and no preventive measures taken.[26] The virus primarily spreads between people through close contact and via respiratory droplets produced from coughs or sneezes It mainly enters human cells by binding to the angiotensin converting enzyme 2 (ACE2).The basic reproduction number) of the virus has been estimated to be around 5.7. This means each infection from the virus is expected to result in 5.7 new infections when no members of the community are immune and no preventive measures are taken. The reproduction number may be higher in densely populated conditions such as those found on cruise ships.[141] Many forms of preventive efforts may be employed in specific circumstances to reduce the propagation of the virus
As of 17 December 2020, there have been 74,210,350 total confirmed cases of SARS-CoV-2 infection in the ongoing pandemic.[137] The total number of deaths attributed to the virus is 1,648,956.[137] Many recoveries from confirmed infections go unreported, but at least 41,977,327 people have recovered from confirmed infections.

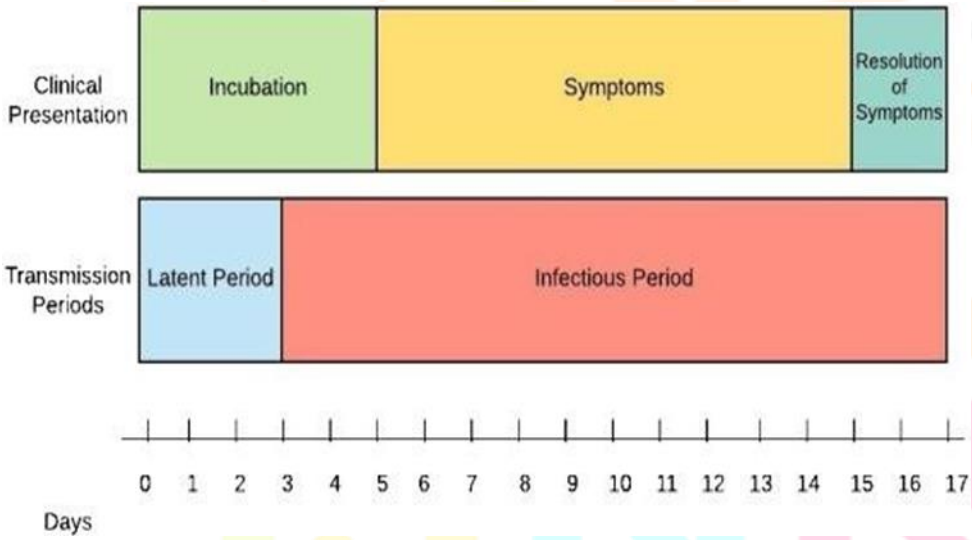

Fig-6
This includes several important country-specific unique mutations in the major proteins of SARS- CoV-2 namely, replicase polyprotein, spike glycoprotein, envelope protein and nucleocapsid protein. Indian strain showed mutation in spike glycoprotein at R408I and in replicase polyprotein at I671T, P2144S and A2798V,. While the spike protein of Spain &south Korea carried F797C and S221W mutation, respectively.

Likewise, several important country specific mutations were analysed. The effect of mutations of these major proteins was also investigated using silico ap.
Main protease (Mpro), the therapeutic target protein of SARS with maximum reported inhibitors, was thoroughly investigated and the effect of mutation on the binding affinity and structural dynamics of Mpro was studied. It was found that the R60C mutation in Mpro affects the protein thereby, affecting the binding of inhibitor within its active site. The implications of mutation on structural characteristics were determined.
SARS-CoV-2 belongs to the family Coronaviridae of genus Beta coronavirus, having positive sense strand RNA genome of 26–32 kb size. SARS-CoV-2 genome has six major open reading frames (ORFs) viz. replication enzyme coding region (ORF 1a and 1b), E gene (envelope protein), M gene (membrane protein), S gene (spike protein), and N gene (nucleocapsid protein)that are common to coronaviruses and a number of other accessory genes (ORF 3a, 6, 7a, 7band 8) (]. The structural proteins: envelope protein, nucleocapsid protein, spike protein and membrane protein are essential for producing the structurally complete viral particle glycoprotein.
Entry of coronavirus into host cells is guided by spike glycoprotein. ORF 1a and 1b encode replication enzyme consisting 16 non-structural proteins (nsp1-16) that are highly conserved among the coronavirus. Main protease (Mpro, also known as 3CLpro) is one of the important nsp encoded by ORF 1a and 1b, play an essential role in the processing of polyproteins and control the replication

of coronavirus [9, 10]. RNA-dependent RNA polymerase(RdRp) also known as nsp12, another important replicase catalyze the replication of RNA using viral genomic RNA template [11].

