WEBVTT

00:00:18.750 --> 00:00:21.340 - All right, I see more people joining

00:00:31.960 --> 00:00:34.760 Jeff, how long do you how long do you have like an hour?

00:00:35.633 --> 00:00:36.466 Less than that?

00:00:36.466 --> 00:00:39.510 - I think I can probably finish in less than an hour.

00:00:39.510 --> 00:00:41.157 - Less than hour, all right.

00:00:58.180 --> 00:01:00.113 I think we should get started.

00:01:01.580 --> 00:01:03.270 So hi, everyone.

00:01:03.270 --> 00:01:06.810 Welcome to our seminar series on COVID-19,

00:01:06.810 --> 00:01:10.260 organized by the Department of Biostatistics.

00:01:10.260 --> 00:01:14.750 I'm very pleased to have here today, Jeff Thompson,

00:01:14.750 --> 00:01:19.750 Professor of biostatistics, Ecology and Evolutionary Biology

00:01:20.330 --> 00:01:22.523 from the Yale School of Public Health.

00:01:23.400 --> 00:01:26.670 Thank you, Jeff, for being here today with us.

00:01:26.670 --> 00:01:29.690 As usual, you're welcome to write questions

00:01:29.690 --> 00:01:34.573 in the chat box or even unmute yourself, if you can,

00:01:34.573 --> 00:01:38.151 and other people are not talking.

00:01:38.151 --> 00:01:42.191 And, Jeff, why don't you take it from here?

00:01:42.191 --> 00:01:44.817 - Okay, thank you very much for the introduction, Laura.

00:01:44.817 --> 00:01:45.650 I'm really pleased to have an opportunity to talk

00:01:45.650 --> 00:01:48.267 about the work that we've been doing.

00:01:49.164 --> 00:01:51.882 I think like many speakers in this series, you know,

00:01:51.882 --> 00:01:54.013 we've been doing a lot of work very hard

00:01:54.013 --> 00:01:56.993 on a short period to try to get some progress on COVID-19.

00:01:57.830 --> 00:01:59.300 Ironically, this is the first work

00:01:59.300 --> 00:02:02.900 I think that I started In response to the COVID-19 epidemic

00:02:02.900 --> 00:02:07.374 and it's turned out to be a lot of work.

00:02:07.374 --> 00:02:08.953 So it's actually gotten the least far.

00:02:11.276 --> 00:02:12.528 So we've done a little bit of work, for instance,

00:02:12.528 --> 00:02:13.673 on epidemic modeling of COVID-19.

00:02:14.948 --> 00:02:17.919 That's already, it's actually been submitted,

00:02:17.919 --> 00:02:20.190 I actually have some other work on quarantine

00:02:20.190 --> 00:02:23.830 and stuff that turns out to be really interesting

00:02:23.830 --> 00:02:25.443 and far along in the research.

00:02:26.380 --> 00:02:27.790 And then this work, which I started early on,

00:02:27.790 --> 00:02:30.762 which is more evolutionary, and looking at the zoonotic

00:02:30.762 --> 00:02:32.390 process has gone a little bit slower.

00:02:32.390 --> 00:02:34.592 So what that means is consistent with

00:02:34.592 --> 00:02:35.480 many other speakers in this series,

00:02:35.480 --> 00:02:37.716 I'm gonna be talking a lot about

00:02:37.716 --> 00:02:40.265 the methods that we're going to be using,

00:02:40.265 --> 00:02:43.089 which are well developed, and what we're planning to do,

00:02:43.089 --> 00:02:44.110 I don't have a lot of results.

00:02:44.110 --> 00:02:47.076 But I think that's consistent with these talks in general.

00:02:47.076 --> 00:02:48.910 So hopefully, that will be of interest to you

00:02:48.910 --> 00:02:53.330 and also be illuminating in terms

00:02:53.330 --> 00:02:58.330 of possible research approaches towards this kind of work.

00:02:58.340 --> 00:03:00.020 So as Laura mentioned,

00:03:00.020 --> 00:03:02.120 I use a lot of evolutionary approaches

00:03:02.120 --> 00:03:04.180 to do my analyses of things.

00:03:04.180 --> 00:03:08.220 And the title of this talk is model averaged estimation

00:03:08.220 --> 00:03:11.500 of molecular evolution and natural selection

00:03:11.500 --> 00:03:14.240 in SARS coronavirus, one and SARS coronavirus two

00:03:14.240 --> 00:03:18.000 two Corona viruses during the zoonotic period.

00:03:18.000 --> 00:03:21.170 So what was attracting my interest in this particular case

00:03:21.170 --> 00:03:24.729 is that it's usually very difficult and challenging to find.

00:03:24.729 --> 00:03:27.480 And I'll get to this later in the talk to figure

00:03:27.480 --> 00:03:29.480 out what's going on during the zoonotic period,

00:03:29.480 --> 00:03:32.233 because you don't usually get much sampling there.

00:03:32.233 --> 00:03:34.700 So, what I wanted to do was apply some techniques

00:03:34.700 --> 00:03:37.700 that I've developed to this problem.

00:03:37.700 --> 00:03:39.400 And I will get to those techniques

00:03:40.659 --> 00:03:42.849 and the application to this problem.

00:03:42.849 --> 00:03:45.736 But I first just wanna give a little bit of introduction,

00:03:45.736 --> 00:03:46.790 I think, maybe from a statistics point of view

00:03:46.790 --> 00:03:49.330 towards some of the methodologies that we're using,

00:03:49.330 --> 00:03:51.210 just so everyone can sort of see on board

00:03:51.210 --> 00:03:53.330 at least how I see this as contributing

00:03:54.643 --> 00:03:57.040 to interesting statistical questions.

00:03:57.040 --> 00:03:59.889 So and in a broad sense, if I can get this to Move forward.

00:03:59.889 --> 00:04:01.320 Here we go.

00:04:01.320 --> 00:04:02.780 I think one of the most intriguing

00:04:02.780 --> 00:04:04.900 and interesting and challenging areas of mathematics

00:04:04.900 --> 00:04:07.610 and statistics is understanding this border

00:04:07.610 --> 00:04:09.280 between the discrete and the continuous.

00:04:09.280 --> 00:04:12.550 So these are just some one particular

00:04:12.550 --> 00:04:15.730 example you can pick out is, if you look at discrete

00:04:15.730 --> 00:04:18.711 and continuous distributions that are frequently

00:04:18.711 --> 00:04:21.360 in use in statistical probabilistic analyses,

00:04:21.360 --> 00:04:25.240 we have the geometric and negative binomial distributions.

00:04:25.240 --> 00:04:27.840 And we have the exponential and gamma distributions.

00:04:29.809 --> 00:04:31.906 These are basically essentially waiting for discrete events

00:04:31.906 --> 00:04:33.340 when you have a probability over time.

00:04:33.340 --> 00:04:35.217 We're waiting for the earth event if you

00:04:35.217 --> 00:04:36.709 have probably over time,

00:04:36.709 --> 00:04:39.160 and they correspond to the distributions on a continuous

00:04:39.160 --> 00:04:42.450 time for the wait for the first event

00:04:42.450 --> 00:04:44.650 or the wait for the alpha event.

00:04:44.650 --> 00:04:46.330 So there's a real clear correspondence

00:04:46.330 --> 00:04:47.670 between these two distributions.

00:04:47.670 --> 00:04:49.690 And you can actually see in the mathematics,

00:04:49.690 --> 00:04:51.183 how they're similar as well.

00:04:52.558 --> 00:04:54.190 And that correspondence is kind of interesting.

00:04:54.190 --> 00:04:56.280 And the reason why I say it's interesting is

00:04:56.280 --> 00:04:59.034 because often many of the biggest problems I think

00:04:59.034 --> 00:05:00.820 we wrestle with in statistics are when we're trying

00:05:00.820 --> 00:05:03.840 to deal with data that is some intermediate

00:05:03.840 --> 00:05:06.600 level between continuous and discrete,

00:05:06.600 --> 00:05:08.470 and where we're trying to figure out which

00:05:08.470 --> 00:05:11.288 approach is the best to use, should we use some sort

00:05:11.288 --> 00:05:12.830 sort of parameterize distribution to address it?

00:05:12.830 --> 00:05:15.290 Or should we use some sort of nonparametric

00:05:16.731 --> 00:05:17.780 approach based on the discrete?

00:05:17.780 --> 00:05:19.300 I'm not sure in any particular case.

00:05:19.300 --> 00:05:21.010 But I just wanna mention

00:05:21.010 --> 00:05:21.843 that I think that's a very interesting area.

00:05:21.843 --> 00:05:23.480 And the technique I'm gonna tell you about

00:05:23.480 --> 00:05:26.910 is definitely wrestling with exactly this kind of question.

00:05:26.910 --> 00:05:28.540 So what kind of question do I mean?

00:05:28.540 --> 00:05:32.050 Well, I mean, questions that deal with state spaces,

00:05:32.050 --> 00:05:35.890 over time, or over any discrete or continuous axis.

00:05:35.890 --> 00:05:39.970 And you can see in this diagram just give you a picture

00:05:39.970 --> 00:05:42.660 of the kinds of problems that one deals with

00:05:42.660 --> 00:05:45.420 between discrete and continuous measures.

00:05:45.420 --> 00:05:47.950 You can have here it's depicted as time,

00:05:47.950 --> 00:05:50.640 you could have a discrete state space,

00:05:50.640 --> 00:05:52.890 state space you're measuring over time,

00:05:52.890 --> 00:05:56.240 you could have a continuous sorry,

00:05:56.240 --> 00:05:59.270 you're gonna have discrete measurements

00:05:59.270 --> 00:06:01.400 over where You've got discrete time

00:06:01.400 --> 00:06:03.480 in a discrete state space,

00:06:03.480 --> 00:06:05.900 you could also have discrete time

00:06:05.900 --> 00:06:08.210 and a continuous state space.

00:06:08.210 --> 00:06:09.960 You can have continuous, continuous

00:06:11.638 --> 00:06:13.012 or you can have discrete, continuous.

00:06:13.012 --> 00:06:15.380 And this two on the bottom are, two on the left,

00:06:15.380 --> 00:06:17.429 sorry, are the relevant ones for

00:06:17.429 --> 00:06:18.520 what I wanna talk to you about.

00:06:18.520 --> 00:06:21.660 In my research, which is largely focused

00:06:21.660 --> 00:06:26.050 on informatik data that we can obtain from sequencing

00:06:26.050 --> 00:06:28.388 or other approaches like that.

00:06:28.388 --> 00:06:30.050 A lot of what we're trying to do is look at these discrete

00:06:30.050 --> 00:06:34.145 linear sequences that have sites DNA sites or amino acid

00:06:34.145 --> 00:06:37.100 sites and trying to understand is there some

00:06:37.100 --> 00:06:39.760 pattern in those sites that allows us to understand

00:06:39.760 --> 00:06:41.450 something about the biology of the organism

00:06:41.450 --> 00:06:44.590 or the biology that we want to know something more about?

