MIE 1624: INTRODUCTION TO DATA SCIENCE AND ANALYTICS

FINAL EXAM PROJECT

NAME – SHREYAS CHOUDHARY

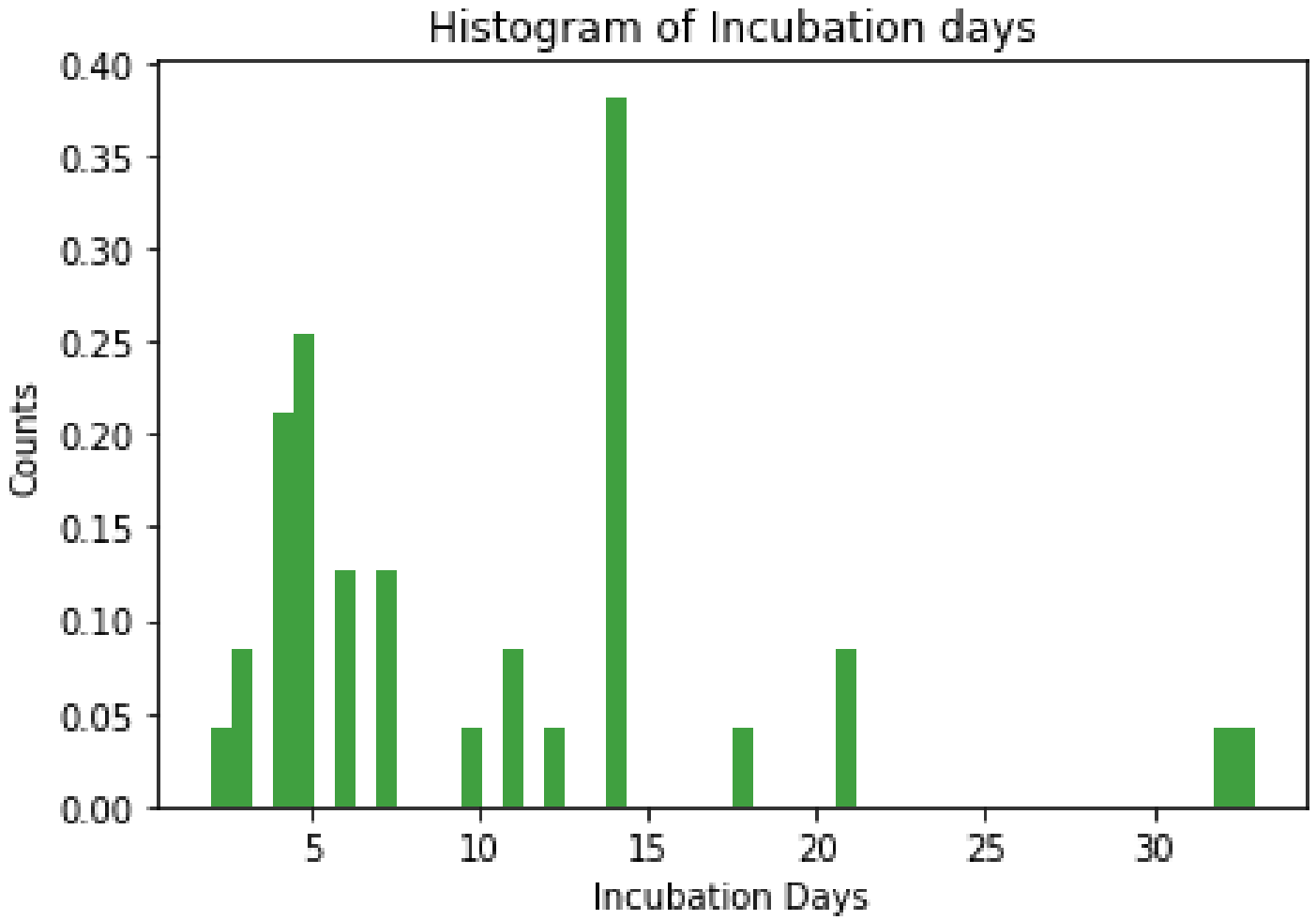
STUDENT NUMBER- 1006376217

DATA CLEANING

- In addition to the project starter code provided, an additional user defined cleaning code was implemented, whose outcome was to come up with a short and precise summary on providing the keywords, the use of which is done in the last part to infer some insights out of the analysis done.