Mutation rate is very high in RNA viruses, up to a million times higher than their host, which enhance their virulence and evolvability (formation of new species) Coronavirus replication is error prone as compared to other RNA viruses and the estimated mutation rate is 4x10-4 nucleotide substitutions/site/year [20]. The rate of SARS-CoV-2 mediated disease spread and the mortality varies from country to country. Of several reasons affecting the rate of disease spread and mortality, mutations within the SARS-CoV-2 strains is also considered one of the major factors. The genome analysis of the SARS-CoV-2 strains from 13 different countries showed a large number of mutations within the major structural proteins. This study provides a deeper insight into the emergence of these mutations within the major structural as well as nsp encoded by the SARS-CoV-2 genome from different countries. Here, molecular dynamics and other in silico studies were also performed to investigate the effect of mutations on the dynamics of Mpro. The findings of this study provide a clue for the futuristic development of potential vaccine candidate or therapeutic design against covid .

Natural reservoir of SARS cov 2: The most likely ecological reservoir for SARS
cov 2 are bats. The virus jumped the species barrier to humans from another intermediate animal host.This intermediate animal host could be a domestic food animal, a wild animal or a domesticated wild animal.

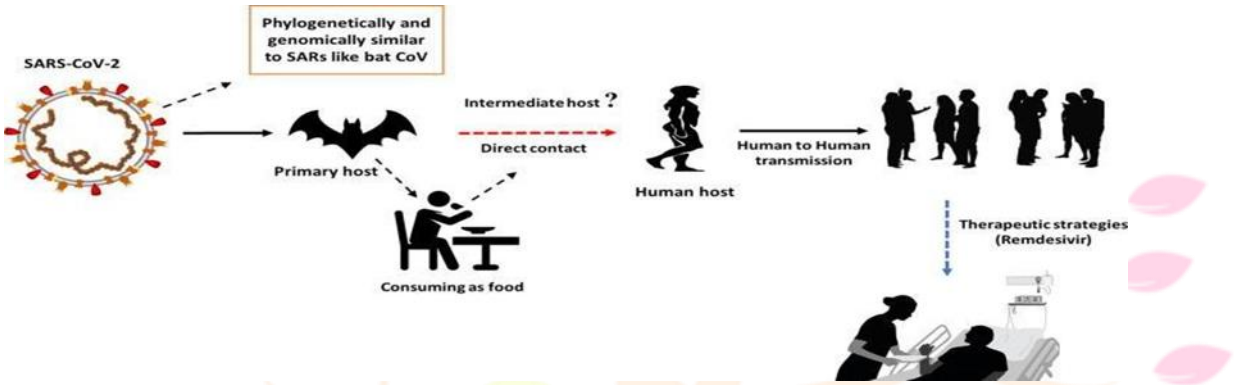

Fig-7 Factors affecting virus pathogenecity-
Co-morbidities are cardiovascular and cerebrovascular disease as well as diabetes.
Several abnormalities also have been observed including cellular immune deficiency, coagulation activation, myocardia injury, hepatic and kidney injury, and secondary bacterial infection. Accessibility of virus to tissue, cell susceptibility to virus multiplication, and virus susceptibility to host defences. Natural selection favours the dominance of low- virulence virus strains.
Sequence analysis and mutation detection.
All major proteins of 13 SARS-CoV-2 sequences and compared with SARS-CoV. ORF 1a and 1b showed 11 changes among all 13 SARS-CoV-2.
Indian SARS-CoV-2 sequence showed three changes at 671 (Isoleucine to Threonine), 2144(Proline to Serine) and 2798 (Alanine to Valine) compared to all other 12 isolates. Here, we also noted two amino acid mutations (in ORF1ab) in each SARS-CoV-2 sequences isolated from China (2708: Asparagine to Serine; 2908: Phenylalanine to Isoleucine).

South Korea(902:Methionine to Isoleucine; 6891: Threonine to Methionine) and Sweden (818: Glycine to Serine; 4321: Phenylalanine to Leucine). Brazil and Vietnam isolate showed only one

Table 1. Amino acid variation in replicase polyprotein of SARS-CoV-2 strains of 13 different countries.