00:06:44.590 --> 00:06:47.884 So what essentially I'm gonna be doing

00:06:47.884 --> 00:06:50.053 is telling you about approach an approach

00:06:50.053 --> 00:06:53.730 that takes essentially discrete items over some X axis

00:06:53.730 --> 00:06:55.760 here, in which case in my case, it's always going to be

00:06:55.760 --> 00:06:58.280 sequence space, like the nucleotides

00:06:58.280 --> 00:07:00.540 or the amino acids of a sequence.

00:07:00.540 --> 00:07:03.920 And turns it into these kinds of more discrete models.

00:07:03.920 --> 00:07:07.142 And then in some, in a procedure that I'm going to tell you

00:07:07.142 --> 00:07:09.090 about actually gives us more of a continuous measure

00:07:10.405 --> 00:07:13.290 over that space, it's not completely continuous,

00:07:13.290 --> 00:07:14.470 it actually is on every site.

00:07:14.470 --> 00:07:17.010 But when you work with hundreds of sites,

00:07:17.010 --> 00:07:18.810 it turns out to look very continuous

00:07:19.727 --> 00:07:20.953 in terms of how it appears.

00:07:22.259 --> 00:07:23.092 But it's done with a discrete model

00:07:23.092 --> 00:07:24.330 that looks over multiple sites.

00:07:24.330 --> 00:07:26.280 So well, I'll tell you how it works in a moment.

00:07:26.280 --> 00:07:28.300 And I hope it's of interest to you guys.

00:07:28.300 --> 00:07:30.640 So just to introduce that, in general,

00:07:30.640 --> 00:07:33.620 the lab has worked on a lot of different kinds of data,

00:07:33.620 --> 00:07:35.950 and including things like gene expression data

00:07:35.950 --> 00:07:39.130 that borders this discrete continuous measurement.

00:07:39.130 --> 00:07:41.710 The old micro arrays we used to use give us

00:07:42.559 --> 00:07:43.900 essentially continuous measures of gene expression.

00:07:43.900 --> 00:07:45.903 Now we get discrete counts

00:07:45.903 --> 00:07:49.230 from our census sequencing approaches.

00:07:49.230 --> 00:07:50.870 Then all the sequence data we work with

00:07:50.870 --> 00:07:53.480 often ends up being essentially clusters

00:07:53.480 --> 00:07:55.750 of sites and various kinds.

00:07:55.750 --> 00:07:58.880 And then we also use a lot of phylogenetic inference,

00:07:58.880 --> 00:08:01.140 which is another kind of just discrete modeling

00:08:01.140 --> 00:08:03.160 in terms of the topology, but the borders

00:08:03.160 --> 00:08:05.780 between these two because we have discrete modeling of the

00:08:06.840 --> 00:08:07.890 topology, there are certain topologies

00:08:09.600 --> 00:08:11.704 that the taxa that we're interested in looking at

00:08:11.704 --> 00:08:13.310 that show their relationship to each other.

00:08:13.310 --> 00:08:15.190 At the same time, there's also a continuous

00:08:15.190 --> 00:08:17.420 measure out of that, which is these branch lengths,

00:08:17.420 --> 00:08:19.210 or how diverge these different tacks

00:08:19.210 --> 00:08:22.193 are from each other and constructing the phylogeny.

00:08:22.193 --> 00:08:23.950 So this sort of border between discrete

00:08:23.950 --> 00:08:27.640 and continuous measures, always sort of plagues

00:08:27.640 --> 00:08:30.090 and intrigues me, I guess it would be the question.

00:08:30.090 --> 00:08:31.680 Okay, so what am I gonna do today?

00:08:31.680 --> 00:08:34.520 What I wannado today is talk about

00:08:34.520 --> 00:08:37.290 maximum likelihood model averaging to profile clustering

00:08:37.290 --> 00:08:39.540 of site types across discrete linear sequences.

00:08:39.540 --> 00:08:40.780 So at the very base level,

00:08:40.780 --> 00:08:43.610 how do we take kind of these discrete sequences

00:08:43.610 --> 00:08:45.760 of amino acids or nucleotides

00:08:45.760 --> 00:08:49.610 and understand whether sites are closer to each other

00:08:49.610 --> 00:08:51.210 or farther apart from each other

00:08:52.115 --> 00:08:52.948 this is the question are they just uniformly

00:08:52.948 --> 00:08:54.760 distributed site types across a sequence?

00:08:54.760 --> 00:08:57.110 Are they clustered close together or far apart?

00:08:58.330 --> 00:09:01.135 Secondly, I'm gonna talk about how we can

00:09:01.135 --> 00:09:03.650 then use that approach to understand whether sites

00:09:03.650 --> 00:09:07.360 are under selection in a gene expressed in a sequence.

00:09:07.360 --> 00:09:09.190 And what I mean by under selection is that,

00:09:09.190 --> 00:09:11.670 in fact, sites are changing in a rapid

00:09:11.670 --> 00:09:14.430 or at a more rapid pace than you'd expect simply

00:09:14.430 --> 00:09:16.199 by mutation alone.

00:09:16.199 --> 00:09:17.929 So mutation, of course, is going to introduce

00:09:17.929 --> 00:09:19.050 variation into a genetic sequence.

00:09:19.050 --> 00:09:21.460 But when you see changes that are happening faster

00:09:21.460 --> 00:09:23.330 over time in a population,

00:09:23.330 --> 00:09:25.997 then mutation alone would produce

00:09:25.997 --> 00:09:28.670 that implies that every time that mutation is happening,

00:09:28.670 --> 00:09:29.503 it's spreading across the population.

00:09:29.503 --> 00:09:31.310 And that's why you see that uptick

00:09:31.310 --> 00:09:33.720 in the rate of change of those sites.

00:09:33.720 --> 00:09:35.610 So we can actually use this clustering approach

00:09:35.610 --> 00:09:38.210 to identify regions of the gene that have

00:09:38.210 --> 00:09:40.750 that sort of uptick and I'll explain how we do that.

00:09:40.750 --> 00:09:43.360 Now lastly, I'm just going to show you a very few slides

00:09:43.360 --> 00:09:44.800 on the title of the talk,

00:09:44.800 --> 00:09:47.540 which is this model average estimation of the molecular

00:09:47.540 --> 00:09:50.600 evolution and natural selection in SARS Coronavirus one

00:09:50.600 --> 00:09:53.493 and SARS Coronavirus two during the zoonosis.

00:09:55.020 --> 00:09:56.800 So by the time we refer to these,

00:09:56.800 --> 00:09:59.440 I'll just let you know we're almost done with the talk.

00:09:59.440 --> 00:10:01.160 AlL right, so to talk about the first one

00:10:01.160 --> 00:10:03.390 maximum likelihood model averaging five clustering

00:10:03.390 --> 00:10:06.153 of sites across the street linear sequences.

00:10:08.860 --> 00:10:11.299 I just want to... (phone ringing)

00:10:11.299 --> 00:10:14.716 Sorry, emphasize that we wanna figure out

00:10:20.430 --> 00:10:22.390 whether site types are clustered within a linear sequence.

00:10:22.390 --> 00:10:24.350 This sounds like a very straightforward

00:10:24.350 --> 00:10:26.831 statistical question seems like something

00:10:26.831 --> 00:10:28.441 that should have been addressed many, many times

00:10:28.441 --> 00:10:29.320 in the statistical literature.

00:10:29.320 --> 00:10:30.470 Much to my surprise,

00:10:30.470 --> 00:10:34.070 it's actually not terribly well explored.

00:10:34.070 --> 00:10:35.645 You have a linear sequence,

00:10:35.645 --> 00:10:37.630 it's so long and you have site types of one type

00:10:37.630 --> 00:10:39.420 or another are they clustered next to each other?

00:10:39.420 --> 00:10:41.600 Well, if you know the bounds of the region of interest,

00:10:41.600 --> 00:10:43.150 and others, if you can describe oh,

00:10:43.150 --> 00:10:45.450 it's I'm interested in this domain right here,

00:10:46.331 --> 00:10:48.228 and it's from site to site 90 or some other description.

00:10:48.228 --> 00:10:49.434 If you know the bounds,

00:10:49.434 --> 00:10:52.090 it's very simple to analyze that kind of data.

00:10:52.090 --> 00:10:54.810 You can just quantify the site type proportions

00:10:54.810 --> 00:10:56.630 within and outside those bounds.

00:10:56.630 --> 00:10:59.419 use something like a straightforward fisher's exact

00:10:59.419 --> 00:11:01.030 test for significance extremely simple problem.

00:11:01.030 --> 00:11:03.590 But what if you don't actually know those bounds?

00:11:03.590 --> 00:11:04.950 What if you don't know even what you're looking for exactly?

00:11:04.950 --> 00:11:07.090 you just know you're interested in concentrations

00:11:07.090 --> 00:11:09.700 of one site type compared to another site type

00:11:09.700 --> 00:11:11.640 across some discrete linear sequence,

00:11:11.640 --> 00:11:14.880 like this series of zeros and ones you see below.

00:11:14.880 --> 00:11:16.970 There's one, zero, zeros, there's one, zero, ones,

00:11:16.970 --> 00:11:19.920 there's periods where ones are closer to each other a series

00:11:19.920 --> 00:11:22.440 of ones are closer or farther apart from each other.

00:11:22.440 --> 00:11:24.220 How should we figure out whether things

00:11:24.220 --> 00:11:25.590 are actually clustered in that site?

00:11:25.590 --> 00:11:26.930 Or are they random?

00:11:26.930 --> 00:11:30.680 So if you don't know exactly where to describe,

00:11:30.680 --> 00:11:33.050 or what size you're looking for,

00:11:33.050 --> 00:11:34.700 the most common solution people use

00:11:34.700 --> 00:11:36.330 is some kind of sliding window,

00:11:36.330 --> 00:11:38.310 they take a window over the series,

00:11:38.310 --> 00:11:40.257 and they slide it across and say,

00:11:40.257 --> 00:11:41.480 "How many are in this window?"

00:11:41.480 --> 00:11:44.100 And then you can come up with based on the sliding window

00:11:44.100 --> 00:11:45.835 a sort of diagram of the clustering.

00:11:45.835 --> 00:11:49.450 And that's an approach that actually does

00:11:49.450 --> 00:11:51.470 give a good metric of the clustering

00:11:51.470 --> 00:11:53.280 in terms of like you see peaks where there's

00:11:53.280 --> 00:11:55.740 a lot of clustering and valleys where there is none.

00:11:55.740 --> 00:11:59.022 However, significance testing with that kind of approach

00:11:59.022 --> 00:12:00.150 is often awkward to construct.

00:12:00.150 --> 00:12:02.400 Due to a strong or autocorrelation

00:12:02.400 --> 00:12:04.490 among this URL overlapping windows.