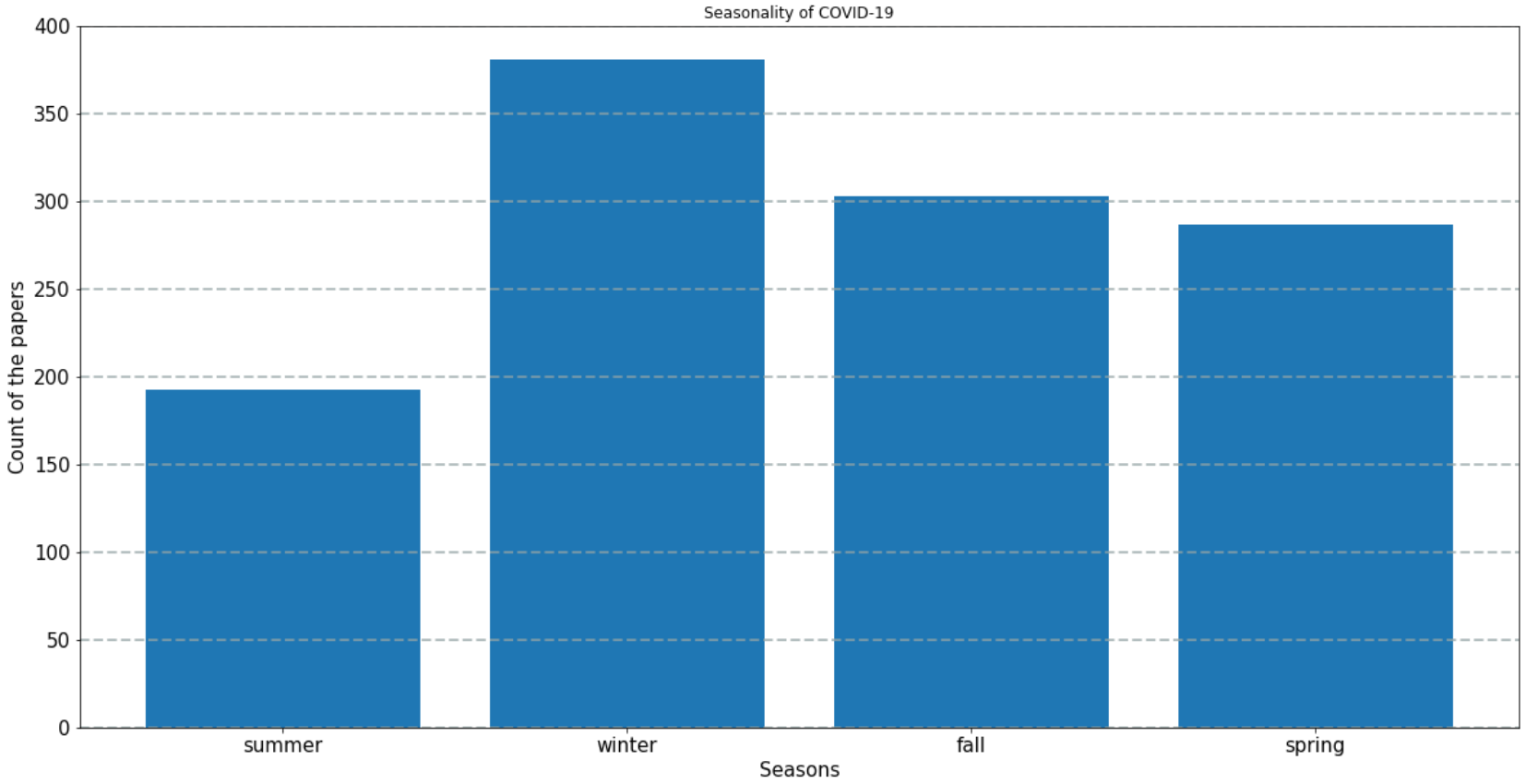
DATA VISUALIZATION AND EXPLORATORY DATA ANALYSIS

- The data visualization done for the incubation period includes filtering out the abstracts description form the original dataset containing the keywords related to incubation and then forming a dictionary of them, defining a list of keywords (in this case , days) and then coming up with the visualization
- This analysis suggests that an incubation period of 14 days is most frequently occurring in the dataset of research papers