Amino acid	India	China	South Korea	Sweden	Vietnam	Brazil	Taiwan	USA	Australia	Japan	Finland	Nepal	Italy
671	T	I	I	I	I	I	I	I	I	I	I	I	I
818	G	G	G	S	G	G	G	G	G	G	G	G	G
902	M	M	I	M	M	M	M	M	M	M	M	M	M
2144	S	P	P	P	P	P	P	P	P	P	P	P	P
2708	N	S	N	N	N	N	N	N	N	N	N	N	N
2908	F	I	F	F	F	F	F	F	F	F	F	F	F
3323	R	R	R	R	C	R	R	R	R	R	R	R	R
3606	L	L	L	L	L	F	L	L	L	L	L	L	X
4321	F	F	F	L	F	F	F	F	F	F	F	F	F
4798	V	A	A	A	A	A	A	A	A	A	A	A	A
6891	T	T	M	T	T	T	T	T	T	T	T	T	T

Fig-8

change at3603 (Leucine to Phenylalanine) and 3323 (Arginine to Cystine), respectively. The viral Mpro controls the replication of coronavirus and is a key protein responsible for its life cycle [29–31]. Mpro is an attractive drug discovery target. The analysis of Mpro reveals that there was only one point mutation (R60C) in the Vietnam strain of SARS-CoV-2 RdRp, which is another important target for antiviral drugs functions by catalyzing the viral RNA synthesis [32]. Only one mutation (A406V) was observed in the RdRp of Indian

SARS-CoV-2 isolate spike proteins are the key surface glycoproteins and are well reported for their prominent role in interaction with host cell receptors . Here, we analysed the mutations in the covid.

- spike protein of SARS-CoV-2 from different countries. It was found that this glycoprotein carried five different amino acid mutations at various positions within the investigated SARS-CoV-2 isolates. For instance, India, Finland, Australia, South Korea and SwedenSARS-CoV-2 isolates showed one amino acid change at 408 (Arginine to Isoleucine).

B49 (Histidine to Tyrosine), 247 (Serine to Arginine), 221 (Serine to Tryptophan) and 797 (Phenylalanine to Cysteine), respectively .The value of ΔΔG show that the mutant R408I (0.49732107 kcal/mol) mutation was having stabilization effect on spike protein. It was found that the mutation on the receptor binding domain (RBD) of spike protein increases the stability.

When these 13 SARS-CoV-2 isolates were compared to SARS-CoV sequence, 1338 changes have been reported (S1 File). The analysis of ORF3a showed 3 mutations within different

- SARS-CoV-2 strains: W128L (South Korea), L140V (Japan), G251V (Australia, South Korea, Brazil, Italy, Sweden).

One amino acid change occurred in each envelope protein of South Korea SARS-CoV-2 isolate at 37 (Leucine to Histidine) and nucleocapsid protein of Japan SARS-CoV-2 isolate at 344 (Proline to Serine) when compared among 13 SARS-CoV-2 isolates while, 5 and 45 changes has been reported in envelop and nucleocapsid proteins, respectively as compared to SARS-CoV. Deletion of Glycine and Serine occurred at position 70.

8 in envelop and nucleocapsid proteins, respectively, in all 13 SARS-CoV-2 isolates when compared to SARS- CoV. MEM glycoprotein did not show any amino ac change among 13 SARS-CoV-2 isolates, while 24 changes occurred when compared to SARS-CoV (S1File). All the other point mutations occurring within the structural proteins of SARS-CoV-2 isolates from different countries were found to decrease protein stability (Table 6).

Steps involved in genomic analysis:

- Genome sequence assembly
- Identify repetitive sequence: mask out
- Gene prediction-train a model for each genome
- Gene annotation: process of attaching biological information
- Metabolic pathway and regulation
- Protein 2D gel electrophoresis
- Functional genomics
- Gene location
- Self comparison of proteome
- Comparative genomics
- Identify cluster of structurally related genes
- Evolutionary modelling: to analyze chromosomal arrangement.

Methods of data collection

Inter and intra specific pan genome analysis

In order to highlight the proteins shared between SARS-CoV-2 and other species of the genus Beta coronavirus, Likewise, the proteins shared on the intra-genomic scale of SARS-CoV-2, we have constructed a pan-genome by clustering the sets of proteins encoded in 115 genomes distributed in 17 species, including 83 genomes belonging to SARS-Cov-2 .. A total of 1,148 proteins were grouped into a pan genome of 94 orthologous protein clusters. Among them, ten protein clusters were shared between SARS-CoV-2 and only three species of the genus Betacoronavirus, including; BatCoV RaTG13, SARS-CoV and Bat Hp-betacoronavirus / Zhejiang2013. The BatCoV RaTG13 genome had more orthologous proteins shared with SARS-CoV-2, followed by SARS-CoV with ten and nine orthologous proteins, respectively . It is interesting to note that among all the strains used of Betacoronavirus, the protein ORF8 was found in orthology only between SARS-RATG13 and SARS-CoV-2. In addition, the ORF10 protein was found as a singleton for SARS-CoV-2 genomes.

These proteomes were used for the construction of pangenome at the inter-specific scale of Betacoronavirus and intra-genomic of SARS-CoV-2. The strategy of best reciprocal BLAST result was implemented to identify all of the orthologous genes using Proteinortho v6.0b [20]. Proteins with an identity above 60% and sequence coverage above 75% with an e-value threshold below 1e-5 were used to be considered as significant hits.