00:12:04.490 --> 00:12:05.610 And of course, if you just sort of

00:12:05.610 --> 00:12:09.070 take windows arbitrarily from one location to another,

00:12:09.070 --> 00:12:12.756 then you're really instituting, (indistinct chatter)

00:12:12.756 --> 00:12:14.364 then that causes problems.

00:12:14.364 --> 00:12:16.140 Because what if the cluster is really on a border

00:12:16.140 --> 00:12:19.205 between two windows, so you have to slide it over and then

00:12:19.205 --> 00:12:20.040 you have the autocorrelation.

00:12:20.040 --> 00:12:21.440 And it becomes actually statistically

00:12:21.440 --> 00:12:23.990 quite challenging to sort of account

00:12:23.990 --> 00:12:25.410 for all of those auto correlations.

00:12:25.410 --> 00:12:27.310 Secondly, they need to specify that window

00:12:27.310 --> 00:12:30.610 size itself presents a user with a procedural ambiguity

00:12:30.610 --> 00:12:33.790 that almost inevitably leads to post hoc selection of window

00:12:33.790 --> 00:12:37.010 size and can mislead inference that is just the fact that

00:12:37.010 --> 00:12:39.030 you have to choose a window size.

00:12:39.030 --> 00:12:41.070 And if you don't actually have a good arbitrary

00:12:41.070 --> 00:12:42.570 outside reason to choose it.

00:12:42.570 --> 00:12:44.480 It's very hard not to choose a window size

00:12:44.480 --> 00:12:48.830 that ends up validating your hypothesis in some way.

00:12:48.830 --> 00:12:50.680 So it'd be better if we could just have an approach

00:12:50.680 --> 00:12:52.980 that does not require us to place in some

00:12:52.980 --> 00:12:55.760 arbitrary parameter that gives us a window size.

00:12:55.760 --> 00:12:57.680 So in order to address this question,

00:12:57.680 --> 00:13:00.710 a postdoc of mine, John John, who you see below work

00:13:00.710 --> 00:13:02.610 with me to address it.

00:13:02.610 --> 00:13:03.950 Oh, I wanted to say one other thing,

00:13:03.950 --> 00:13:07.390 which is that, yes, this has been addressed with some

00:13:07.390 --> 00:13:09.840 nonparametric methods that people have developed,

00:13:10.750 --> 00:13:14.270 including some rather famous people like Sam Carlin.

00:13:14.270 --> 00:13:17.360 And these are methods that do not assume prior knowledge.

00:13:17.360 --> 00:13:19.690 And they've been suggested to detect this clustering

00:13:19.690 --> 00:13:20.860 and discrete linear sequences.

00:13:20.860 --> 00:13:22.420 So you can do runs tests that look for

00:13:22.420 --> 00:13:25.700 the longest unbroken run, or the variance of the run

00:13:25.700 --> 00:13:27.290 links across the entire sequence.

00:13:27.290 --> 00:13:29.640 Both of these are indicators of clustering.

00:13:29.640 --> 00:13:32.170 Unfortunately, both of those are using

00:13:32.170 --> 00:13:34.110 are not sufficient tests.

00:13:34.110 --> 00:13:36.290 And those they don't use enough of the information

00:13:36.290 --> 00:13:38.860 to say that you're actually have as much power as you'd

00:13:38.860 --> 00:13:40.080 like to do the analysis.

00:13:40.080 --> 00:13:41.730 And that's because if you use like

00:13:41.730 --> 00:13:43.700 the longest run link, for instance, of course,

00:13:43.700 --> 00:13:45.200 you're only really using a little bit

00:13:45.200 --> 00:13:47.260 of information about the entire sequence.

00:13:47.260 --> 00:13:49.450 And of course, you're really missing anything

00:13:49.450 --> 00:13:52.340 like the cluster of ones that are have a bunch of small

00:13:52.340 --> 00:13:54.200 clusters that are all next to each other interspersed

00:13:54.200 --> 00:13:55.710 with a few of the other type,

00:13:55.710 --> 00:13:58.740 so the longest unbroken run doesn't work well.

00:13:58.740 --> 00:14:00.970 If you use the In terms of power,

00:14:00.970 --> 00:14:03.701 if you use the variance of long run link

00:14:03.701 --> 00:14:05.160 that gets rid of the fact that you're looking for just one.

00:14:05.160 --> 00:14:07.440 But unfortunately, a variance doesn't tell you anything

00:14:07.440 --> 00:14:09.290 about the relative position of site

00:14:11.102 --> 00:14:14.060 that are of the same type across the sequence.

00:14:14.060 --> 00:14:17.535 So the fact that this one, one, one, one here is close

00:14:17.535 --> 00:14:19.828 to the one, one here, and the one another is,

00:14:19.828 --> 00:14:22.335 and this the fact that these are all close to each other,

00:14:22.335 --> 00:14:25.210 does not give us the power that it should

00:14:25.210 --> 00:14:26.590 for understanding this region may

00:14:26.590 --> 00:14:30.250 be under maybe cluster.

00:14:30.250 --> 00:14:33.210 So variants of run length is also an underpowered approach.

00:14:33.210 --> 00:14:36.170 The most powerful approach that's been published out there,

00:14:36.170 --> 00:14:38.140 aside from the ones we've been working on,

00:14:38.140 --> 00:14:40.620 is the empirical cumulative distribution functions

00:14:40.620 --> 00:14:43.410 to sick that's where you sort of go across the sequence

00:14:43.410 --> 00:14:46.728 and just say, "oh, okay, we're accumulating ones here,

00:14:46.728 --> 00:14:47.561 we're shooting more accumulating more."

00:14:48.873 --> 00:14:49.830 And there's fortunately a number

00:14:51.502 --> 00:14:53.153 of highly developed statistical approaches

00:14:53.153 --> 00:14:55.400 to look at the empirical distribution and figure

00:14:55.400 --> 00:15:00.030 out whether you see an increase beyond

00:15:00.030 --> 00:15:02.950 expected during some period during that ECDF,

00:15:02.950 --> 00:15:04.950 the power is better than either the previous methods,

00:15:04.950 --> 00:15:06.700 but it's still not very strong.

00:15:06.700 --> 00:15:08.340 It's not clear that it includes all the

00:15:08.340 --> 00:15:10.180 information that it should.

00:15:10.180 --> 00:15:11.756 And it can be affected.

00:15:11.756 --> 00:15:13.730 Research has shown that it can be affected

00:15:13.730 --> 00:15:16.060 by the location of the cluster, which is not desirable.

00:15:16.060 --> 00:15:17.930 So if you have a cluster on an end,

00:15:17.930 --> 00:15:20.640 that has less the ECDF will have less power

00:15:20.640 --> 00:15:23.320 or more power compared to a cluster in the middle.

00:15:23.320 --> 00:15:26.300 It's also challenging to interpret in the end,

00:15:26.300 --> 00:15:28.830 for reasons I'm not gonna go into right away.

00:15:28.830 --> 00:15:29.970 So what did we do?

00:15:29.970 --> 00:15:32.420 What we did was develop a tripartite divide

00:15:32.420 --> 00:15:34.920 and conquer approach to model variant sites

00:15:34.920 --> 00:15:36.930 based on iterative sub clustering.

00:15:36.930 --> 00:15:38.820 And I'll describe it in detail right now.

00:15:38.820 --> 00:15:40.370 I'll just tell you the plus and the minus

00:15:40.370 --> 00:15:42.150 of this approach at the beginning,

00:15:42.150 --> 00:15:44.620 which is it's sort of a bioinformatics approach

00:15:44.620 --> 00:15:47.930 and that are bioinformatics statisticians approach

00:15:47.930 --> 00:15:50.380 and that it uses intensive computation

00:15:50.380 --> 00:15:52.480 to solve the problem instead of giving

00:15:52.480 --> 00:15:54.373 a strict analytical result.

00:15:55.409 --> 00:15:57.810 And in fact, what it does is it just says,

00:15:57.810 --> 00:16:00.160 Well, if we're interested in clustering in any case,

00:16:00.160 --> 00:16:03.226 clusters should be represented by increases in

00:16:03.226 --> 00:16:05.680 the probability within some cluster central region

00:16:05.680 --> 00:16:08.310 compared to some side regions.

00:16:08.310 --> 00:16:10.810 And if we define CS and CE to be anything

00:16:10.810 --> 00:16:13.600 from the very beginning to the very end of the sequence,

00:16:13.600 --> 00:16:16.700 it encompasses all possible single clusters

00:16:16.700 --> 00:16:19.404 within a sequence.

00:16:19.404 --> 00:16:22.360 So, for instance, if the cluster were on the far left

00:16:22.360 --> 00:16:24.600 we can just define CS to be at zero,

00:16:24.600 --> 00:16:28.220 the left hand cluster is nothing and the right hand cluster,

00:16:28.220 --> 00:16:33.220 right hand area that has depressed in variant type intensity

00:16:35.220 --> 00:16:38.240 would be the other category.

00:16:38.240 --> 00:16:41.600 Anyway, so, what we can do is divide any sequence

00:16:41.600 --> 00:16:43.890 into three sections, just count up the number

00:16:43.890 --> 00:16:46.460 of site types in each one, estimate the maximum

00:16:46.460 --> 00:16:50.040 likelihood probability for the site type

00:16:50.040 --> 00:16:51.970 to be of the variant type of interest,

00:16:51.970 --> 00:16:54.900 say it's a glycine amino acids within a protein

00:16:54.900 --> 00:16:59.900 or add mean nucleotides limited gene, whatever it is.

00:16:59.960 --> 00:17:02.580 So then you can just come up with a null hypothesis,

00:17:02.580 --> 00:17:06.060 which is the likelihood under the hypothesis

00:17:06.060 --> 00:17:09.490 that these things are located at random

00:17:09.490 --> 00:17:11.320 across the whole sequence.

00:17:11.320 --> 00:17:13.660 And then an alternate hypothesis that allows

00:17:13.660 --> 00:17:17.520 that is invoking a model which involves more parameters,

00:17:17.520 --> 00:17:20.990 which then separate separates into a clustered

00:17:20.990 --> 00:17:22.890 versus non-clustered state.

00:17:22.890 --> 00:17:24.600 So that would be fine if what we really

00:17:24.600 --> 00:17:26.944 expected in a sequence was one cluster,

00:17:26.944 --> 00:17:29.094 compared to nothing else,

00:17:29.094 --> 00:17:33.120 compared to the sort of baseline rate of clustering,

00:17:33.120 --> 00:17:35.414 sort of baseline rate of variant types.

00:17:35.414 --> 00:17:39.040 And but what we really want is an approach

00:17:39.040 --> 00:17:41.590 that can take clustering at many, many levels.

00:17:41.590 --> 00:17:43.470 So what if there's a cluster within the cluster

00:17:43.470 --> 00:17:44.780 or cluster within left?