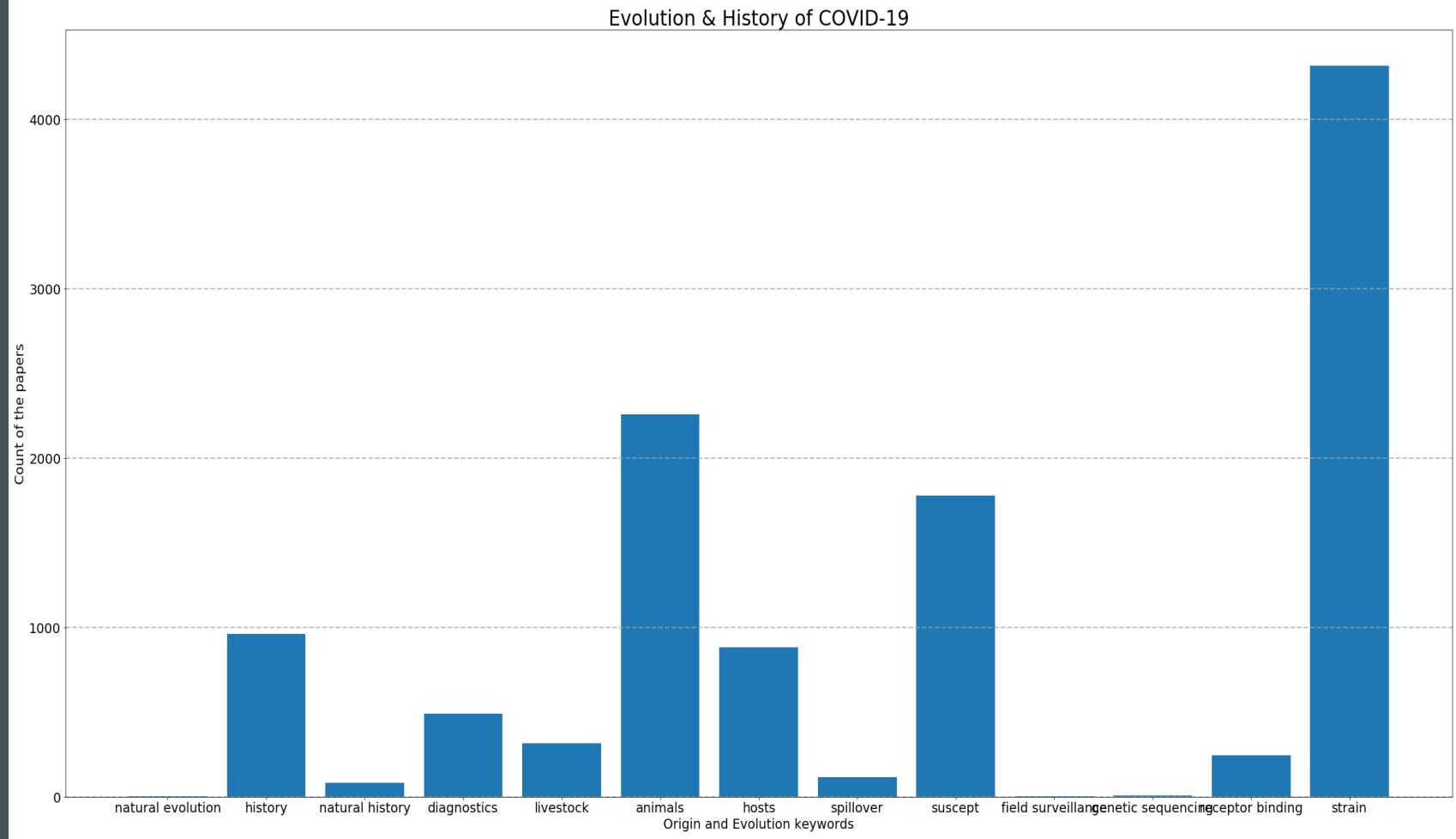
DATA VISUALIZATION AND EXPLORATORY DATA ANALYSIS

- The data visualization done for the seasonality of the virus includes filtering out the abstracts description form the original dataset containing the keywords related to season and then forming a dictionary of them, defining a list of keywords (in this case, season names) and then coming up with the visualization
- This analysis suggests that winter season is most frequently occurring in the dataset of research papers, during which the virus tend to be more proactive.



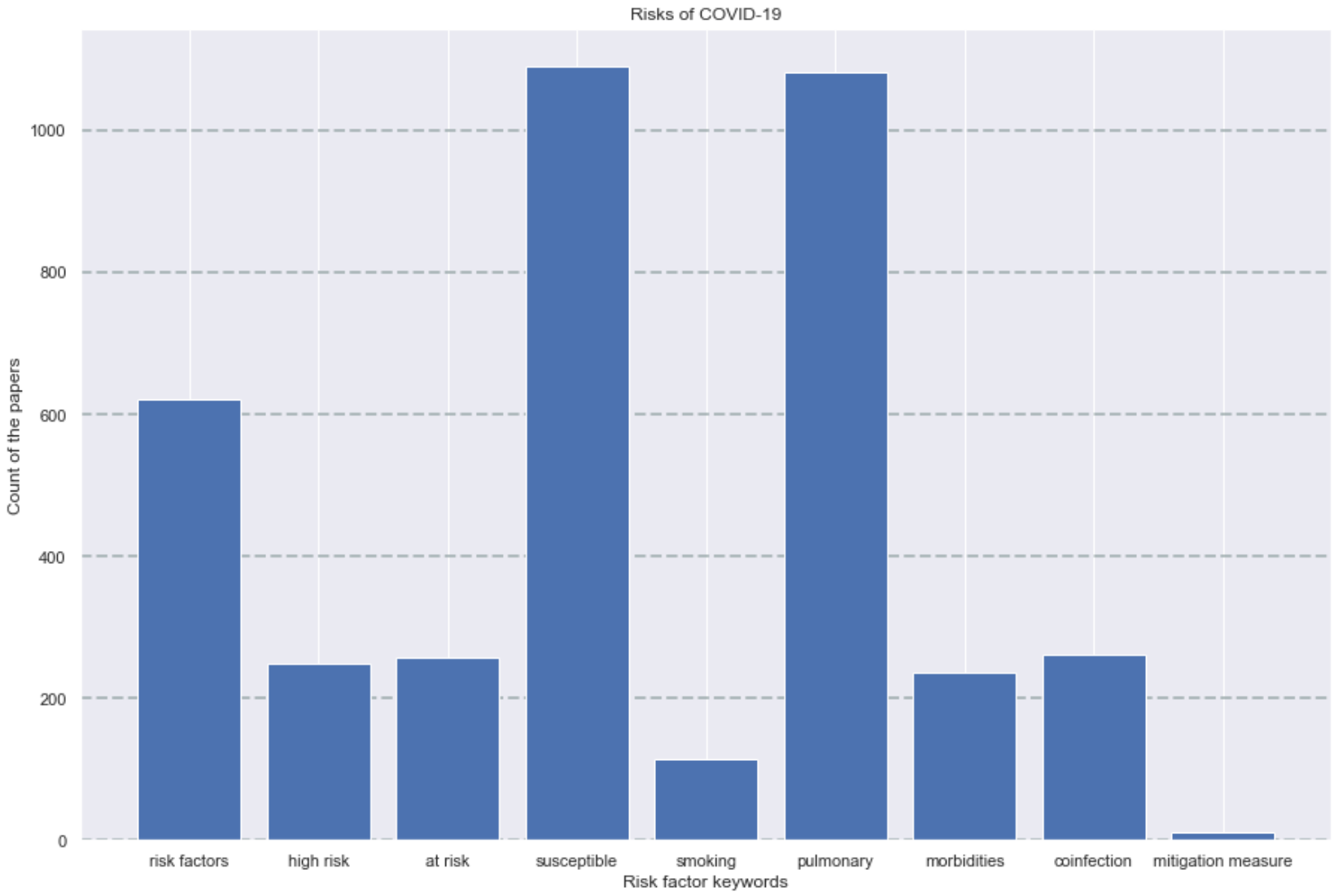
DATA VISUALIZATION AND EXPLORATORY DATA ANALYSIS

- The data visualization done for the origin and evolution of the virus includes filtering out the abstracts description form the original dataset containing the keywords related to origin and evolution and then forming a dictionary of them, defining a list of keywords (in this case, sources of common virus) and then coming up with the visualization.
- This analysis suggests that strains (different species of the virus) is most frequently occurring in the dataset of research papers, followed by the keywords suscept and animals which suggest that the origin and evolution of COVID-19 is usually strains of the virus.