Selective pressure analysis

Selection pressure means factors that contribute to selection which variations will provide the individual with an increase chance of surviving over others. Because of selective pressures, organisms with certain phenotypes have an advantage when it comes to survival and reproduction. Over time, this leads to evolution..Selective pressure on orf1ab, gene harboured a high rate of mutations and on the Spike gene, indicated a single alignment-wide ω ratio of 0.571391 and 0.75951 for spike and or1ab, respectively. Most sites for both genes had ω <1 values, indicating purifying selection. In orf1ab, we estimated eight sites under negative selection pressure (696, 1171, 2923, 3003, 3715, 5221, 5704 and 6267) and three sites under positive selection pressure (1473, 2244 and 3090). For spike, we found seven sites under negative selection pressure (215, 474, 541, 809, 820, 921 and 1044), and only one site under negative selection pressure. None of the hotspot mutations was identified under negative selection, this is mostly due sampling size and early date of sample collection.

Types of selective pressure include:

- Resource availability – Presence of sufficient food, habitat (shelter / territory) and mates.

- Environmental conditions – Temperature, weather conditions or geographical access.
- Biological factors – Predators and pathogens (diseases)

Phylogenetic genomic analysis of sars cov 2

The phylogenetic tree based on the whole genome alignment demonstrates that SARS-CoV-2 is wildly disseminated across distinct geographical location. The results showed that several strains are closely related even though they belong to different countries. Which indicate likely transfer events and identify routes for geographical dissemination.

. The phylogenetic analysis revealed two main clades C1 and C2; the original clade C1 harbouring the mutation F3606L and starting since the beginning of the pandemic contains mainly Chinese strains from Dec to mid-Feb. After this period, the clade has emerged in other countries all over the globe. C1 is also composed of two subclades, SCB 1 sharing the mutation G251V (ORF3a) first identified in strains from china and further emerged in European strains, such as England and Iceland. The second subclade SCB2 also stared in China at the beginning of Jan and harbored the mutation L84S (ORF8). Following the first appearance it started emerging in other European countries mainly in Spain, this clade has also emerged in the USA in mid-Jan and gives birth to a new cluster containing 444 strains all sharing a C17747T mutation (Leu5828Leu, ORF1ab) starting from mid-Feb. Strains from the second clad C2 shared the spike mutation D614G (S) and harbored three subclades, this clade started in shanghai end of Jan. However, it contains mainly strain from Europe and North America. The first sub cluster SCB3 harboured strain sharing two mutation R203K (N) and G204R (N) harbouring largely strains from Europe and some strains from North Africa (France and USA). The second sub cluster SCB4 harboured strain from Europe with the Q57H (ORF3a) mutation,

Distribution of the 3,067 genomes used in this study by country and date of isolation.

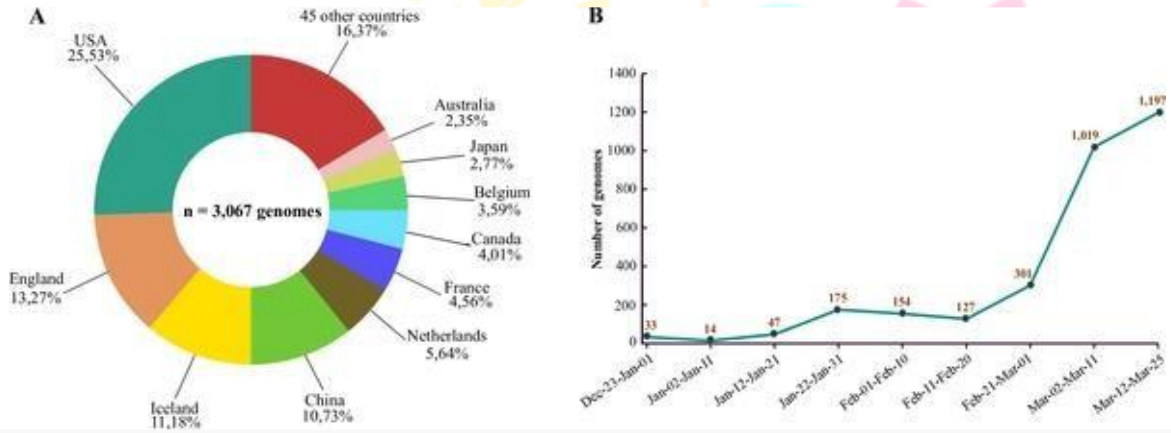

Fig-9

A) The pie chart represents the percentage of genomes used in this study according to their geographic origins. The colours indicate different countries.