00:17:44.780 --> 00:17:46.450 So what you can do is then take each

00:17:46.450 --> 00:17:49.680 of these sub clusters you've identified and actually

00:17:49.680 --> 00:17:52.560 do the same process on them looking for whether there's

00:17:52.560 --> 00:17:56.030 a higher likelihood of the data given another cluster

00:17:56.030 --> 00:17:59.358 somewhere within this sequence, et cetera, et cetera.

00:17:59.358 --> 00:18:03.730 Now, if you think so this sort of dictates a procedure,

00:18:03.730 --> 00:18:06.890 which is that you start, you input the sequence,

00:18:06.890 --> 00:18:08.900 you start at, you know, the first at

00:18:08.900 --> 00:18:10.770 the left and move all the way to the right,

00:18:10.770 --> 00:18:13.200 essentially, you find the most likely cluster

00:18:13.200 --> 00:18:15.110 among all the possible clusters.

00:18:15.110 --> 00:18:17.200 If the cluster is statistically significant,

00:18:17.200 --> 00:18:20.920 you then sub sequence each of those three parts,

00:18:20.920 --> 00:18:23.730 the left hand part, the central center part

00:18:23.730 --> 00:18:25.870 and the right hand part, find the most

00:18:25.870 --> 00:18:27.480 likely clusters within each of them.

00:18:27.480 --> 00:18:29.560 And proceed doing this until you reach a point

00:18:29.560 --> 00:18:31.830 where you can no longer find any statistical evidence

00:18:31.830 --> 00:18:33.760 that there is continued clustering within it.

00:18:33.760 --> 00:18:35.600 And that's the point at which you stop.

00:18:35.600 --> 00:18:36.670 And then what you can do.

00:18:36.670 --> 00:18:38.500 And this, I think, is sort of a key because

00:18:38.500 --> 00:18:41.780 at the end of that, what you get is one discrete diagram,

00:18:41.780 --> 00:18:43.520 kind of like that diagram I showed you initially,

00:18:43.520 --> 00:18:45.750 where it proceeds flat, goes up,

00:18:45.750 --> 00:18:47.243 proceeds flat goes down, et cetera.

00:18:47.243 --> 00:18:49.890 I'll show you an example of that in a moment.

00:18:49.890 --> 00:18:52.835 But what you really wanna do possibly,

00:18:52.835 --> 00:18:54.795 right, what I think is really appealing about

00:18:54.795 --> 00:18:55.760 this approach is that then you can take

00:18:55.760 --> 00:18:58.720 that as one model, the most likely model and you can look

00:18:58.720 --> 00:19:00.290 at all the other possible models

00:19:00.290 --> 00:19:01.660 that you could have constructed.

00:19:01.660 --> 00:19:04.730 And you can use AIC weighting to actually figure

00:19:04.730 --> 00:19:09.730 out how much you should believe what is the weight

00:19:11.375 --> 00:19:13.039 for every possible model.

00:19:13.039 --> 00:19:14.470 And then you can average across those models

00:19:14.470 --> 00:19:16.742 to give you a continuous description

00:19:16.742 --> 00:19:18.180 of how much clustering you see across the sequence.

00:19:18.180 --> 00:19:20.430 And again, the advantage that I mentioned

00:19:20.430 --> 00:19:21.530 early on about this,

00:19:21.530 --> 00:19:23.870 from my standpoint is I haven't put in anything

00:19:23.870 --> 00:19:26.350 about how big a window how big a cluster,

00:19:26.350 --> 00:19:28.300 I put in nothing about what I'm expecting

00:19:28.300 --> 00:19:29.610 to see out of the sequence.

00:19:29.610 --> 00:19:32.220 I'm just asking, what's the most likely description

00:19:32.220 --> 00:19:36.560 of this given the assay penalty for parameterization

00:19:36.560 --> 00:19:38.940 and what the result gives me.

00:19:38.940 --> 00:19:41.400 So then we have a bunch of different weights

00:19:41.400 --> 00:19:43.003 for all our different models.

00:19:44.251 --> 00:19:45.250 And what it gives us something like this.

00:19:45.250 --> 00:19:47.820 So on the top, I've shown you the AIC model selection

00:19:47.820 --> 00:19:48.900 which is the first thing I showed you

00:19:48.900 --> 00:19:51.420 if I just took the most likely description

00:19:51.420 --> 00:19:52.890 of this particular sequence.

00:19:52.890 --> 00:19:54.820 It's not important what it is it's PRF

00:19:54.820 --> 00:19:59.430 ADHD, which has been widely studied in evolutionary biology.

00:19:59.430 --> 00:20:02.420 But if you take this model selection would,

00:20:02.420 --> 00:20:04.610 the most likely description

00:20:04.610 --> 00:20:06.670 given that sub clustering looks something like this

00:20:06.670 --> 00:20:09.660 where we have a region with fairly high concentration

00:20:09.660 --> 00:20:13.730 of polymorphism, in this case, a valley,

00:20:13.730 --> 00:20:15.700 a region, an intermediate level,

00:20:15.700 --> 00:20:18.520 a point where we have a lot of polymorphism.

00:20:18.520 --> 00:20:21.260 And then it moves and changes across the sequence.

00:20:21.260 --> 00:20:24.700 Now, if you then instead take not just that one model,

00:20:24.700 --> 00:20:27.500 but a series of models and do the AIC model average,

00:20:27.500 --> 00:20:29.750 you get a much more continuous description across

00:20:29.750 --> 00:20:32.790 the sequence of what the probability

00:20:32.790 --> 00:20:34.983 of sight types being different is.

00:20:35.845 --> 00:20:37.280 And that enables us to ask a question

00:20:37.280 --> 00:20:41.050 that's a little bit more interesting in many cases,

00:20:41.050 --> 00:20:43.080 and I'll show you how it enables us to ask questions

00:20:43.080 --> 00:20:45.400 about natural selection in a moment.

00:20:45.400 --> 00:20:47.900 So in particular, it allows us to get an estimate,

00:20:48.975 --> 00:20:50.353 you know of what the probability

00:20:50.353 --> 00:20:51.186 is across the entire sequence.

00:20:51.186 --> 00:20:52.310 Even though we don't have

00:20:52.310 --> 00:20:54.480 observations within the central region

00:20:54.480 --> 00:20:56.420 or this barren region here.

00:20:56.420 --> 00:20:59.600 We can still estimate what the model average,

00:20:59.600 --> 00:21:02.130 probably of a change of hearing in different places

00:21:02.130 --> 00:21:04.590 have this gene are and that enables us

00:21:04.590 --> 00:21:07.640 to ask questions that we otherwise could not do.

00:21:07.640 --> 00:21:11.160 All right, so that's an introduction of MACML.

00:21:11.160 --> 00:21:14.010 I'll just mention, and I could give you more detail on this.

00:21:14.010 --> 00:21:16.010 It's like this is actually published work,

00:21:16.010 --> 00:21:17.220 so you can find it.

00:21:17.220 --> 00:21:19.080 But compared to the ECDF statistics,

00:21:19.080 --> 00:21:21.140 that approach I just showed you has greater power

00:21:21.140 --> 00:21:23.090 to detect heterogeneous clusters

00:21:23.090 --> 00:21:25.710 it identifies clusters with greater accuracy and precision

00:21:25.710 --> 00:21:28.410 based on the Kullback-Liebler divergence between

00:21:28.410 --> 00:21:31.450 the actual distribution of the observed distribution,

00:21:31.450 --> 00:21:32.950 sorry, the actual distribution

00:21:34.201 --> 00:21:35.615 and the inferred distribution.

00:21:35.615 --> 00:21:36.610 It has better power and accuracy across

00:21:36.610 --> 00:21:37.920 different levels of clustering,

00:21:37.920 --> 00:21:39.520 better power and accuracy across

00:21:40.357 --> 00:21:41.315 different sequence links,

00:21:41.315 --> 00:21:43.071 and better power and accuracy and finding

00:21:43.071 --> 00:21:44.540 multiple clusters compared to a single cluster.

00:21:44.540 --> 00:21:46.560 The disadvantage is, it's extraordinarily

00:21:46.560 --> 00:21:49.160 computationally intensive, and it is prohibitively

00:21:49.160 --> 00:21:50.720 so for very long sequences.

00:21:50.720 --> 00:21:53.160 So for genes a very long length,

00:21:53.160 --> 00:21:55.210 we can't actually run it on the full-length gene

00:21:55.210 --> 00:21:58.270 and we have to do some more heuristic processes

00:21:58.270 --> 00:22:00.620 to crunch those genes into smaller size.

00:22:00.620 --> 00:22:02.820 Which we then can analyze and then build them up.

00:22:02.820 --> 00:22:04.880 Again, I won't go into those at the moment.

00:22:04.880 --> 00:22:07.100 But the point is that at certain links,

00:22:07.100 --> 00:22:09.430 it gets just computationally too intensive to go

00:22:09.430 --> 00:22:12.909 through all the possible models that could explain the data.

00:22:12.909 --> 00:22:17.030 Now, I've talked about the maximum-likelihood averaging

00:22:17.030 --> 00:22:18.890 to profile clustering of site types

00:22:18.890 --> 00:22:21.210 across discrete linear sequences,

00:22:21.210 --> 00:22:24.030 introduced that methodology to now I'm gonna talk about

00:22:24.030 --> 00:22:26.200 how we can at apply that methodology

00:22:26.200 --> 00:22:29.250 to get us a better idea of which sites are under selection

00:22:29.250 --> 00:22:32.120 using a what's called a pause on random fields approach.

00:22:32.120 --> 00:22:33.980 And don't worry about that terminology.

00:22:33.980 --> 00:22:37.170 You might know it from statistics,

00:22:37.170 --> 00:22:39.700 it has to do with a particular observation

00:22:39.700 --> 00:22:42.078 in molecular evolutionary biology,

00:22:42.078 --> 00:22:42.911 which is why they're using it

00:22:44.433 --> 00:22:45.530 and it's not really important for this talk,

00:22:45.530 --> 00:22:46.740 why it's called that.

00:22:48.385 --> 00:22:51.110 So let's go on and go ahead and do that talk

00:22:51.110 --> 00:22:53.155 about the model-averaged site selection

00:22:53.155 --> 00:22:54.377 using Poisson random fields.

00:22:54.377 --> 00:22:56.383 So first, I need to give you a little bit of background

00:22:56.383 --> 00:22:57.620 in the evolutionary biology for those of you

00:22:59.071 --> 00:23:00.465 who haven't had a lot of biology,

00:23:00.465 --> 00:23:01.570 so you understand how this fits in with

00:23:01.570 --> 00:23:03.020 what we tend to do another strategy.

00:23:03.020 --> 00:23:04.906 Of course, evolutionary biologists

00:23:04.906 --> 00:23:05.960 are often very interested in understanding

00:23:05.960 --> 00:23:07.190 what things are under selection.