DATA VISUALIZATION AND EXPLORATORY DATA ANALYSIS

- The data visualization done for the risk of the virus includes filtering out the abstracts description form the original dataset containing the keywords related to risks and then forming a dictionary of them, defining a list of keywords (in this case, some risk keywords) and then coming up with the visualization.
- This analysis suggests that risk factor keyword pulmonary (relating to the respiratory system) is most frequently occurring in the dataset of research papers, followed by the keywords morbidities and coinfection which suggest that the risk factors of COVID-19 are mainly related to the respiratory system and can turn out as bad as a coinfection (getting infected with another virus than COVID-19 at the same time)



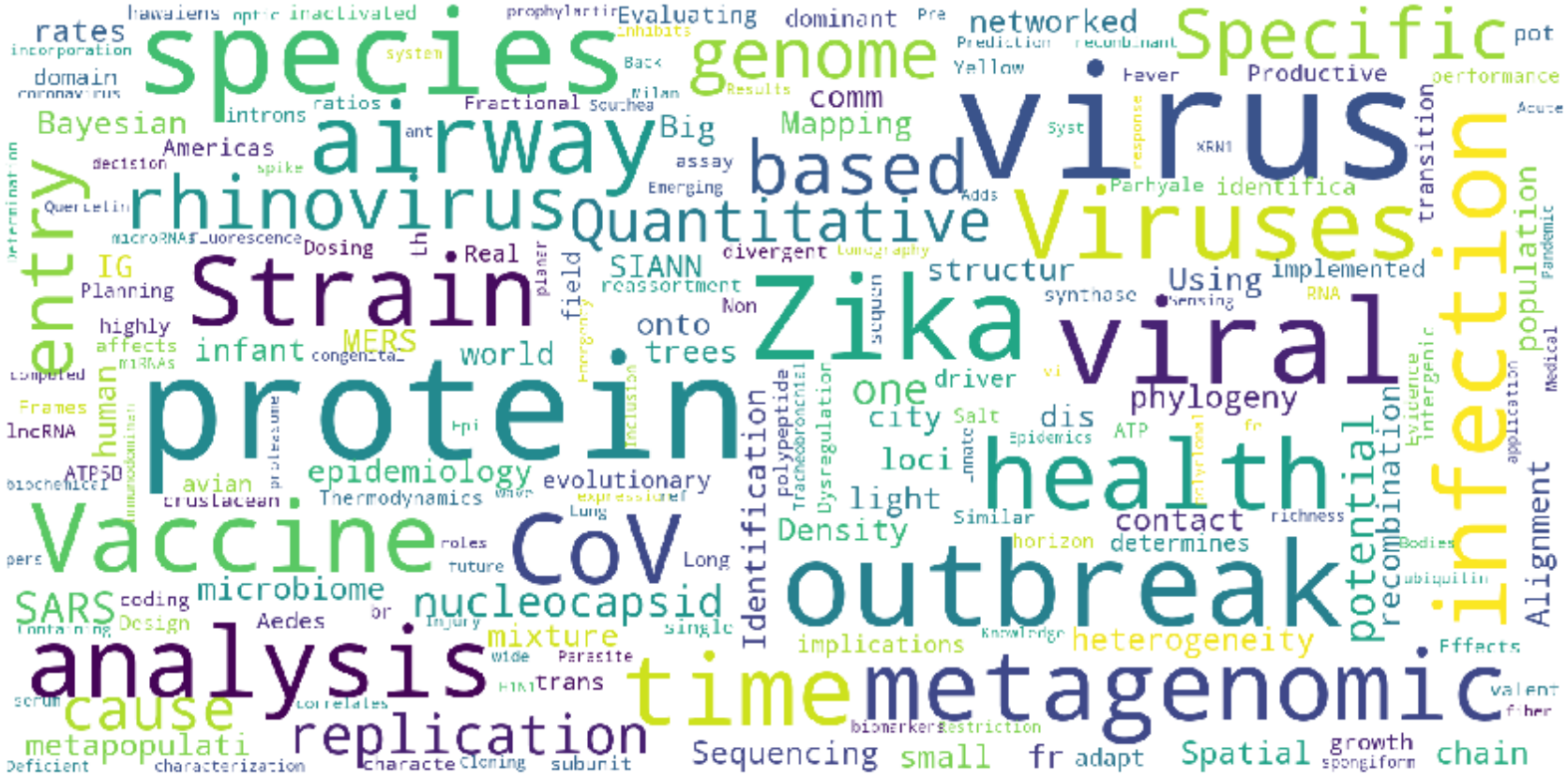
DATA VISUALIZATION AND EXPLORATORY DATA ANALYSIS

- Word cloud visualization for the abstract column in the dataset is given below. This was used for deriving some insights in the last section

Research papers abstract wordcloud

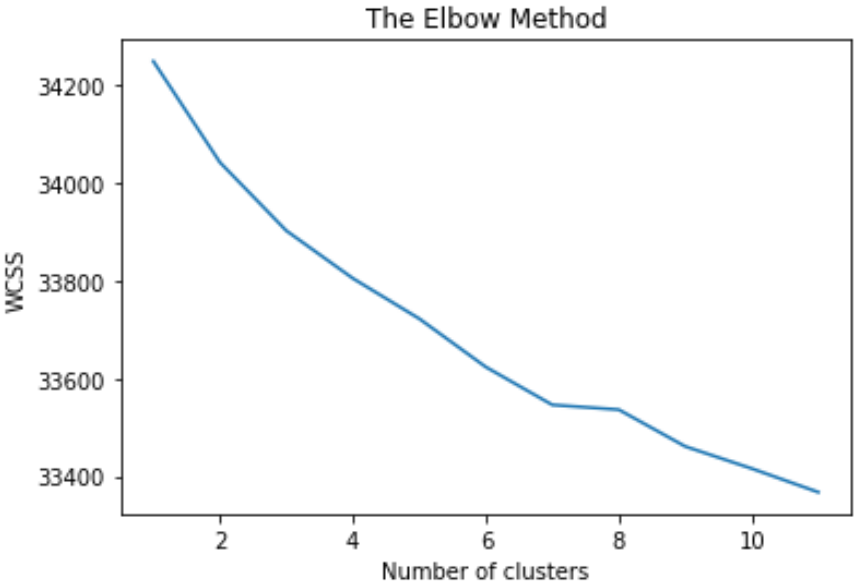
DATA VISUALIZATION AND EXPLORATORY DATA ANALYSIS

- Word cloud visualization for the title column in the dataset is given below. This was used for deriving some insights in the last section