B) Number of genomes of complete pathogens, distributed over a period of 3 months from the end of December to the end of March.

Amongst all known RNA viruses, coronaviruses consist of the largest genome (26.4 to 31.7kb)

The large genome size provides more plasticity in accommodating and modifying genes [36–38]. Mutation frequency is very high in RNA viruses, which enhances virulence and responsible for the formation of new species [19]. The high frequency of mutation within the viral genome at different geographical locations may be one of the reasons that SARS-CoV-2 is responsible for change in mortality rate and symptom of the disease.

The RBD of spike protein is the region which specifically interact with ACE2 leading to viral entry into the host cell.

The Indian isolate of SARS-CoV-2 showed mutation within this region where at 408 position, Arginine is replaced by Isoleucine.]. Earlier findings suggest that a single point mutation at RBD is responsible for disrupting the antigenic structure, thereby, affecting the binding of RBD to ACE2.

- Single amino acid mutation was observed in both Mpro (R60C) of SARS-CoV-2 Vietnam isolate and RdRp (A408V) of SARS-CoV-2 India isolate. The in silico findings revealed that the mutations in both strains decrease the stability of protein.
- total of 782 variant sites were detected compared to the Wuhan-Hu-1/2019 reference sequence, of which 65.98% having a non-synonymous effect, 28.39% synonymous mutations, and 5.63% are distributed regions intergenic.
- Mutational frequency analysis revealed the presence of 68 mutations with a frequency greater than 0.06% of the total genomes, which corresponds to at least 20/3067 genomes.

It was also revealed that among all the genomes studied in this study, the Indian SARSCoV-2 isolates were carrying maximum mutation. The Indian isolates were carrying the R408I on the spike protein while A406V on the RdRp and several mutations on the replicase polyprotein of SARS-CoV-2. It is expected that these large number of mutations among the SARSCoV-2 may affect the vaccine/inhibitor development against these isolates.

Newest strain of covid the UK health official announced that the country has identified a new variant of Coronavirus that spreads faster than the other strains. Up More than 40 countries including India have banned travel to and from the UK, a move several

researchers said was necessary in the view of the rapid spread of the new strain, which was first detected on September 21.
It was previously predicted that 10 subgenomic RNAs make up the viral particle structure. However, the research team confirmed that 9 subgenomic RNAs actually exist, invalidating the remaining one subgenomic RNA. Researchers also found that there are dozens of unknown subgenomic RNAs, owing to RNA fusion and deletion events. "Though it requires further investigation, these molecular events may lead to the relatively rapid evolution of coronavirus. It is unclear yet what these novel RNAs do, but a possibility is that they may assist the virus to avoid the attack from the host," says Prof. Kim.

They believe if they figure out the unknown characteristics of RNA, the findings may offer a new clue for combatting the new coronavirus. Newly discovered features will also help to understand the life cycle of the virus and develop new strategies for antiviral therapy.

Methods:

1. sequencial retrieval and allignment: To decipher the genetic variations, we retrieved two thousand four hundred and ninety-two (n=2,492) complete or near-complete genomes of SARS-CoV-2 available at the global initiative on sharing all infuenza data (GISAID) up to 30 March 2020. We divided the SARS-CoV-2 genome sequence data according to their geographic origins from six continents such as Europe, Asia, North America, South America, Africa, and Australia, and five related climatic zones including temperate, tropical, diverse, dry and continental. To estimate the case fatality (mortality) rates of SARS-CoV-2 infections, we collected information on total infected cases, and total reported deaths in these countries from the World Health Organization (WHO) COVID-19 Reports 35 up to March 30, 2020. Tese SARS-CoV-2 sequences belonged to the infected patients from 58 countries of six continents.
The spike protein is a key protein for SARS-CoV-2 viral entry and a target for vaccine development. We, therefore, wanted to find amino acid conservation between other coronavirus sequences in the spike protein. We used the basic local alignment search tool BLAST (National Centre for Biotechnology Information [NCBI], Bethesda, United States)23 followed by the constraint-based multiple alignment tool COBALT (NCBI, Bethesda, United States).24 We carefully investigated

mutations within the receptor binding domain and predicted B-cell epitopes.25,26 The mutations were further analysed to identify cross species conservation and to understand the nature of amino acid changes. We visualized the aligned sequence using the open source software ale.