00:23:07.190 --> 00:23:08.730 And in the context of this talk,

00:23:08.730 --> 00:23:09.860 why is that important?

00:23:09.860 --> 00:23:12.035 Well, we'd really like to know what things

00:23:12.035 --> 00:23:13.800 are under selection in the COVID epidemic,

00:23:13.800 --> 00:23:15.860 because we'd like to know what sites

00:23:15.860 --> 00:23:17.760 are actually causing the COVID epidemic

00:23:17.760 --> 00:23:21.380 to spread more or not, and what sites may have

00:23:21.380 --> 00:23:23.580 been important in it prior to zoonosis,

00:23:23.580 --> 00:23:26.270 MSN, perhaps, especially in the context of this talk,

00:23:26.270 --> 00:23:27.660 what sites were selected during

00:23:27.660 --> 00:23:30.610 that zoonotic process that made this virus perhaps able

00:23:30.610 --> 00:23:32.590 to infect humans in the first place.

00:23:32.590 --> 00:23:34.312 So what we're doing is,

00:23:34.312 --> 00:23:36.080 so to give you an introduction,

00:23:36.080 --> 00:23:38.560 I just wanna mention that they're sort of ways

00:23:38.560 --> 00:23:40.270 to look at ancient times and understand

00:23:40.270 --> 00:23:41.890 whether selection was happening.

00:23:41.890 --> 00:23:44.145 And that's this approach that's called

00:23:44.145 --> 00:23:45.080 that looks at phylogenetic divergence,

00:23:45.080 --> 00:23:47.397 looking at multiple sites and saying,

00:23:47.397 --> 00:23:49.340 "Oh, we have a whole bunch of phylogeny

00:23:49.340 --> 00:23:51.070 of how these organisms are related."

00:23:51.070 --> 00:23:54.910 And then we have a bunch of sites that are for each taxon.

00:23:54.910 --> 00:23:56.700 When we see sites like this, for instance,

00:23:56.700 --> 00:23:59.660 that's having A and then a couple C's and then a G

00:23:59.660 --> 00:24:02.870 and another tacks on, we know that this site changed twice

00:24:02.870 --> 00:24:04.690 on that phylogeny, at least right?

00:24:04.690 --> 00:24:08.770 So it changed to probably change from C ancestrally

00:24:08.770 --> 00:24:11.460 to an A in this lineage and to a G

00:24:11.460 --> 00:24:13.060 in this lineage independently.

00:24:13.060 --> 00:24:15.510 And so the fact that it changed twice means

00:24:15.510 --> 00:24:18.210 that it's got an elevated rate of change.

00:24:18.210 --> 00:24:19.500 And that elevated rate of change is an indication

00:24:19.500 --> 00:24:21.810 that there's been positive selection for change.

00:24:21.810 --> 00:24:24.920 It's especially likely in sort of pathogen hosts

00:24:24.920 --> 00:24:27.690 interactions that high rates of high change are

00:24:27.690 --> 00:24:30.124 because pathogens are changing in order

00:24:30.124 --> 00:24:32.590 to not be recognizable by their hosts.

00:24:32.590 --> 00:24:34.510 And often the host has recognition proteins

00:24:34.510 --> 00:24:36.470 that are changing to still recognize the pathogen,

00:24:36.470 --> 00:24:38.040 even the pathogen is changing.

00:24:38.040 --> 00:24:39.560 So these high rates of evolution

00:24:39.560 --> 00:24:41.788 are very strong indicators of selection

00:24:41.788 --> 00:24:44.880 in host pathogen situations.

00:24:44.880 --> 00:24:48.460 So this is one way to study a natural selection.

00:24:48.460 --> 00:24:52.030 It does depend, though, on having a lot of data going back

00:24:52.030 --> 00:24:54.630 in time because you're actually reliant on these changes

00:24:54.630 --> 00:24:57.820 are occurring in multiple places on multiple lineages.

00:24:57.820 --> 00:25:02.230 Now, a more recent level, and I'm going to go back

00:25:02.230 --> 00:25:03.530 to the middle in a moment.

00:25:04.837 --> 00:25:05.740 But a very recent time, you may have

00:25:06.648 --> 00:25:08.294 heard of selective sweep detection,

00:25:08.294 --> 00:25:10.812 a couple of methods people use are tajima's D,

00:25:10.812 --> 00:25:13.700 or IHS, there's a bunch of other methods that are out now.

00:25:13.700 --> 00:25:16.100 And the idea there is to look at polymorphism.

00:25:16.100 --> 00:25:19.550 And if you look at an individual, before selection,

00:25:19.550 --> 00:25:21.540 this is sort of just a idea diagram,

00:25:21.540 --> 00:25:22.840 not what you look at.

00:25:22.840 --> 00:25:26.380 But so if you look at an individual who has a variant,

00:25:26.380 --> 00:25:30.110 and what you see in a population is that

00:25:30.110 --> 00:25:33.290 one individual with variant, a variant that's important

00:25:33.290 --> 00:25:35.380 as somehow swept across the population.

00:25:35.380 --> 00:25:37.240 So if you see this would be before selection,

00:25:37.240 --> 00:25:39.280 there's a lot of variation at a particular locus

00:25:39.280 --> 00:25:41.410 in the genome after selection,

00:25:41.410 --> 00:25:44.255 that one individuals variant which contributed

00:25:44.255 --> 00:25:46.430 to the reproductive fitness would then imply

00:25:46.430 --> 00:25:50.310 that they would spread across the population.

00:25:50.310 --> 00:25:51.950 And if they spread across the population,

00:25:51.950 --> 00:25:53.980 then the genetic variants that were present

00:25:53.980 --> 00:25:56.210 in that original individual spread across

00:25:56.210 --> 00:25:59.700 the population as well along with this selected site,

00:25:59.700 --> 00:26:03.820 and so you can look for this kind of partial or speedy.

00:26:03.820 --> 00:26:07.469 And the selection is going on neither

00:26:07.469 --> 00:26:08.991 of the approaches that I just talked about

00:26:08.991 --> 00:26:09.890 or the approach that I'm doing today.

00:26:09.890 --> 00:26:12.036 So I just wanted to introduce those,

00:26:12.036 --> 00:26:12.869 so you knew those are different.

00:26:12.869 --> 00:26:15.299 And they're different because we're looking

00:26:15.299 --> 00:26:16.495 at a more intermediate timescale.

00:26:16.495 --> 00:26:18.790 That's like the sweet detection is purely

00:26:18.790 --> 00:26:20.880 dependent on polymorphism in the population,

00:26:20.880 --> 00:26:23.720 like what's happening in a population right now.

00:26:23.720 --> 00:26:25.720 The phylogenetic divergence is purely dependent

00:26:25.720 --> 00:26:28.400 on this ancient changes that you get from a phylogeny

00:26:28.400 --> 00:26:31.409 understanding how different species are related

00:26:31.409 --> 00:26:33.010 to each other at an intermediate level,

00:26:33.010 --> 00:26:35.487 our methods use that use both the polymorphism

00:26:35.487 --> 00:26:37.260 and the divergence.

00:26:37.260 --> 00:26:39.990 And the idea here in the McDonald-Kreitman approach,

00:26:39.990 --> 00:26:41.980 and the master approach I'm going to tell you

00:26:41.980 --> 00:26:45.600 about is that the polymorphism what you see generally

00:26:45.600 --> 00:26:48.298 in the population is sort of consistent with this.

00:26:48.298 --> 00:26:51.240 Sorry, if I go back to this slide.

00:26:51.240 --> 00:26:53.420 With this before selection, you know,

00:26:53.420 --> 00:26:54.970 all of these blue sites are assumed

00:26:54.970 --> 00:26:56.510 to not be under selection,

00:26:56.510 --> 00:26:59.290 and that generally what we believe in evolutionary biology,

00:26:59.290 --> 00:27:01.960 because of empirical data that validates it

00:27:01.960 --> 00:27:05.220 is that most sites that you find varying in populations

00:27:05.220 --> 00:27:06.640 are not under strong selection.

00:27:06.640 --> 00:27:07.930 If they were on stronger selection,

00:27:07.930 --> 00:27:10.273 they would probably fix it, everyone would have them.

00:27:11.441 --> 00:27:13.116 And if they were under negative selection,

00:27:13.116 --> 00:27:13.949 they wouldn't rise to a high frequency.

00:27:13.949 --> 00:27:16.706 So generally speaking sites that you actually see

00:27:16.706 --> 00:27:18.330 change differences between us and our genetics

00:27:18.330 --> 00:27:20.170 typically are not affecting anything.

00:27:20.170 --> 00:27:22.584 Of course, we spend in our...

00:27:22.584 --> 00:27:23.850 In the media, you only hear about the changes

00:27:23.850 --> 00:27:25.060 that actually affect things.

00:27:25.060 --> 00:27:26.470 And that's because those are important to us,

00:27:26.470 --> 00:27:28.429 the ones that don't change anything

00:27:28.429 --> 00:27:29.417 we don't really care about.

00:27:29.417 --> 00:27:30.250 So nobody talks about that much.

00:27:30.250 --> 00:27:32.750 But most of the changes within population or differences

00:27:32.750 --> 00:27:35.175 within population don't have much material effect.

00:27:35.175 --> 00:27:37.100 So under that hypothesis,

00:27:37.100 --> 00:27:38.960 then when you look at polymorphism,

00:27:38.960 --> 00:27:41.240 most polymorphism is just an indication

00:27:41.240 --> 00:27:42.760 of the underlying mutation rate,

00:27:42.760 --> 00:27:44.970 some mutation happened didn't have any effect.

00:27:44.970 --> 00:27:47.410 It's drifting up and down in the population.

00:27:47.410 --> 00:27:49.810 And so the advantage of that is if you know

00:27:49.810 --> 00:27:52.040 that polymorphism is signal is a signature

00:27:52.040 --> 00:27:53.966 of just random mutation, it gives us an estimate

00:27:53.966 --> 00:27:57.160 of the underlying mutation rate, which we can then compare

00:27:57.160 --> 00:27:59.610 to the divergence and using that comparison,

00:27:59.610 --> 00:28:02.350 we can understand how organisms are related.

00:28:02.350 --> 00:28:05.207 So whether organisms are under selection

00:28:05.207 --> 00:28:07.104 or not, if the divergence is high compared

00:28:07.104 --> 00:28:08.940 to the polymorphism, that indicates a lot of selection.

00:28:08.940 --> 00:28:12.211 That means (indistinct chatter)

00:28:12.211 --> 00:28:14.180 in the timescale of the analysis you're doing,

00:28:14.180 --> 00:28:17.280 we have a lot of change the population,

00:28:17.280 --> 00:28:19.520 and on the other hand, you have a lot of polymorphism

00:28:19.520 --> 00:28:22.100 and not that much divergence, then that indicates

00:28:22.100 --> 00:28:23.350 you've got a lot of change going on,

00:28:23.350 --> 00:28:25.809 but it's not actually being directionally

00:28:25.809 --> 00:28:27.340 selected because the divergence is much lower.

