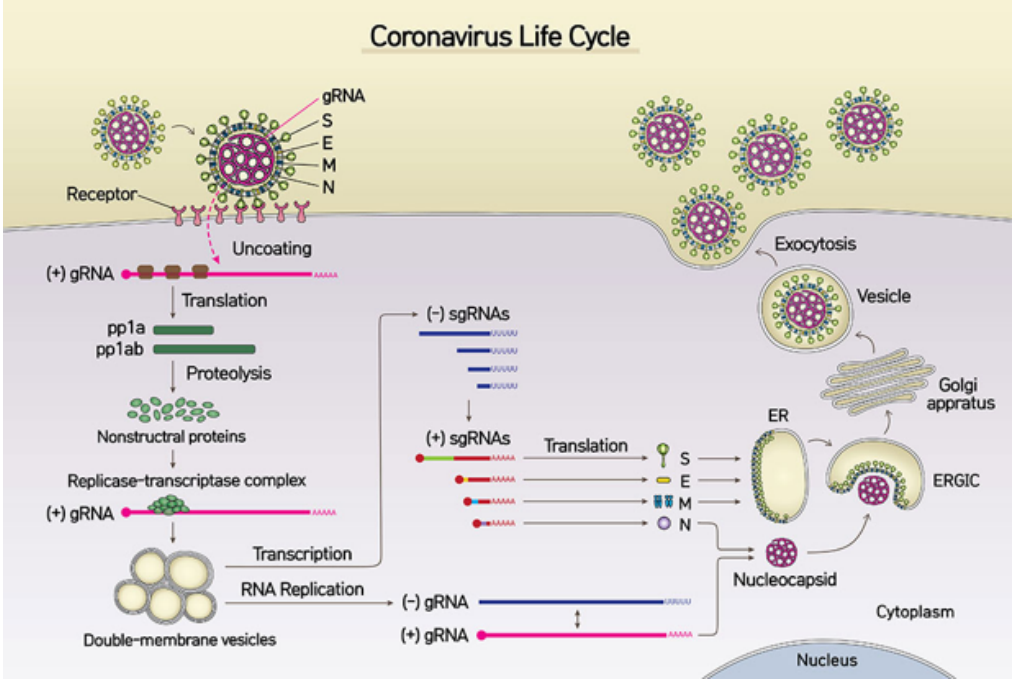
The Hidden Threat: Analyzing protein sequences of animals to identify potential intermediate hosts of SARS-CoV-2 – Background and Approach (Part 1 of 2)

Author: John Pace, PhD
Senior Data Scientist at Mark III Systems

Yesterday I saw the following headline for a piece in the prestigious journal Nature – "Bats are a key source of human viruses — but they're not special." Since SARS-CoV-2, the novel coronavirus that causes COVID-19, was discovered in December 2019, there has been a tremendous amount of research related to it. Much of the work has understandably been clinical since, as of April 15, over 128,000 people have died worldwide from COVID-19. However, much research in the basic science realm has also been conducted.

On March 30, Lan et. al, published the x-ray crystallography determined structure of the SARS-CoV-2 spike receptor binding domain bound to the human ACE2 receptor. Just 3 days before, Yan et. al, published the same structure determined by cryo-EM. Interestingly, the results were slightly different, particularly regarding the amino acids involved in the binding of the two proteins.

If you don't know how significant the determination of these structures is, let me give a little background. The SARS-CoV-2 virus has a portion of its RNA genome that serves as the template to make a protein known as the S protein, which has a section known as the Receptor Binding Domain, or RBD. In humans and other animals, some cells make a cell surface protein called angiotensin-converting enzyme 2 (ACE2). To gain entry into the cell, the RBD of the virus binds to a part of the ACE2 protein and a series of events occurs allowing the virus to (1) gain entry into the cell, (2) hijack the cellular machinery, (3) get replicated repeatedly by the infected cell, then finally (4) be released by the cell and spread. It's a very elegant, yet frightening process where the cell does all the work that will eventually lead to its own death without even realizing it. Such is the trickery of viruses. By determining how the RBD binds to the ACE2 receptor at the amino acid level, researchers can target drugs that will interfere with binding and prohibit the virus from entering the cell.