2. Mutation analysis: For this study, we used the sequence of established SARS-CoV-2 reference genome, NC_045512.17 This genome was sequenced in December 2019. Each sample was first aligned to the reference genome in a pairwise manner using EMBOSS needle (Hinxton, Cambridge, England), with a default gap penalty of 10 and extension penalty of 0.5.1
Then, we developed a custom script in Python (Python Software Foundation, Wilmington, United States of America) to extract the differences between the genome variants and the reference genome. Nucleotide variants in the coding regions were converted to corresponding encoded amino acid residues. For the open reading frame 1 (ORF1), we used the protein coordinates from YP_009724389.119 for translation.
Finally, we carefully investigated stop-gained and frameshift variants causing deletions and insertions to detect potential artefacts caused by undetermined or ambiguous bases. The results are provided in a list of variants (available in the data repository.
Using the identified recurrent variants, we performed hierarchical clustering in SciPy library, Python, to identify clades. First, a binary matrix of samples and distinct variants was created. Then, we did hierarchical clustering using the Ward's method
We used MEGA 7for multiple sequence alignments to differentiate the SARS-CoV-2 genomes according to their open reading frames (ORF). Sequence cleaner was used to remove all ambiguous and low-quality sequences. SeqKit38 toolkit was used to intercept gap containing strains for deletion analysis. Internal stop codon containing sequences were removed by using Sequence Dataset builder (SEDA; . Amino-acid heterogeneity analysis was performed with Fingerprint, a web-based protein profile analysis tool39. Amino-acid mutation analysis was done by simple bio-python program with pairwise alignment . We used custom Venn diagrams server to create the Venn diagrams.
Technique used in mutation detection:
1. Allele specific oligomer hybridisation
2.Array technology
3.High density oligonucleotide array
4.Melt curve analysis

Conclusion:
Tis study reveals a number of unreported mutations, which cover both mismatches and deletions in translated and untranslated regions of the SARS-CoV-2 genomes. Moreover, the geo-climate distribution of the mutations deciphered higher unique mutations as well as disease severity in the European temperate countries. Further investigations should focus on structural validations and subsequent phenotypic consequences of the deletions and/or mismatches in transmission dynamics of the current epidemics and the immediate implications of these genomic markers to develop potential prophylaxis and mitigation for tackling the crisis of pandemic COVID-19.

Moreover, the identification of the conformational changes in mutated protein structures and untranslated cascading elements is of signifcance for studying the virulence, pathogenicity and transmissibility of SARS-CoV-2.
In total, we analysed 10 022 SARS CoV-2 genomes (sequences are available from the data repository)20 from 68 countries. Most genomes came from the United States of America (3543 samples), followed by the United Kingdom of Great Britain and Northern Ireland (1987 samples) and Australia (760 samples; Box 1). We detected in total 65776 variants with 5775 distinct variants. The 5775 distinct variants consist of 2969 missense mutations, 1965 synonymously mutations, 484 mutations in the non-coding regions, 142 non-coding deletions, 100 in-frame deletions, 66 non-coding insertions, 36 stop- gained variants, 11 frameshift deletions and two in-frame insertions The number of people with confirmed COVID-19 has rapidly increased over the last five months with no

sign of decline in the near future. The fight against COVID-19 will be long, until vaccines and other effective therapies are developed. To facilitate rapid therapeutic development, clinic pathological, genomic and other societal information must be shared with researchers, physicians and public health officials. Given the evolving nature of the SARS-CoV-2 genome, drug and vaccine developers should continue to be vigilant for emergence of new variants or sub-strains of the virus

References:

1. Drosten C, Gunther S, Preiser W, van der Werf S, Brodt HR, Becker S, et al. Identification of a novel corona virus in patients with severe acute respiratory syndrome. N ENGmedical;348(20):1967.
2. Zaki AM, van Boheemen S, Bestebroer TM, Osterhaus AD, Fouchier RA. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. N Engl J Med. 2012; 367(19):1814–20.
3. Lindner HA, Fotouhi-Ardakani N, Lytvyn V, Lachance P, Sulea T, Menard R. The papain-like protease from the severe acute respiratory syndrome coronavirus is a deubiquitinating enzyme. J Virol. 2005; 79 (24):15199–208.
4. Coronaviridae Study Group of the International Committee on Taxonomy of V. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2.NAT Microbiol. 2020; 5(4):536–44.
5. Webb B, Sali A. Comparative Protein Structure Modeling Using MODELLER. Curr Protoc Bioinformatics. 2016; 54:5.6.1–5.6.37.
6. Wang M, Yan M, Xu H, Liang W, Kan B, Zheng B, et al. SARS-CoV infection in a restaurant from palm civet. Emerg Infect Dis. 2005; 11(12):1860–5.
7. Ruch TR, Machamer CE. The coronavirus E protein: assembly and beyond. Viruses. 2012; 4(3):363–399
8. Nieto-Torres JL, DeDiego ML, Verdia-Baguena C, Jimenez-Guardeno JM, Regla-Nava JA, FernandezDelgado R, et al. Severe acute respiratory syndrome coronavirus envelope protein ion channel activity promotes virus fitness and pathogenesis. PLoS Pathog.2014;10(5):e1004077.
9. Hsin WC, Chang CH, Chang CY, Peng WH, Chien CL, Chang MF, et al. Nucleocapsid protein-dependent assembly of the RNA packaging signal of Middle East respiratory syndrome coronavirus. J Biomed Sci. 2018; 25(1):47.