00:28:27.340 --> 00:28:29.640 So how does that test work in practice?

00:28:29.640 --> 00:28:31.820 Well, just to step back for one moment,

00:28:31.820 --> 00:28:33.770 so we're gonna apply that kind of test.

00:28:34.664 --> 00:28:36.210 In this talk I'm applying that test

00:28:36.210 --> 00:28:39.450 to the emergence of COVID-19.

00:28:39.450 --> 00:28:43.600 I'm actually applying it but also to SARS, which is fairly

00:28:43.600 --> 00:28:46.170 closely related the SARS coronavirus one

00:28:46.170 --> 00:28:48.040 because we have similar data and can apply

00:28:48.040 --> 00:28:51.820 the same test in the same way to that data set.

00:28:51.820 --> 00:28:54.250 And we're using in addition the SARS like

00:28:55.340 --> 00:28:57.870 Coronavirus in a sample that had been sequence

00:28:57.870 --> 00:28:59.870 basically collected from bats.

00:28:59.870 --> 00:29:01.930 Over the past 20 years or so,

00:29:01.930 --> 00:29:05.199 so what you can see here is a phylogeny,

00:29:05.199 --> 00:29:09.160 which includes COVID-19 epidemic ongoing now in humans,

00:29:09.160 --> 00:29:12.790 the SARS epidemic, which caused some 400 deaths

00:29:12.790 --> 00:29:17.610 or so back in the early 2000s.

00:29:17.610 --> 00:29:21.260 And what we're doing is analyzing both and looking at,

00:29:21.260 --> 00:29:24.890 in particular, the very short internode here

00:29:24.890 --> 00:29:29.890 were between the most closely related non human infections

00:29:30.950 --> 00:29:33.200 and the human infection set that we can see.

00:29:33.200 --> 00:29:36.040 And this internode here, also,

00:29:36.040 --> 00:29:39.040 between these non human infections and the human

00:29:39.040 --> 00:29:41.770 infections we can see here, because the changes

00:29:41.770 --> 00:29:45.010 that may have enabled, we don't know,

00:29:45.010 --> 00:29:47.230 there may be no changes that enabled it,

00:29:47.230 --> 00:29:48.780 maybe this virus throughout

00:29:48.780 --> 00:29:50.620 its entire history could have infected humans,

00:29:50.620 --> 00:29:53.420 but it just never managed to or never did.

00:29:53.420 --> 00:29:55.970 But if there are changes that are unique to this virus

00:29:55.970 --> 00:29:58.890 that happened during zoonosis, enabling it to infect us,

00:29:58.890 --> 00:30:00.430 they happened on this lineage,

00:30:00.430 --> 00:30:03.280 and so we're interested in seeing what those changes are.

00:30:04.200 --> 00:30:06.100 And so that's what we're gonna do is we're gonna run

00:30:06.100 --> 00:30:10.030 this polymorphism and divergence approach on this lineage.

00:30:10.030 --> 00:30:13.190 And what I just want to make (indistinct chatter)

00:30:13.190 --> 00:30:14.390 clear to you is the reason

00:30:14.390 --> 00:30:17.510 why the polymorphism divergence approach is important is

00:30:17.510 --> 00:30:20.482 the phylogenetic approach, the ancient approach

00:30:20.482 --> 00:30:22.180 relies on a large clade of data, which we don't have

00:30:22.180 --> 00:30:24.248 for that particular lineage here,

00:30:24.248 --> 00:30:25.600 we just have the human infection,

00:30:25.600 --> 00:30:26.433 which is no longer zoonotic.

00:30:26.433 --> 00:30:27.500 And we have this one lineage.

00:30:27.500 --> 00:30:29.890 And so what we can do is ancestrally reconstruct

00:30:29.890 --> 00:30:32.710 the ancestor of this lineage, which is right here,

00:30:32.710 --> 00:30:34.190 actually on the phylogeny,

00:30:34.190 --> 00:30:36.700 and also the ancestor right here,

00:30:36.700 --> 00:30:40.090 and then use mass PRF, this approach that's based

00:30:40.090 --> 00:30:42.600 on polymorphism in the room, so I'll explain to you

00:30:42.600 --> 00:30:45.560 on the divergence between that ancestor

00:30:45.560 --> 00:30:48.390 and the first ancestor of all the human infections.

00:30:48.390 --> 00:30:51.050 And we can take that as the near zoonosis time

00:30:51.050 --> 00:30:52.620 and figure out what mutations might

00:30:52.620 --> 00:30:54.290 have happened during that time.

00:30:54.290 --> 00:30:56.410 All right, so we're gonna do that in both

00:30:56.410 --> 00:30:58.163 the COVID-19 and SARS cases.

00:30:59.130 --> 00:31:01.620 Now, how does this work in principle?

00:31:01.620 --> 00:31:02.660 Well, there's an old approach,

00:31:02.660 --> 00:31:04.590 which is not what we're using.

00:31:04.590 --> 00:31:05.960 But I have to compare it to in order to

00:31:05.960 --> 00:31:08.653 sort of reference it in terms of the literature.

00:31:09.490 --> 00:31:11.480 And that is that when you assume

00:31:11.480 --> 00:31:13.480 that polymorphism is neutral,

00:31:13.480 --> 00:31:15.530 we expect a different proportion of replacement

00:31:15.530 --> 00:31:18.070 to synonymous divergence compared to replacement

00:31:18.070 --> 00:31:21.150 to synonymous polymorphism in a gene.

00:31:21.150 --> 00:31:23.450 So it's just a two by two table here, again,

00:31:23.450 --> 00:31:25.360 very simple statistics, where we look at

00:31:25.360 --> 00:31:27.730 the number of replacement sites that are divergent

00:31:27.730 --> 00:31:30.113 the number of synonymous sites replacement,

00:31:30.113 --> 00:31:31.725 again, is when an amino acid change

00:31:31.725 --> 00:31:32.580 occurs in a DNA sequence.

00:31:32.580 --> 00:31:35.070 DNA sequence changes can either change the amino acid

00:31:35.070 --> 00:31:38.620 or not depending on what the sequence of the code on

00:31:38.620 --> 00:31:41.600 the three base pair code on in the DNA sequences.

00:31:41.600 --> 00:31:43.680 So if there's a replacement, we tally it here,

00:31:43.680 --> 00:31:45.730 if it's a synonymous change, that doesn't change the amino

00:31:45.730 --> 00:31:48.473 acid, we tally it here, these ones are preserved.

00:31:48.473 --> 00:31:49.760 Sometimes changes are presumably neutral because

00:31:49.760 --> 00:31:52.370 they don't change anything about your protein.

00:31:52.370 --> 00:31:55.690 And then the if it's a polymorphic replacement,

00:31:55.690 --> 00:31:57.210 then we see it here.

00:31:57.210 --> 00:31:58.920 And if it's a synonymous polymorphism we see it here.

00:31:58.920 --> 00:32:01.460 So under the hypothesis that I mentioned,

00:32:01.460 --> 00:32:03.930 all three of these cells should occur, it should

00:32:03.930 --> 00:32:06.330 be sort of changing in exactly the same way

00:32:06.330 --> 00:32:08.720 because polymorphic sites, whether they're replacement

00:32:08.720 --> 00:32:10.840 are synonymous, we're assuming are neutral,

00:32:10.840 --> 00:32:12.380 synonymous sites, whether the divergent

00:32:12.380 --> 00:32:15.084 or polymorphic, we're assuming is neutral.

00:32:15.084 --> 00:32:16.330 The only one that apparently that under

00:32:17.191 --> 00:32:19.021 assumption is not neutral are these replacement

00:32:19.021 --> 00:32:21.690 changes at replacement divergence sites.

00:32:21.690 --> 00:32:25.390 So, if this replacement divergence, if the marginals

00:32:25.390 --> 00:32:28.510 add up so that this replacement divergence is sort of in

00:32:28.510 --> 00:32:30.415 line with all these others, then we assume nothing important

00:32:30.415 --> 00:32:33.060 is happening in that gene, it's probably not selected,

00:32:33.060 --> 00:32:35.460 it's just neutral changes that are happening there.

00:32:35.460 --> 00:32:37.924 If this divergence is higher, though,

00:32:37.924 --> 00:32:39.391 then we might conclude that it's under

00:32:39.391 --> 00:32:40.860 selection for changes at a rapid pace.

00:32:40.860 --> 00:32:43.770 So neutrality yields a DN over DS that's equal

00:32:43.770 --> 00:32:45.945 to the PN over PS positive selection means

00:32:45.945 --> 00:32:49.680 that the DN DS is greater than the PN PS and negative

00:32:49.680 --> 00:32:53.010 selection where changes are actually being selected against

00:32:53.010 --> 00:32:56.130 at a high level indicates the DN DS

00:32:56.130 --> 00:32:57.913 is gonna be less than PN PS.

00:32:58.840 --> 00:33:01.010 All right now Let's get to a little bit of the

00:33:01.010 --> 00:33:04.245 complexity on this thing that I mentioned that's called

00:33:04.245 --> 00:33:05.078 Poisson random field theory, quantitatively estimates

00:33:05.078 --> 00:33:09.270 gene-wide selection intensity.

00:33:09.270 --> 00:33:10.820 So what you can do is take that

00:33:12.108 --> 00:33:13.880 same two by two table, and you can say under a model of

00:33:13.880 --> 00:33:17.675 selection, what do we actually think is happening here.

00:33:17.675 --> 00:33:19.877 And that gives us the ability to estimate the selection

00:33:19.877 --> 00:33:21.760 coefficient, which is a basically the rate at which that

00:33:21.760 --> 00:33:25.420 change allows the virus to increase its reproductive ability

00:33:25.420 --> 00:33:27.382 or survival ability in the host.

00:33:27.382 --> 00:33:31.700 And that that is this gamma term right here

00:33:31.700 --> 00:33:34.070 in these terms, and this, these look complicated,

00:33:34.070 --> 00:33:36.350 but essentially, these formulas are just saying

00:33:36.350 --> 00:33:38.880 that the expectation for a synonymous sorry,

00:33:38.880 --> 00:33:41.385 the synonymous and replacement have reversed

00:33:41.385 --> 00:33:43.061 on this chart compared to the last,

00:33:43.061 --> 00:33:44.538 so don't be confused by that.

00:33:44.538 --> 00:33:45.480 But the expectation under synonymous

00:33:45.480 --> 00:33:47.613 changes is essentially the mutation rate.

00:33:48.487 --> 00:33:50.220 And these terms are just about the sampling properties

00:33:50.220 --> 00:33:52.470 of when you sequence how many of these things you get,

00:33:52.470 --> 00:33:54.600 I don't need to go into the detail about that here.