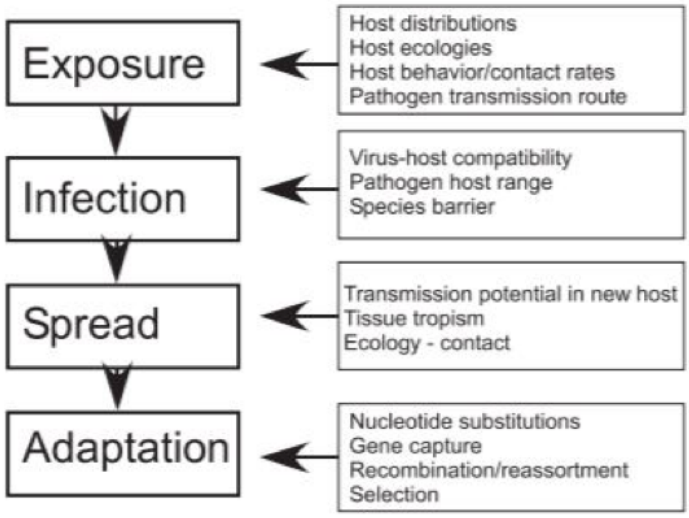


Source Linked

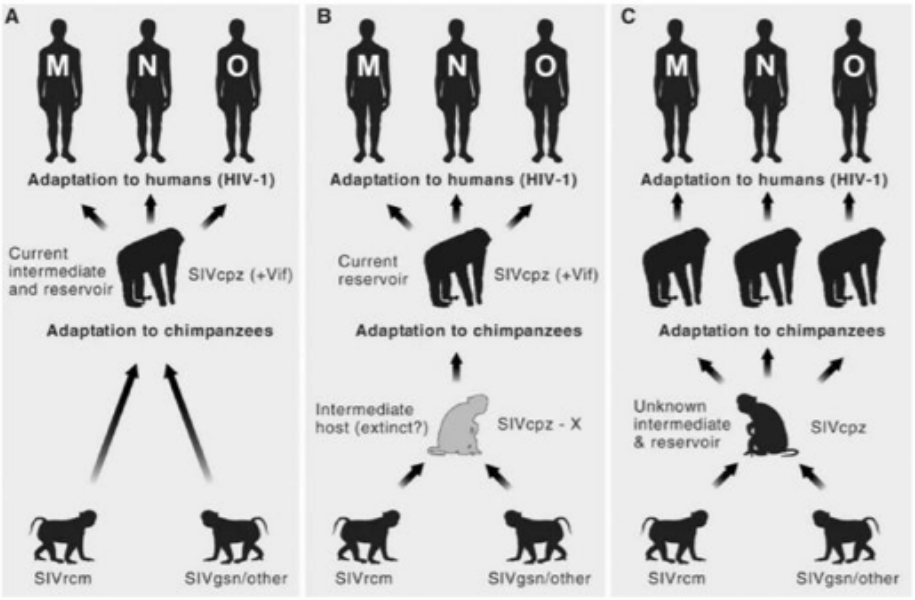
Along with the clinical research, determining where the virus originated and how it was transferred to humans is also critical. Why? Because we must figure out which animal, or animals, is the natural reservoir of the virus. In other words, where is the virus currently residing, mutating, then spreading to other animals from? If we can find this elusive animal, or animals, we can put guidelines in place for humans to avoid them. Unfortunately, there is something that severely complicates this process.

Research has repeatedly shown that viruses can "jump hosts," or experience cross-species transfer from one animal to another. Bats are very often the culprit. Are any of the viruses that cross-species ones that we should be concerned about? If you consider HIV-AIDS, SARS, measles, smallpox, Ebola, bird flu, swine flu, and rabies to be important, then we should be concerned. All evidence is now pointing to SARS-CoV-2, and thus COVID-19, to have entered humans by cross-species transfer. Thus, we HAVE to find the reservoir host since COVID-19 is in the same category as the viruses just listed.

Here is where it gets complicated. Viruses that infect humans don't always cross species boundaries in one step. It's not always bat-to-human direct transfer as with Ebola. At times, there is an animal, or animals, known as "intermediate hosts" (IH), that acquires the virus from the reservoir animal, such as a bat. The version of the virus in the reservoir animal may not be able to directly infect humans. However, once the virus has crossed species into the IH, it evolves and potentially acquires the ability to infect humans. So not only is it critical to find the reservoir host, it is also critical to find the intermediate host since both are involved in the transfer of the virus to humans. This is not an easy process! You can't just go out and test every animal in the world for the presence of a virus. You must narrow the search down considerably. Here are a couple of figures from Parrish et. al that show the steps in the transfer of a virus from reservoir host to intermediate host to humans.