MODEL IMPLEMENTATION

- WCSS is the sum of squares of the distances of each data point in all clusters to their respective centroids. The idea is to minimize the sum. Suppose there are n observation in a given dataset and we specify n number of clusters (k = n) then WCSS will become zero since data points themselves will act as centroids and the distance will be zero and ideally this forms a perfect cluster, however this doesn't make any sense as we have as many clusters as the observations. Thus there exists a threshold value for K which we can find using the Elbow point graph.
- We randomly initialize the K-Means algorithm for a range of K values and will plot it against the WCSS for each K value. For the above-given graph, the optimum value for K would be 7. As we can see that with an increase in the number of clusters the WCSS value decreases. We select the value for K on the basis of the rate of decrease in WCSS. For example, from cluster 1 to 6 in the graph below we see a sudden and huge drop in WCSS. After 7 the drop is minimal and hence we chose 7 to be the optimal value for K.



MODEL IMPLEMENTATION

K-means clustering was implemented after vectorizing the text using tf-idf vectorizer. The number of clusters were chosen based on the elbow method as shown before.

The clusters formed gave some keywords that provided insights from the dataset, which will be discussed in the next section.

INSIGHTS FROM THE ANALYSIS

CLUSTER 0 KEYWORDS:
protein, proteins, virus cells, cell binding, activity, fusion, viral membrane

CLUSTER 1 KEYWORDS:
health, public, disease, outbreak, epidemic, transmission, infectious, diseases, cases, china

Cluster 0

- abstract severe acute respiratory syndrome (sars) is a respiratory disease caused by a newly found virus, called sars coronavirus this risk has been recently reinforced by human epidemics in Singapore of sars coronavirus, 2009 pandemic h1n1 influenza a virus, and enterovirus 71 possible dispersal mechanisms as to how coronaviruses arrived on Madagascar are discussed.
- additional new respiratory viruses detected during follow-up of these 15 patients included rhinovirus (3), metapneumovirus (2), coronavirus (1), respiratory syncytial virus (1), parainfluenza virus (1), and adenovirus (1) we propose that coronavirus ntds originated from a host galectin and retained sugar-binding functions in some contemporary coronaviruses, but evolved new structural features in mhv for mceacam1a binding.
- the present work presents the first review of the fatal novel coronavirus cases in china this review gives an example of one such vaccine platform, replication-deficient simian adenoviruses, and describes progress in human and livestock vaccine development for three outbreak pathogens, ebola virus, rift valley fever virus and middle east respiratory syndrome coronavirus. other viruses, porcine transmissible gastroenteritis coronavirus, avian infectious bronchitis coronavirus, porcine reproductive and respiratory syndrome virus, classic swine fever virus and porcine pseudorabies virus, were unreactive.
- in 2004, a second coronavirus was discovered (cov-nl63) and in 2005 a third new coronavirus was described (cov-hku1) abstract human coronavirus nl63 was identified in 2004 in the Netherlands overall, our findings highlight the role of the host immune response in contri buting to the pathogenesis of coronavirus-induced respiratory disease. the sars coronavirus (sars cov) affects multiple organ systems with severe viral pneumonia as its main clinical manifestation but with diarrhea, lymphopenia, and mild liver dysfunction being common extra -pulmonary manifestations the virus is genetically and antigenically distinct from enteric canine coronavirus; therefore, specific tests are required for diagnos is.
- the gc content of the hcov-nl63 genome is extremely low (34%) compared to other coronaviruses, and we therefore performed additional analysis of the nucleotide composition ,middle east respiratory syndrome coronavirus (mers-cov) is a major emerging zoonotic infectious disease this is the first identification of phosphorylated sites for a group ii coronavirus n protein.

Cluster 1

- summary background in december, 2019, a pneumonia associated with the 2019 novel coronavirus (2019-ncov) emerged in wuhan, china background: since late december 2019, novel coronavirus-infected pneumonia (ncp) emerged in wuhan, hubei province, china recently, some media outlets inappropriately labelled the coronavirus by race, using such headlines as 'chinese virus pandemonium' the outbreak of a novel coronavirus infection can lead to 15% ~ 30% patients developing into acute respiratory distress syndrome (ards) using as an example the coronavirus disease 2019 (covid-19) epidemic, we (a) examine challenges associated with what we term ?global information crises? coronavirus has a history of causing epidemics in human and animals sars-cov-2 is a coronavirus associated with the epidemiological outbreak in late 2019 the newly identified coronavirus in the most-trafficked mammal could represent a continuous threat to public health if wildlife trade is not effectively controlled. the severe acute respiratory syndrome coronavirus 2 (sars-cov-2) was subsequently identified as responsible of this condition, defined coronavirus disease (covid-19).
- background an outbreak of a novel coronavirus (sars-cov-2)-infected pneumonia (covid-19) was first occurred in wuhan, china, in december 2019 and then spread rapidly to other regions overall, this study enhances our understanding of the evolution of coronavirus rbds, provides insights into receptor recognition by mers-cov, and may help control the transmission of mers-cov in humans. the current outbreak of coronavirus disease 2019 (covid-19) has become a global crisis due to its quick and wide spread over the world with this perspective, we give suggestions regarding a potential candidate for the rapid detection of the coronavirus disease 2019 (covid-19), as well as factors for the preparedness and response to the outbreak of the covid-19.
- this review gives an example of one such vaccine platform, replication-deficient simian adenoviruses, and describes progress in human and livestock vaccine development for three outbreak pathogens, ebola virus, rift valley fever virus and middle east respiratory syndrome coronavirus