10. Wang L, Veenstra DL, Radmer RJ, Kollman PA. Can one predict protein stability? An attempt to do so for residue 133 of T4 lysozyme using a combination of free energy derivatives, PROFEC, and free energy perturbation methods. Proteins. 1998; 32(4):438–58. PMID: 9726415
11. Stoye J, Blomberg J, Coffin J, Fan H, Hahn B, Neil J. International Committee on Taxonomy of Viruses.ICTV 9th Report(2011)
12. Sanner MF. Python: a programming language for software integration and development. J Mol Graph Model. 1999; 17(1):57–61. PMID: 10660911.
13. Cotten, M. et al. Transmission and evolution of the Middle East respiratory syndrome coronavirus in Saudi Arabia: a descriptive genomic study. Lancet 382(9909), 1993–2002 (2013).
14. ZZhu, N. et al. A novel coronavirus from patients with pneumonia in China, 2019. N. Eng. J. Med. 382(8), 727–733.
15. Walls, A. C. et al. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. Cell 180, 1–12
.
16. Ahmed, S. F., Quadeer, A. A. & McKay, M. R. Preliminary identifcation of potential vaccine targets for the COVID-19 coronavirus (SARS-CoV-2) based on SARS-CoV immunological studies. Viruses 12(3), 254 (2020).
17. Phan, T. Genetic diversity and evolution of SARS-CoV-2. Infect. Genet. Evol. 81, 104260 (2020).
18 . Rahman, M. S. et al. Epitope-based chimeric peptide vaccine design against S, M and E proteins of SARS- CoV-2, the etiologic agent of COVID-19 pandemic: an in silico approach. PeerJ 8, e9572 (2020).
19. Chiara, M., Horner, D. S. & Pesole, G. Comparative genomics suggests limited variability and similar evolutionary patterns between major clades of SARS-Cov-2. bioRxiv (2020).
20. Sardar, R., Satish, D., Birla, S. & Gupta, D. Comparative analyses of SAR-CoV2 genomes from diferent geographical locations and other coronavirus family genomes reveals unique features potentially consequential to host-virus interaction and pathogenesis. bioRxiv (2020).
21. Armijos-Jaramillo, V., Yeager, J., Muslin, C. & Perez-Castillo, Y. SARS-CoV-2, an evolutionary perspective of interaction with human ACE2 reveals undiscovered amino acids necessary for complex stability. bioRxiv (2020).
22. 10. Shen, Z. et al. Genomic diversity of SARS-CoV-2 in Coronavirus Disease 2019 patients. Clin. Infect. Dis. Ciaa 203 (2020).
23. Pachetti, M. et al. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. J. Transl. Med. 18, 179 (2020).
24Pybus, O. G., Tatem, A. J. & Lemey, P. Virus evolution and transmission in an ever more connected world. Proceed. Biol. Sci. 282(1821), 20142878 (2015).
25.. Mahy, B. W. J. Te evolution and emergence of RNA viruses. Emerg. Infect. Dis. 16(5), 899 (2010).
26. Yin, C. Genotyping coronavirus SARS-CoV-2: methods and implications. Genomics S0888–7543(20), 30318– 30319 (2020).

27 emergend, J. et al. Te importance of naturally attenuated Sars-Cov-2 in the fght against Covid-19. Environ. Microbiol. 22(6), 1997–2000 (2020).
28. Holland, L. A. et al. An 81 nucleotide deletion in SARS-CoV-2 ORF7a identifed from sentinel surveillance in Arizona (Jan–Mar 2020). J. Virol. (2020).
29. Bal, A. et al. Molecular characterization of SARS-CoV-2 in the frst COVID-19 cluster in France reveals an amino acid deletion in nsp2 (Asp268del). Clin. Microbiol. Infect. 26(7), 960–962 (2020).
30. Kim, Y., Jedrzejczak, R., Maltseva, N. I., Wilamowski, M., Endres, M., Godzik, A. et al. Crystal structure of Nsp15 endoribonuclease NendoU from SARS-CoV-2. Protein Sci. PMC7264519 (2020).
34. Angeletti, S. et al. COVID-2019: the role of the nsp2 and nsp3 in its pathogenesis. J. Med. Virol. 92(6), 584– 588 (2020).
35 Harvey, C. What Could Warming Mean for Pathogens like Coronavirus? E&E News, March 9 (2020).

