APS COVID Webinar: October 7, 2020

What we know and don't know about SARS-CoV-2: Origins and Evolution

Raul Rabadan, director of the Program for Mathematical Genomics and director of the Center for Topology of Cancer Evolution and Heterogeneity at Columbia University

PART 1: *Evolution of Coronaviruses*

- COVID-19 is the disease caused by the SARS-CoV-2 virus in humans
    - So in order to understand evolution of the disease, we need information about both the virus and the human (i.e genomic info) – this talk is about the virus information
- How do coronaviruses evolve?
    - 1. Mutations
        - SARS is an RNA virus
        - Smaller genome size (bp) leads to faster mutation rate
            - Thus viruses mutate at a much higher rate than eukaryotes and bacteria
        - Use phylogenetic trees to represent the evolution/history of how the virus evolves
        - Mutation tracking:
            - 100,000 genomes of this virus that have been collected around the world since December
            - Branches show the relationship between measured genomes
            - Tells us that most of the viruses are very similar, which tell us that this virus is new in the population (the Wu-Han outbreak is very close to the origin of the virus); has not been circulating in humans for very long
            - Can see by the clusters where the virus branched from
                - Example: US cluster in Washington came from one branch from China
                - Suggests that international travel played a very instrumental role in early spread
    - 2. Recombination
        - Can create highly non-local jumps in the branches
        - One branch can quickly acquire traits of different branches
- Where is SARS-CoV-2 coming from?
    - Is a sarbecovirus (subset of beta-coronaviruses)
    - SARS is a subtype of coronavirus called beta-coronaviruses
        - We know of 4 of these in humans:
            - SARS-CoV-1 in 2002-2003
                - Coronaviruses in bats are very similar to what we saw in humans for SARS-CoV-1 and current SARS-CoV-2
            - MERS-CoV
                - Come from other animals (pangolins)
            - OC43 and HKU1 have been circulating for long time
                - Usually kids and usually asymptomatic
                - Distantly related to current virus
- Recombination is pervasive in beta-coronaviruses (extremely frequent)
    - Showed plots that show a jump in recombination events in a particular region

        - Looking more closely at the phylogenetic structure of this particular region…
            - When looking at whole genome see relation to virus in pangolins
            - When looking at specific region, see more similarity to SARS-CoV-1 and virus in bats
        - This region appears to be highly important because of receptor-binding domain (RBD)
- Topological inconsistencies in SARS-CoV-2
    - Seems that at some point there was an exchange of information between a close ancestor of SARS-CoV-1 and 2 (recombination event)
    - Studied Rosetta binding energy for each branch of the sarbecoviruses
        - Shows difference in binding affinity for various branches of the virus
    - Receptor-binding domain (RBD) recombination events study suggest that many events that happened ~2007 and ~2010 mutated to create SARS-CoV-2
- SUMMARY:
    - Coronavirus mostly evolve through point mutations and recombinations
    - Recombinations are extremely frequent in beta-coronaviruses
    - Reconstruction of ancestral states to CoV-2 suggest a two hit scenario model:
        - Recombination picks up SARS-like RBD - > enables binding to human ACE2 receptor
        - Further mutations refine the interactions between RBD and receptor
    - Ancestral reconstruction analysis indicates active circulation of potentially human infecting sarbecoviruses for the last 20 years

PART 2: *Tools and Methods*

- Are phylogenetic trees a good way of thinking about viral data?
    - Standard method, but have shown that looking as specific regions of the genome can be very impactful (looking at subsections of the trees is important)
- Practical problem: Given a set of genomes and a way of comparing them, how do we represent their relationship without any assumptions about species or non-vertical genetic exchange?
    - 3 important features to consider:
        - Type of evolution
        - Frequency
        - Scale of exchange
- Could topological analysis be a better method?
- Topological data analysis
    - Homology groups capture global properties about shapes of spaces
        - Count the number of n dim objects that are not a boundary of an n+1 object
        - Ranks are called Betti numbers (b0: connected, b1: holes, b2: voids/cavities)
    - How do we uncover underlying topological invariants with real (noisy) samples?
        - Examine the persistent homology
            - Calculate the number of connected components across different scales
- What is the topology of space sampled by genomes?
    - Represent evolutionary relationships of a large set of genomes by persistence complex
        - Simulation without recombination only gives 0-dim connections
        - With recombination then results in 0 and 1 dimensional connections
    - Genomic data -> set of evolutionary complexes -> properties of the evolution process
    - 3 examples:
        - 1. Influenza A: reassortments

        - Several are circulating, every ~30 years we get a pandemic due to a recombination jump from a branch that has been circulating in animals
        - H1N1: Trying to explore the origin of the reassortments of the swine, human, and avian versions of the virus
            - Found no recombination (all b0 connections) within one segment of the virus
            - Found higher dimension connections when looking at combinations of segments
            - Explored whether particular segments like to travel together
    - 2. Sarbecoviruses
        - Find many recombination events in a particular region
    - 3. HIV
        - Show many voids
- SUMMARY:
    - Evolution occurs in a high dimensional space (genomics sample this space)
    - Want to understand the structure of this space from sample points
    - Algebraic topology (TDA) provides a way to capture properties about the shapes of the evolutionary spaces
        - Type: including complex exchange of genomic material keeping track of the scale of the process
        - Scale: keeps track of the evolutionary scale
        - Numbers: statistics of shapes (how many objects and at what scale)

Persistent homology	Viral Evolution
Filtration value ε	Genetic distance (evolutionary) scale
0-dimensional Betti number at filtration value ε	Number of clusters at scale ε
Generators of 0-dimensional homology	A representative element of the cluster
Hierarchical relationship among generators of 0-dimensional homology	Hierarchical clustering
1-dimensional Betti number	Number of irreducible recombination/reassortment events
Generators of 1-dimensional Homology	Recombinant/reassortant events
Generators of 2-dimensional Homology	Complex Horizontal Genomic Exchange
Number of higher dimensional generators in time frame	Lower bound on Recombination/Reassortment rate
Non-zero high dimensional homology (topological obstruction to phylogeny)	No phylogenetic representation

QUESTIONS:

- How can you find shared ancestry using topological data?
    - Topology has no time direction, so we have to impose additional restrictions to incorporate these dynamic concepts
- How do you examine functional impact of mutations and recombinations?
    - Still needs much study
- How much evidence do we have to find the exact mutations and their locations?
- Transmission of virus between humans
- Does the virus evolve more or less in humans who are more affected by the virus (more symptoms)?
    - No evidence to show this. Yet? Only small amount of data

- Would SARS-CoV-1 cause illness if reintroduced to the population now? Or do we have vaccinations?
    - Evidence that it was more confined to the hospital environment than this one
- Comparing to SARS-CoV-1, very few people didn't have symptoms… can we find a location in the genome that explains why this is different in CoV-2?
    - Need more experimental testing to understand this better
- Noise in the data… how much does the noise really affect your studies? There is a lot of sampling
    - Noise in this technology is very low. Noise is a minor contribution in the technology
    - Noise discussion earlier was about biological noise, not technological
- Is there a way to extrapolate to future predictions of mutations?
    - Can see that the number of mutations will continue to increase with time
    - Can predict that recombinations will certainly happen – don't know how frequent, likely very rare