INSIGHTS FROM THE ANALYSIS

CLUSTER 2 KEYWORDS:

cov, sars, mers, coronavirus, respiratory syndrome, protein, severe, east, middle

CLUSTER 3 KEYWORDS:

rna, viruses, viral, virus, host, replication, genome, proteins, protein, cellular

Cluster 2

- abstract severe acute respiratory syndrome (sars) is a respiratory disease caused by a newly found virus, called sars coronavirus this risk has been recently reinforced by human epidemics in Singapore of sars coronavirus, 2009 pandemic h1n1 influenza a virus, and enterovirus 71 possible dispersal mechanisms as to how coronaviruses arrived on Madagascar are discussed.
- other viruses, porcine transmissible gastroenteritis coronavirus, avian infectious bronchitis coronavirus, porcine reproductive and respiratory syndrome virus, classic swine fever virus and porcine pseudorabies virus, were unreactive. the sars coronavirus (sars cov) affects multiple organ systems with severe viral pneumonia as its main clinical manifestation but with diarrhea, lymphopenia, and mild liver dysfunction being common extra-pulmonary manifestations the virus is genetically and antigenically distinct from enteric canine coronavirus; therefore, specific tests are required for diagnosis. the gc content of the hcov-nl63 genome is extremely low (34%) compared to other coronaviruses, and we therefore performed additional analysis of the nucleotide composition abstract middle east respiratory syndrome coronavirus (mers -cov) is a major emerging zoonotic infectious disease this is the first identification of phosphorylated sites for a group ii coronavirus n protein.

Cluster 3

- moreover, these compounds also inhibit the replication of sars coronavirus and human coronavirus 229e zika virus (zikv), ebola virus (ebov), severe acute respiratory syndrome coronavirus (sars-cov), and middle east respiratory syndrome coronavirus (mers-cov) as (re-)emerging viral pathogens and other enveloped viruses could be efficiently inactivated by both who formulations, implicating their use in healthcare systems and viral outbreak situations.
- additional new respiratory viruses detected during follow-up of these 15 patients included rhinovirus (3), metapneumovirus (2), coronavirus (1), respiratory syncytial virus (1), parainfluenza virus (1), and adenovirus (1) we propose that coronavirus ntds originated from a host galectin and retained sugar-binding functions in some contemporary coronaviruses, but evolved new structural features in mhv for mceacam1a binding.
- infection with gamma herpesviruses, alpha herpesviruses, and beta coronaviruses can result in widespread miRNA degradation, in each case initiated predominantly by a single viral factor human coronaviruses cause respiratory infections that range in seriousness from common colds to severe acute respiratory syndrome it suggests that nonhuman coronaviruses may be attractive new therapeutic agents against human tumors. other viruses, porcine transmissible gastroenteritis coronavirus, avian infectious bronchitis coronavirus, porcine reproductive and respiratory syndrome virus, classic swine fever virus and porcine pseudorabies virus, were unreactive. this well-established multiplex real-time rt-qpcr assay provided a rapid, efficient, specific, and sensitive tool for detection of swine enteric coronaviruses the discovery of bat sars-like coronaviruses and the great genetic diversity of coronaviruses in bats have shed new light on the origin and transmission of sars coronaviruses.
- three miscellaneous positive-strand rna viruses are described briefly with an emphasis on the genome structure: caliciviruses, togaviruses, and coronaviruses background: the genome of coronaviruses contains structural and non-structural genes, including several so-called accessory genes thus, the major ifn-induced antiviral activities that are specifically inhibited by mhv, and possibly by other coronaviruses, remain to be identified.
- many genetic and mechanistic features distinguish the coronavirus replication machinery from that encoded by most other rna viruses this study evaluated survival of two surrogate coronaviruses, transmissible gastroenteritis (tgev) and mouse hepatitis (mhv) conclusion: the possible roles of these genes, and their importance in feline coronaviruses infection, are discussed.

INSIGHTS FROM THE ANALYSIS

CLUSTER 4 KEYWORDS:
patients, respiratory, influenza, children, infections, viruses, rsv, virus, viral, clinical

CLUSTER 5 KEYWORDS:
cells, ifn, immune, infection, cell, mice, il, virus, response, responses