36. Brassey, J., Heneghan, C., Mahtani, K. R. & Aronson, J. K. Do weather conditions infuence the transmission of the coronavirus (SARS-CoV-2)? (Centre for Evidence-Based Medicine, Nufeld Department of Primary Care Health Sciences, University of Oxford, Oxford, 2020).
37 Su, S. et al. Epidemiology, genetic recombination, and pathogenesis of coronaviruses. Trends Microbiol. 24(6), 490–502 (2020).
38 Yuan, M. et al. A highly conserved cryptic epitope in the receptor-binding domains of SARS-CoV-2 and SARS- CoV. Scienceeabb7269 (2020).
39. Sokalingam, S., Raghunathan, G., Soundrarajan, N. & Lee, S. G. A study on the efect of surface lysine to arginine mutagenesis on protein stability and structure using green fuorescent protein. PLoS ONE 7(7), e40410 (2012).
40. Casadevall, A. & Pirofski, L. A. Te convalescent sera option for containing COVID-19. J. Clin. Invest. 130(4), 1545–1548 (2020).
41. Liu, Z., Zheng, H., Yuan, R., Li, M., Lin, H. & Peng, J. Identifcation of a common deletion in the spike protein of SARS-CoV-2. bioRxiv. (2020).
42. Lau, S. Y. et al. Attenuated SARS-CoV-2 variants with deletions at the S1/S2 junction. Emerg. Microb. Infect. 9(1), 837–842 (2020).
43. Li, W. et al. Bats are natural reservoirs of SARS-like coronaviruses. Science 310, 676–679 (2005).
44. Xu, Y. et al. Characteristics of pediatric SARS-CoV-2 infection and potential evidence for persistent fecal viral shedding. Nat. Med.
45. Liu, S. et al. Inter Protein structure prediction and structural genomics". Science. 294 (5540): 93–6.action between heptad repeat 1 and 2 regions in spike protein of SARS-associated coronavirus: implications for virus fusogenic mechanism and identifcation of fusion inhibitors. Lancet 363(9413), 938–947 (2004).
46.. Deshwal, V. K. COVID 19: a comparative study of Asian, European, American continent. Int. J. Sci. Res. Eng. Dev. 3(2), 436–440 (2020).

47. Khailany, R. A., Safdar, M. & Ozaslan, M. Genomic characterization of a novel SARS-CoV-2. Gene Rep. 19, 100682 (2020).
48. Taiaroa, G., Rawlinson, D., Featherstone, L., Pitt, M., Caly, L., Druce, J. et al. Direct RNA sequencing and early evolution of SARSCoV-2. BioRxiv (2020).
49.. Te World Health Organization (WHO). Novel Coronavirus (2019-nCoV) Situation Reports (WHO, Geneva, 2020).
50. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucl. Acids Res. 30(14), 3059–3066 (2002).
51. Kumar, S., Stecher, G. & Tamura, K. MEGA7: molecular evolutionary genetics analysis version 70 for bigger datasets. Mol. Biol. Evol. 33(7), 1870–1874 (2016).
52. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q fle manipulation.
PLoS ONE 11(10), e0163962 (2016).
53. Goyal, A. et al. Identifcation of an ideal-like fngerprint for a protein fold using overlapped conserved residues based approach. Sci. Rep. 4, 5643 (2015).
54. Okonechnikov, K., Golosova, O. & Fursov, M. Unipro UGENE: a unifed bioinformatics toolkit. Bioinformatics 28(8), 1166–1167
.55.Zeng Y., Zhen Y.. Chinese medical staff request international medical assistancein fighting against COVID-19.
The Lancet Global health 2020.
56. Calefi A.S., de Queiroz Nunes C.A., da Silva Fonseca J.G., Quinteiro-Filho W.M.,Ferreira A.J.P., Palermo-Neto J.. Heat stress reduces Eimeria spp. infection and interferes with C. perfringens infection via activation of the hypothalamic-pituatary adrenal axis. Research in veterinary science 2019;123:273–80.
57 Scorza F.A., Albuquerque R., Arida R.M., et al. What are the similarities between stress, sudden cardiac death in Gallus gallus and sudden unexpected death in people with epilepsy. Arquivos de neuro-psiquiatria 2010;68(5):788–90.
58. Novel Coronavirus Pneumonia Emergency Response Epidemiology T. [The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19) in China]. Zhonghua Liu Xing Bing Xue Za Zhi 2020;41(2):145–51.
59. E M., A U.-E., K R., et al. Sudden cardiac arrest during sports activity in middle age. Circulation 2015;131(16):1384–91.
60. Z W., JM M.. Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72 314 Cases From the Chinese Center for Disease Control and Prevention. Jama 2020.