The steps involved in the emergence of host-switching viruses, showing the host and viral processes that can be involved in the transfer and adaptation process. (Source for text and image linked)



Origins of HIV-1 in humans from related viruses in chimpanzees, possible pathways of origin from other primates. (Source for text and image linked)

Back to the Nature article. Bats have been proven to be reservoir hosts for some viruses that can be directly transmitted to humans, i.e., Ebola. Birds and rodents are often implicated as well. A significant open question about SARS-CoV-2 is if the virus was transmitted directly from bats to humans or if there was an intermediate host. As the article states, "Bats are…a prime suspect as the source of SARS-CoV-2, the virus responsible for the current pandemic." However, as of now, the evidence is inconclusive. Other animals that are thought to potentially be intermediate hosts include the pangolin, cat, cow, buffalo, sheep, goat, pigeons, ferrets, civets, and horseshoe bats. Again, the evidence is inconclusive.

Pangolin

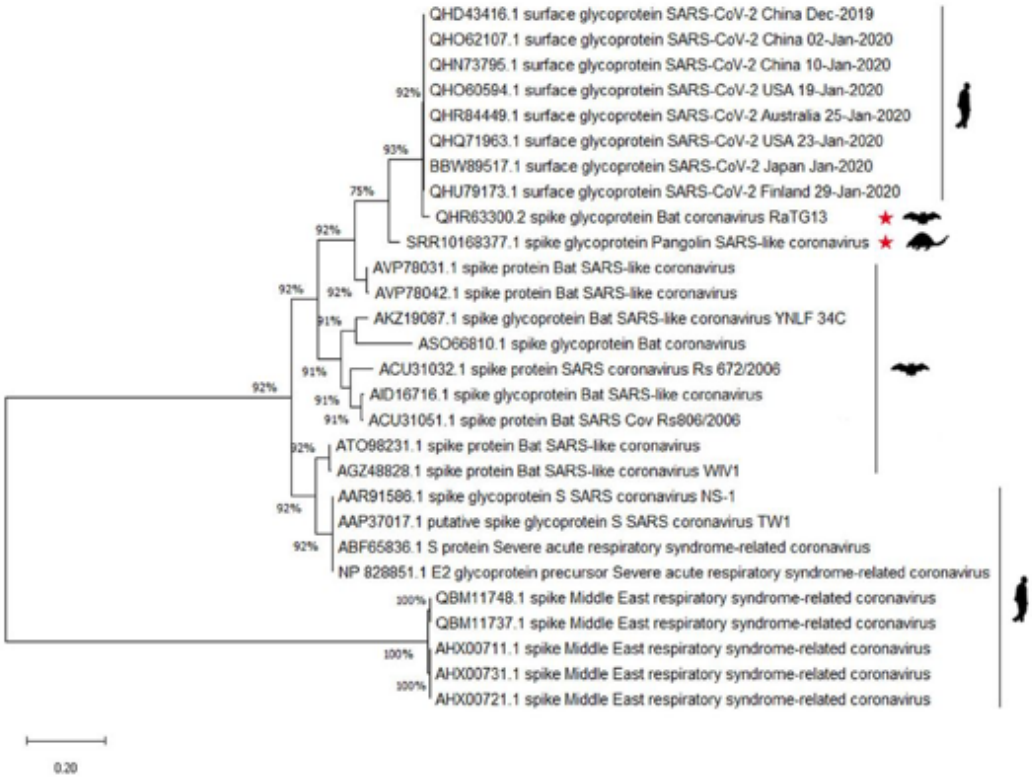
Ferret

Masked palm civet

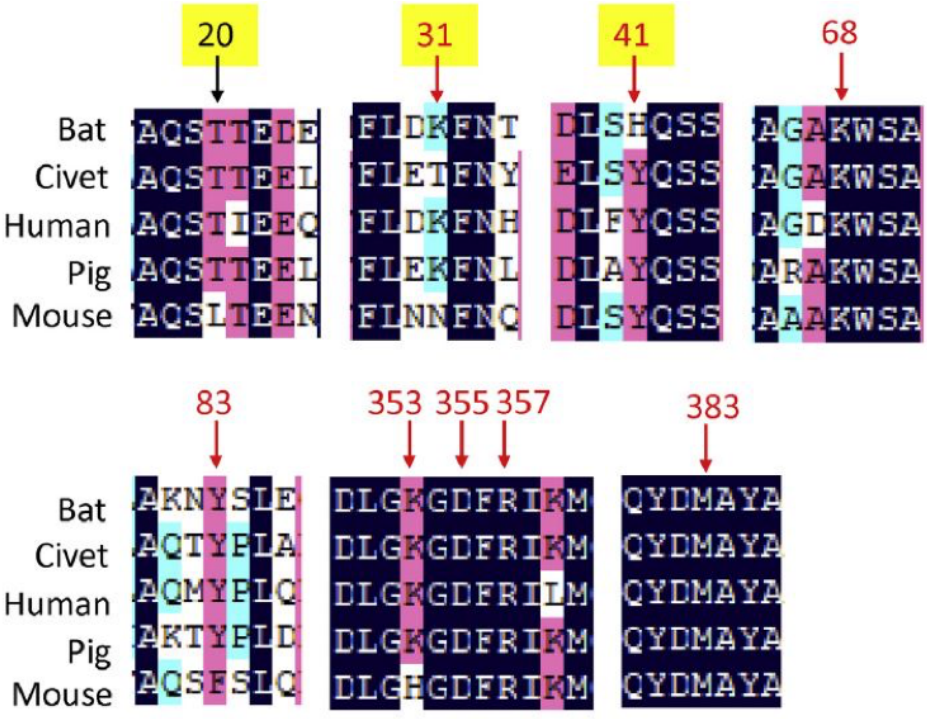
Intermediate horseshoe bat

Two of the lines of research to determine the potential intermediate host have focused on complementary techniques. There are other methods being used as well, such as testing for ACE2 expression in HeLa cells by Zhou et. al, but I'm just going to focus on the first two. In each case, both have used evolutionary comparative genomics. I find this particularly interesting because this is exactly what I worked on for my PhD and what my dissertation was about. The first line of research is looking at the evolutionary history of the viral DNA or proteins sequence of the RBD (see phylogenetic tree below), while the other is looking at the evolution of the host DNA or protein sequences of the ACE2 receptor (see sequence alignment below). Both have yielded some interesting results, but nothing conclusive. A

few studies that focus on the viral approach include Wan et. al, Liu et. al, and Li et. al. The host approach is demonstrated in Hou et. al and Qiu et. al.