00:33:54.600 --> 00:33:56.680 Similarly, the polymorphic sequence

00:33:56.680 --> 00:33:59.850 is just basically dependent on the mutation rate.

00:33:59.850 --> 00:34:02.060 How the replacement sequences are a little bit more

00:34:02.060 --> 00:34:06.680 complicated in that they have to account

00:34:06.680 --> 00:34:09.683 for kinds of selection that may be going on.

00:34:10.780 --> 00:34:12.450 For reasons that I don't wanna get into

00:34:12.450 --> 00:34:15.820 the polymorphic selection, so both of them are depending

00:34:15.820 --> 00:34:17.990 on the mutation rate for replacement sites,

00:34:17.990 --> 00:34:20.045 and both of them depend on

00:34:20.045 --> 00:34:22.620 how much each variant is selected.

00:34:22.620 --> 00:34:24.810 Selection doesn't pack the polymorphism

00:34:24.810 --> 00:34:27.000 to a certain degree in the sense that if variants

00:34:27.000 --> 00:34:29.520 are moving through the population very fast,

00:34:29.520 --> 00:34:32.180 that can change how much polymorphism you see.

00:34:32.180 --> 00:34:35.750 But then if you use these sampling formulas, and the formula

00:34:35.750 --> 00:34:38.050 for the estimate of the strength of selection,

00:34:38.050 --> 00:34:40.850 given how many variants we see changing,

00:34:40.850 --> 00:34:43.560 you get these formulas for how much replacement

00:34:44.409 --> 00:34:46.697 divergence and polymorphism you expect to see.

00:34:46.697 --> 00:34:48.830 So this is a population genetics that was worked

00:34:48.830 --> 00:34:52.420 out by Stan Sawyer and Dan Hurley in 1992.

00:34:52.420 --> 00:34:55.860 The only change I'm making in this is pure F,

00:34:55.860 --> 00:35:00.400 instead of using a year which was how many grants

00:35:00.400 --> 00:35:04.190 that you see in the the McConnell Craven uses it,

00:35:04.190 --> 00:35:07.680 I'm taking the probabilities of replacement divergence

00:35:07.680 --> 00:35:10.695 and the probabilities of some polymorphism

00:35:10.695 --> 00:35:12.286 and putting them in here.

00:35:12.286 --> 00:35:13.250 And the advantage here is that what

00:35:13.250 --> 00:35:15.170 I can do with that is what I mentioned earlier,

00:35:15.170 --> 00:35:17.750 I can go back to the old mass MACML

00:35:17.750 --> 00:35:20.320 approach sequence clustering approach

00:35:20.320 --> 00:35:23.070 that I mentioned before, estimating those probabilities

00:35:24.665 --> 00:35:26.530 across the entire gene, I can then estimate action across

00:35:26.530 --> 00:35:30.370 the entire gene by using these probability single site,

00:35:30.370 --> 00:35:32.430 I don't have changes for single site.

00:35:32.430 --> 00:35:33.850 So what this allows

00:35:33.850 --> 00:35:37.709 us to estimate this gamma, minimizing likelihood of what

00:35:37.709 --> 00:35:41.900 gamma is to blame those problems exist, see.

00:35:41.900 --> 00:35:46.360 So this is a very complex diagram of how this all works,

00:35:46.360 --> 00:35:50.050 again, is a pretty elaborate method of computation.

00:35:50.050 --> 00:35:53.190 But again, has the nice properties that I'm not putting

00:35:53.190 --> 00:35:55.090 in any I'm not using assumptions

00:35:55.090 --> 00:35:56.480 and not putting in any parameters.

00:35:56.480 --> 00:35:57.934 They go in.

00:35:57.934 --> 00:36:00.740 I just take the polymorph at the end analyze it for

00:36:00.740 --> 00:36:03.860 weather sites are clustered into four different categories.

00:36:03.860 --> 00:36:05.690 Again, replacement polymorphism.

00:36:05.690 --> 00:36:07.050 That's this arc here.

00:36:07.050 --> 00:36:11.233 So polymorphisms anonymous divergence, placement divergence,

00:36:12.427 --> 00:36:15.300 we cluster within all four of those categories.

00:36:15.300 --> 00:36:16.990 We calculate the model average probability,

00:36:16.990 --> 00:36:20.200 all those clusters and merge the data together.

00:36:20.200 --> 00:36:21.560 I'm not going to go through the details.

00:36:21.560 --> 00:36:24.890 But just if you were to do essentially the KML,

00:36:24.890 --> 00:36:27.050 like clustering on those four categories

00:36:27.050 --> 00:36:29.570 for a particular gene polymorphisms

00:36:29.570 --> 00:36:32.690 and Ana's polymorphisms, monster and placement divergence

00:36:32.690 --> 00:36:36.550 if you plug those in, to the formulas I showed you before,

00:36:36.550 --> 00:36:39.354 you're basically plugging into these categories,

00:36:39.354 --> 00:36:40.904 you can estimate those formulas.

00:36:40.904 --> 00:36:42.000 And in the end, what you get is

00:36:42.000 --> 00:36:46.763 an estimate of gamma across nucleotide positions in a gene.

00:36:48.750 --> 00:36:50.870 I won't go into what this result here,

00:36:50.870 --> 00:36:52.770 it's an interesting result for reasons

00:36:53.920 --> 00:36:55.180 that are only of interest mostly to evolutionary

00:36:56.146 --> 00:36:58.150 biologist, but you can see here in this particular gene

00:36:58.150 --> 00:37:02.360 that there's a lot of variation in the selection

00:37:02.360 --> 00:37:04.140 intensity across the gene.

00:37:04.140 --> 00:37:05.590 Now, that is actually really

00:37:05.590 --> 00:37:07.560 consistent with what we'd expect.

00:37:07.560 --> 00:37:10.223 From a sort of basic biology standpoint.

00:37:11.340 --> 00:37:13.210 Different parts of a gene are gonna either

00:37:13.210 --> 00:37:15.230 be very strongly selected to stay the same

00:37:15.230 --> 00:37:18.321 or they're gonna change, you shouldn't really expect

00:37:18.321 --> 00:37:19.770 that all parts of gene are equally likely to change.

00:37:19.770 --> 00:37:22.129 And this gives a very nice diagram

00:37:22.129 --> 00:37:23.185 that allows you to understand how

00:37:23.185 --> 00:37:24.730 it's different across the gene.

00:37:24.730 --> 00:37:27.070 So if we compare this kind of approach

00:37:27.070 --> 00:37:30.451 to the McDonald kreitman tests, which again,

00:37:30.451 --> 00:37:33.460 are just putting in the DN DS, PN PS values

00:37:33.460 --> 00:37:35.666 into this two by two table,

00:37:35.666 --> 00:37:38.520 and I went through that, the important difference is that

00:37:38.520 --> 00:37:41.760 the Mk test assumes this intergenic homogeneous selection

00:37:41.760 --> 00:37:44.070 that in fact, a gene has the same selection

00:37:44.070 --> 00:37:45.570 across the entire sequence.

00:37:45.570 --> 00:37:48.350 The problem with that is if you have one small

00:37:48.350 --> 00:37:49.983 region that's under selection,

00:37:49.983 --> 00:37:52.633 the averaging out process across that entire gene

00:37:52.633 --> 00:37:53.910 can mean that you don't detect the selection there,

00:37:53.910 --> 00:37:57.160 even though it may be very strong for that small region.

00:37:57.160 --> 00:38:00.540 And so the hope is that mastery graph can

00:38:00.540 --> 00:38:02.120 identify those regions much better

00:38:02.120 --> 00:38:04.290 than MK for instance, would.

00:38:04.290 --> 00:38:07.173 And in fact, I went through this already.

00:38:08.528 --> 00:38:11.673 I'll just skip past this because I went through it already.

00:38:12.900 --> 00:38:17.820 And this it does do that.

00:38:17.820 --> 00:38:20.830 So this is an example of McDonnell Craven

00:38:20.830 --> 00:38:23.290 tests here applied to a Drosophila gene,

00:38:23.290 --> 00:38:27.200 what you see is this high evolution of a high level

00:38:27.200 --> 00:38:29.750 of replacement divergence, which turns out

00:38:29.750 --> 00:38:32.760 to indicate high selection.

00:38:32.760 --> 00:38:35.370 And you can see here that the DN DS ratio

00:38:35.370 --> 00:38:38.410 is about eight to one word as the PN PS ratio

00:38:38.410 --> 00:38:39.880 is almost even.

00:38:39.880 --> 00:38:42.390 So this is a gene that's under very strong selection

00:38:42.390 --> 00:38:44.970 based on the McDonald kreitman test.

00:38:44.970 --> 00:38:46.820 Now, interestingly, so this one works

00:38:46.820 --> 00:38:49.000 with a homogeneity.

00:38:49.000 --> 00:38:53.427 And then if you analyze the ACP 26 AA gene

00:38:55.220 --> 00:38:57.900 and look for the probability of all four categories.

00:38:57.900 --> 00:39:00.960 These are the four categories and of course,

00:39:00.960 --> 00:39:03.622 the replacement divergence here is the one

00:39:03.622 --> 00:39:05.720 that's most likely to drive selection.

00:39:05.720 --> 00:39:08.773 What do you get when you estimate gamma using this?

00:39:08.773 --> 00:39:09.840 Well, interestingly, what you see is not something

00:39:09.840 --> 00:39:12.710 that's under very strong selection across the entire gene,

00:39:12.710 --> 00:39:14.970 but something that's on moderately strong selection,

00:39:14.970 --> 00:39:16.740 basically in the second half of the gene,

00:39:16.740 --> 00:39:18.780 and then one peak of very strong

00:39:18.780 --> 00:39:20.850 selection right around the middle of the gene.

00:39:20.850 --> 00:39:23.060 And this is visible in currents because

00:39:23.060 --> 00:39:25.690 of a number of changes that occur

00:39:25.690 --> 00:39:28.280 in one particular domain of the gene here.

00:39:28.280 --> 00:39:30.370 Now, if you look at just the replacement divergence,

00:39:30.370 --> 00:39:32.176 you wouldn't be able to figure this out.

00:39:32.176 --> 00:39:33.710 Because you see there are other

00:39:33.710 --> 00:39:34.722 peaks along here.

00:39:34.722 --> 00:39:36.180 Those don't turn out to be so important.

00:39:36.180 --> 00:39:37.960 And the reason why they don't turn out to be so important

00:39:39.206 --> 00:39:40.820 is that the synonymous divergence synonymous by morphism

00:39:40.820 --> 00:39:42.110 replacement polymorphism.

00:39:42.110 --> 00:39:44.370 Tell us more about the underlying mutation rate

00:39:44.370 --> 00:39:46.650 that says those elevations are probably have

00:39:46.650 --> 00:39:49.300 something to do with mutation rate, and not necessarily

00:39:49.300 --> 00:39:52.340 to do with added divergence.