Cluster 4

- pneumoniae, rhinovirus, respiratory syncytial virus (rsv), influenza virus, metapneumovirus, adenovirus', parainfluenza virus and coronavirus in acute respiratory tract infections in children the most commonly acquired respiratory viruses were human rhinovirus, followed by human coronaviruses and influenza a virus, in decreasing order cordata has much potential for the development of antiviral agents against coronavirus and dengue infections.
- the pathogens, including influenza virus, respiratory syncytial virus (rsv), rhinovirus (hrv), adenovirus (adv), herpes simplex virus (hsv), human coronavirus (hcov), streptococcus pneumoniae and haemophilus influenzae, were detected by real-time pcr future studies should explore links between the timing of coronavirus infections and subsequent development of schizophrenia and other disorders with psychotic symptoms.
- purpose to investigate the clinical, laboratory, and imaging findings of emerging coronavirus 2019-ncov pneumonia in humans during our study no prrsv or coronaviruses were detected the identification of picornaviruses and coronaviruses and concurrent typing of influenza a virus by rvp, which are not currently included in our diagnostic testing algorithm, will improve our diagnosis of respiratory tract infections. the receptor-interacting site is conserved in all coronavirus s glycoproteins that engage 9-o-acetyl-sialogycans, with an architecture similar to the ligand-binding pockets of coronavirus hemagglutinin esterases and influenza virus c/d hemagglutinin-esterase-fusion glycoproteins the identified factors could be interesting targets for the development of host-directed antiviral therapy to treat infections with sars-cov or other pathogenic coronaviruses.
- we illustrate the power of these methods by: 1) identifying the sites explaining sars coronavirus differences between human, bat and palm civet samples; 2) showing how cross species jumps of rabies virus among bat populations can be readily identified; and 3) de novo identification of likely functional influenza host discriminant markers. previous studies have reported epidemiological and clinical characteristics of coronavirus disease 2019 (covid-19) this state revealed by cryo-em first time could provide an important information for the identification and relevant clinical research of this new coronavirus. although the precise molecular mechanism of deaminase-dependent inhibition of coronavirus replication remains elusive, our results further our understanding of apobec-mediated restriction of rna virus infections.

Cluster 5

- the pipeline has been validated on human immunodeficiency virus, human parainfluenza virus 1-4, human metapneumovirus, human coronaviruses human enteroviruses/rhinoviruses, measles virus, mumps virus, hepatitis a-e virus, chikungunya virus, dengue virus, and west nile virus, as well the human polyomaviruses bk/jc/mcv, human adenoviruses, and human papillomaviruses these data greatly extend our knowledge of wildlife reservoirs of alphacoronaviruses. soe also inhibited simian immunodeficiency virus infection but failed to block vesicular stomatitis virus, severe acute respiratory syndrome coronavirus, and influenza h5n1 pseudoviruses it is unclear whether newly reported viral respiratory pathogens (such as the middle east respiratory syndrome coronavirus, will be more of a problem in hiv-infected individuals than the general population. objective: to explore imaging characteristics of children with 2019 novel coronavirus (2019-ncov) infection intratracheal instillation of rats with sdav coronavirus caused an acute, self-limited infection that is a useful model for studying the early events of the innate immune response to respiratory coronavirus infections in lungs of the natural virus host.
- the epidemic of 2019 novel coronavirus, later named as severe acute respiratory syndrome coronavirus 2 (sars-cov-2), is still gradually spreading worldwide this unexpected activity for a coronavirus papain-like protease suggests a novel viral strategy to modulate the host cell ubiquitination machinery to its advantage. the receptor-interacting site is conserved in all coronavirus s glycoproteins that engage 9-o-acetyl-sialogycans, with an architecture similar to the ligand-binding pockets of coronavirus hemagglutinin esterases and influenza virus c/d hemagglutinin-esterase-fusion glycoproteins in huh7 cells, 11r exhibits three-digit picomolar activity against middle east respiratory syndrome coronavirus. while overexpression of gilt inhibited the entry mediated by envelope glycoproteins of sars coronavirus (sars-cov), ebola virus (ebov) and lassa fever virus (lasv), depletion of gilt enhanced the entry mediated by these viral envelope glycoproteins here, we report that recombinant human interferon (ifn)-β 1a potently inhibits sars coronavirus replication in vitro.
- recent studies have illuminated the intricately complex replicative organelles of coronaviruses, a group that includes the largest known rna virus genomes moreover, it will be important to discriminate this previously undescribed coronavirus from hcov 229e and oc43 and the severe acute respiratory syndrome coronavirus.

INSIGHTS FROM THE ANALYSIS

CLUSTER 6 KEYWORDS:
Virus, disease, samples, clinical, study, infection, abstract, strains, pedv, vaccine

VACCINES ANALYSIS KEYWORDS:
vaccines, therauptics, vaccine

Cluster 6

- sars-cov-2 is the novel coronavirus responsible for this disease abstract cultures of human rhabdomyosarcoma (rd) and human glioblastoma (u87-mg) were compared for their ability to sustain a persistent infection with coronavirus oc43 these cross-reactive mabs may serve as tools useful for sars-cov-2 research as well as for the development of diagnostic assays for its associated coronavirus disease covid-19.
- the present study was to characterize turkey coronavirus associated with turkey poult enteritis and mortality the more recent design of wide-spectrum inhibitors targeting the coronavirus main proteases may lead to the discovery of new antivirals against multiple cor onavirus induced diseases. we study epidemiological and clinical outcome of 55 asymptomatic carriers who were laboratory-confirmed positive for the sars-coronavirus-2 by testing the nucleic acid of the pharyngeal swab samples abstractsevere acute respiratory syndrome coronavirus 2 is rapidly spreading around the world the present study reports the first complete genome sequence of a feline coronavirus from brazil.
- much progress has been made in understanding the role of structural and accessory proteins in the pathogenesis of severe acute respiratory syndrome coronavirus (sars-cov) infections abstract a competition elisa utilizing a mab directed towards a peplomer protein epitope common to tgev, prcv and related feline and canine coronaviruses is described. coronavirus disease 2019 (covid-19) is a respiratory disorder caused by the highly contagious sars-cov-2 mycobacteriosis should be included in the differential diagnosis for ocular, respiratory, and gastrointestinal diseases; in particular, it should be differentiated from systemic coronavirus infection.
- middle east respiratory syndrome coronavirus (mers-cov) is a recently isolated betacoronavirus identified as the etiologic agent of a frequently fatal disease in western asia, middle east respiratory syndrome the pac-man approach is potentially a rapidly implementable pan-coronavirus strategy to deal with emerging pandemic strains. the severe acute respiratory syndrome coronavirus 2 (sars-cov-2) was subsequently identified as responsible of this condition, defined coronavirus disease (covid-19).