Phylogenetic tree showing relationships of coronaviruses. Notice that the SARS-CoV-2 sequences group together at top and that RaTG13 is in the same grouping. (Source Linked)



Alignment of ACE2 amino acid sequences from bat, civet, human, pig, and mouse. (Source Linked)

My research on SARS-CoV-2 is focusing on finding a group of animals that may be an intermediate host. I am examining the evolution of the host protein sequences to see which ACE2 receptors could potentially bind to the RBD based on structural information. I have significantly expanded the range of animals that could be intermediate hosts. I think that if you limit the search to just animals that have previously been found to carry SARS-like viruses, then a massive amount of data, as well as other viable animals can be overlooked. With the wealth of DNA and protein sequences in GenBank, the problem is difficult but tractable. In order to speed the work, I am utilizing NVIDIA V100 graphics processing units (GPUs) running on a couple different GPU-accelerated server platforms, including NVIDIA DGX-1, for BLASTp searches and analysis of the sequences. I don't want to go into too much detail now, but suffice it to say, the results have been quite interesting. Some of my results implicate animals listed above, but many new animals are also predicted. I will submit the manuscript for publication this month and it will include all of the details.

Ultimately, SARS-CoV-2 research is aimed at stopping the spread of the pandemic and saving lives. I'll skip all the COVID-19 buzzwords and just say that by using the technology we possess, such as DNA sequencing, cryo-EM, x-ray crystallography, artificial intelligence, and ultra-high powered computing resources, we can make a difference and save lives. Multiple, varied disciplines - computer science, basic and clinical sciences, epidemiology, statistics, and even human intuition - working together are stronger than any one discipline. Together, we can – and will – make a difference.

Feel free to share links to other research that you think is important in the fight against COVID-19.

Contact us:
contact@markiiisys.com
www.markiiisys.com