Vaccines Analysis

- the spike (s) glycoprotein of severe acute respiratory syndrome coronavirus (sars-cov) mediates the receptor interaction and immune recognition and is considered a major target for vaccine design it is hoped that this approach will lead to the production of a superior commercial vaccine for the protection of neonatal calves against enteric coronavirus infection.
- developing effective and safe vaccines is urgently needed to prevent infection by severe acute respiratory syndrome (sars)–associated coronavirus (sars-cov) this study reveals a role for endou activity as a virulence factor in pedv infection and provides an approach for generating live-attenuated vaccine candidates for emerging coronaviruses. interpretation the gls-5300 mers coronavirus vaccine was well tolerated with no vaccine-associated serious adverse events conclusion: thus, monitoring of bovine coronavirus in ireland is important as the current isolates in circulation in the south of ireland may be diverging from the available vaccine strain, which may have implications regarding futurebcov vaccine efficacy.
- in this study, we demonstrated that the virus inactivation treatment disrupts its genome integrity seriously when using porci ne epidemic diarrhea virus (vaccine), a kind of coronavirus, as a model more generally, our approach may allow the development of vaccines againstinfections with other pathogenic coronaviruses, including that causing severe acute respiratory syndrome in humans. approaches to the development of mers vaccines are discussed herein, including a summary of previous efforts to develop vaccines useful against human and non-human coronaviruses thus, delta inulin adjuvants may offer a unique ability to develop safer and more effective coronavirus vaccines.
- the vaccine was found to be safe and efficacious in one population of cats that had low antibody titre against feline coronavirus (fcov) at the time of vaccination coupled with increased safety and reduced pathogenesis, the study highlights the potential for 2'o methyltransferase attenuation as a major component of future live attenuated coronavirus vaccines.
- bovine coronavirus (bcov) is antigenically related to ecov; it is therefore possible that bcov vaccine will induce antibodies against ecov in horses this chapter will describe the current state of development of sars vaccines, the issue of coronavirus-associated eosinophilic lung immunopathology and how adjuvants can be used to reduce the risk of this complication.

INSIGHTS FROM THE ANALYSIS

SOURCES OF COVID-19 ANALYSIS KEYWORDS:
livestock, animals, hosts, spillover, suscept

GENETICS OF COVID-19 ANALYSIS KEYWORDS:
field surveillance, genetic sequencing, receptor binding

Sources of COVID-19 Analysis

- coronaviruses (covs) are an important cause of illness in humans and animals importance coronaviruses cause widespread respiratory, gastrointestinal, and central nervous system diseases in humans and other animals, threatening human health and causing economic loss these data indicate that, depending upon route of immunization, mice can become susceptible to reinfection with the same coronavirus strain over time. dromedary camels are important reservoir hosts of various coronaviruses, including middle east respiratory syndrome coronavirus (mers-cov) that cause human infections ghanaian domestic livestock are not likely intermediate hosts of hcov-nl63-related coronaviruses.
- source identification requires detailed epidemiological studies of the infected patients and enhanced surveillance of mers-cov or similar coronaviruses in humans and animals these findings may provide important insights into devising therapeutic strategies and selection of antiviral compounds for further development for important coronaviruses in animals and humans. this event occurred near a region that has been implicated to be the human receptor binding site and may have been directly responsible for the switch of host of the sars coronavirus from animals to humans.
- as various coronaviruses (covs) that affect humans emerged from bats, our study raises the question whether covs such as the one detected in our work are yet-to-be-detected pathogens of humans and animals other than bats. background: studies have reminded that cardiovascular metabolic comorbidities made patients more susceptible to suffer 2019 novel corona virus (2019-ncov) disease (covid-19), and exacerbated the infection together, these data suggest that shrews are important and longstanding hosts for coronaviruses that merit additional research and surveillance. fcv3-70 monoclonal antibody produced immunolabelling of group 1 coronavirus antigen in tissue samples from eight animals, the antigen being present in the cytoplasm of macrophages in the different types of granulomatous lesions.

Genetics of COVID-19 Analysis

- like other coronaviruses such as the sars-cov, the 2019-ncov uses the receptor binding domain (rbd) of the surface spike glycoprotein (s protein) to engage ace2 this phenomenon could be extended to other beta coronaviruses utilizing ctd1 of the s1 subunit for receptor binding, which provides new insights into the intermediate states of coronavirus pre-fusion spike trimer during infection.
- we postulate that the versatility of cell receptor binding strategies has immediate implications on therapeutic strategies.one sentence summary molecular dynamics simulations reveal a temporal dimension of coronaviruses interactions with the host receptor.
- moreover, cleavage was mapped to the same region where, in coronaviruses carrying furin-activated spikes, the receptor binding subunit of the protein is separated from the membrane-anchored fusion subunit.

INSIGHTS FROM THE ANALYSIS

STRAINS OF COVID-19 ANALYSIS KEYWORDS:

strain

Strains of COVID-19 Analysis

- as the subgroup 2 strains have not been isolated for at least 20 years, it appears likely that an unknown avian coronavirus that was the donor of the s1 glycoprotein sequence of n1/88 in the 1980s is still recombining with ibv strains in the field.
- thus, monitoring of bovine coronavirus in ireland is important as the current isolates in circulation in the south of ireland may be diverging from the available vaccine strain, which may have implications regarding future bcov vaccine efficacy.
- sialylation of red blood cells with limiting amounts of sialic acid indicated that strain jhb/1/66 of influenza c virus requires less neu5,9ac2 for agglutination of erythrocytes than the two coronaviruses, both of which were found to be similar in their reactivity with neu5,9ac2-containing receptors.