00:39:52.340 --> 00:39:53.860 You can sort of see this elevation

00:39:53.860 --> 00:39:55.940 on the right hand side over here compared

00:39:55.940 --> 00:39:58.930 to the small dip right here and up here

00:39:58.930 --> 00:40:01.803 and the way it all works out mathematically

00:40:01.803 --> 00:40:04.110 is we can really see that there's strong selection here.

00:40:04.110 --> 00:40:06.230 We can also get what I call model intervals for this.

00:40:06.230 --> 00:40:08.010 If you look across all the models,

00:40:08.010 --> 00:40:10.580 what are the estimates of selection?

00:40:10.580 --> 00:40:14.480 Possibly, what do we get is the 95% model interval for this?

00:40:14.480 --> 00:40:17.391 And that's what these very faint gray lines you

00:40:17.391 --> 00:40:18.910 may be able to see are those allow us to detect whether

00:40:18.910 --> 00:40:21.560 these are significant, least significant,

00:40:21.560 --> 00:40:24.080 statistically significant differences in selection.

00:40:24.080 --> 00:40:26.650 All right, I'm gonna skip through this

00:40:26.650 --> 00:40:28.572 just because I want to spend the time

00:40:28.572 --> 00:40:29.405 but the point is, you can do this for other genes,

00:40:29.405 --> 00:40:31.530 and it shows similar results that allow us

00:40:31.530 --> 00:40:34.324 to understand where sites are under selection in that gene.

00:40:34.324 --> 00:40:36.920 I'll just cover a few more examples

00:40:36.920 --> 00:40:38.970 of how we've used this to give you an idea

00:40:38.970 --> 00:40:41.740 of what it can look like in a comparison between humans

00:40:41.740 --> 00:40:43.870 and chimpanzees where we've run this just to understand

00:40:43.870 --> 00:40:45.973 how we've diverged from chimpanzees.

00:40:46.870 --> 00:40:49.660 We see a bunch of different examples here.

00:40:49.660 --> 00:40:51.530 Again, doing a little bit of comparison to

00:40:51.530 --> 00:40:54.066 that traditional McDonald kreitman test

00:40:54.066 --> 00:40:55.640 and the mass PRF test.

00:40:55.640 --> 00:40:59.995 Here you see a gene, which is statistically significant

00:40:59.995 --> 00:41:01.246 people's point of view.

00:41:01.246 --> 00:41:03.640 Based on the Mk tests, the four categories

00:41:03.640 --> 00:41:06.780 of the four tallies of which are indicated here.

00:41:06.780 --> 00:41:09.710 Here's the MASS -PRF profile, and it shows us again

00:41:09.710 --> 00:41:11.880 a particular region within this SLC AA

00:41:11.880 --> 00:41:14.110 one gene that is under selection.

00:41:14.110 --> 00:41:17.106 There are interesting stories behind all of these,

00:41:17.106 --> 00:41:18.523 but I'm not gonna take the time to go through them.

00:41:19.440 --> 00:41:21.800 Here's another example where and this is an example

00:41:21.800 --> 00:41:23.450 where the McDonald pregnant test

00:41:23.450 --> 00:41:24.790 comes out is not significant.

00:41:24.790 --> 00:41:26.450 There's just not that much divergence

00:41:26.450 --> 00:41:28.060 compared to the other categories.

00:41:28.060 --> 00:41:31.640 But if you do this, spatially with the MASS-PRF test,

00:41:31.640 --> 00:41:34.010 you actually see that a very central region there

00:41:34.010 --> 00:41:37.200 has very strong selection, and then the rest of the gene

00:41:37.200 --> 00:41:40.640 is under almost zero selection or almost no selection.

00:41:40.640 --> 00:41:42.660 So this is an example I talked about,

00:41:42.660 --> 00:41:44.660 where you could have some very small portion

00:41:44.660 --> 00:41:46.580 of the gene under very strongest selection.

00:41:46.580 --> 00:41:49.136 And McDonald-Kreitman test wouldn't detect it

00:41:49.136 --> 00:41:50.910 because it's averaging over the entire gene.

00:41:50.910 --> 00:41:52.350 Similarly, you'll get some genes.

00:41:52.350 --> 00:41:53.950 Oops, I didn't mean to do that.

00:41:53.950 --> 00:41:58.200 Some jeans, here's M gamma over here, where there's a...

00:41:58.200 --> 00:41:59.270 Well, let me turn to that one last.

00:41:59.270 --> 00:42:01.580 Actually, let me look at TPH First,

00:42:01.580 --> 00:42:06.340 there's no statistical selection according to the Mk tests.

00:42:06.340 --> 00:42:07.810 And in fact, in our MASS-PRF,

00:42:07.810 --> 00:42:09.240 there's no specific selection either

00:42:09.240 --> 00:42:12.440 the error bars are entirely overlapping zero here,

00:42:12.440 --> 00:42:14.590 which indicates no selection.

00:42:14.590 --> 00:42:16.180 Lastly, here's M gamma.

00:42:16.180 --> 00:42:18.370 This is the one of the very few examples

00:42:18.370 --> 00:42:21.369 we were able to find where McDonald test did detect

00:42:21.369 --> 00:42:23.740 selection where, where MASS-PRF didn't.

00:42:23.740 --> 00:42:25.620 As you can see, there's quite high tallies here,

00:42:25.620 --> 00:42:27.080 which means there's a lot of power

00:42:27.080 --> 00:42:28.389 to detect selection if it's there,

00:42:28.389 --> 00:42:30.040 but it's probably not very strong,

00:42:30.040 --> 00:42:31.880 because the numbers are not all that different

00:42:31.880 --> 00:42:32.723 from each other.

00:42:34.364 --> 00:42:36.250 And McDonald-Kreitman says it's statistically significant.

00:42:36.250 --> 00:42:38.600 Now the reason why McDonald Kreitman is telling

00:42:39.502 --> 00:42:40.820 it's statistic's nothing compared to mass PRF

00:42:40.820 --> 00:42:43.940 is that actually, I didn't explain this in detail to you.

00:42:43.940 --> 00:42:46.540 But McDonald- Kreitman doesn't actually assume

00:42:46.540 --> 00:42:48.370 that there's an elevation of rate here.

00:42:48.370 --> 00:42:50.830 And so the significance here is actually driven by

00:42:50.830 --> 00:42:53.310 the high polymorphic replacement level.

00:42:53.310 --> 00:42:55.800 So there's a lot of polymorphic replacements in there.

00:42:55.800 --> 00:42:58.450 And what that means is there's some other

00:42:59.641 --> 00:43:00.900 kind of selection that isn't a directional selection.

00:43:00.900 --> 00:43:02.270 I won't go into the details there.

00:43:02.270 --> 00:43:04.380 But the nice thing is that in the examples

00:43:04.380 --> 00:43:06.740 where we find that McDonald kreitman is statistically

00:43:06.740 --> 00:43:09.790 significant and MASS-PRF isn't examples

00:43:09.790 --> 00:43:11.970 where in fact MASS-PRF is not designed to detect

00:43:11.970 --> 00:43:14.063 that kind of selection and MK test is.

00:43:15.300 --> 00:43:18.138 In general MASS-PRF turned out to be significant

00:43:18.138 --> 00:43:21.207 in almost every case math MK tests were not.

00:43:21.207 --> 00:43:23.610 Okay, so how can we use this, apply this

00:43:23.610 --> 00:43:26.880 to instances like COVID-19, the point of this whole talk,

00:43:26.880 --> 00:43:29.130 and I'm just gonna give you one example first

00:43:30.085 --> 00:43:32.128 to justify why we think it's a good idea,

00:43:32.128 --> 00:43:33.844 because we don't have results on doing it,

00:43:33.844 --> 00:43:35.790 at least not many results on doing it to COVID-19

00:43:35.790 --> 00:43:38.810 yet, and that is that we applied this influenza before,

00:43:38.810 --> 00:43:42.970 which has some similarities to COVID-19, as everyone knows

00:43:42.970 --> 00:43:46.370 and in influenza, again, we're interested in looking across

00:43:46.370 --> 00:43:48.340 the gene are there sites that are under selection

00:43:48.340 --> 00:43:50.380 because those sites that are under selection

00:43:50.380 --> 00:43:53.480 are candidates where we need to be aware that

00:43:53.480 --> 00:43:56.600 in fact, vaccines need like for every year they design

00:43:57.554 --> 00:43:58.387 a new influenza vaccine, right?

00:43:58.387 --> 00:43:59.910 And what they're trying to do is accommodate

00:43:59.910 --> 00:44:02.500 the fact that these changes occur on the sites

00:44:02.500 --> 00:44:04.430 that are actually susceptible

00:44:04.430 --> 00:44:08.430 to your immune system recognizing the influenza virus.

00:44:08.430 --> 00:44:10.590 So we need to understand those sites that are changing

00:44:10.590 --> 00:44:13.390 and where they are in in order to design

00:44:13.390 --> 00:44:16.060 more universal vaccines that maybe could target sites

00:44:16.060 --> 00:44:18.880 that won't change rapidly because they can't change

00:44:18.880 --> 00:44:21.870 because they're structurally constrained in the virus.

00:44:21.870 --> 00:44:25.312 So what we did was apply this MASS-PRF approach

00:44:25.312 --> 00:44:28.950 to influenza similarly on a phylogeny

00:44:28.950 --> 00:44:30.350 to like I described for Coronavirus.

00:44:30.350 --> 00:44:32.550 I don't have the phylogeny in the slide set,

00:44:33.400 --> 00:44:36.280 but the point is just looking at the ancestral influenza

00:44:36.280 --> 00:44:40.110 and it's divergent sites within a particular region.

00:44:40.110 --> 00:44:42.850 And what we were able to do is identify a set of sites

00:44:42.850 --> 00:44:45.600 that are under select---ion using mass PRF

00:44:45.600 --> 00:44:47.930 that are beyond what people had prophesied

00:44:47.930 --> 00:44:49.920 as positive selection sites in the past.

00:44:49.920 --> 00:44:52.630 So there's a paper by Westgeest al 2012

00:44:52.630 --> 00:44:55.350 which is essentially the gold standard for this

00:44:55.350 --> 00:44:57.830 and they found a bunch of sites that are all

00:44:57.830 --> 00:45:00.120 these circled sites in gray MASS-PRF.

00:45:00.120 --> 00:45:02.590 Also found those the orange diagram here

00:45:02.590 --> 00:45:06.570 is the MASS-PRF for this gene.

00:45:08.550 --> 00:45:10.140 And it also identified other sites

00:45:10.140 --> 00:45:11.790 that are under selection as well.

00:45:13.756 --> 00:45:15.931 And we're in the process of understanding

00:45:15.931 --> 00:45:17.040 better how those can be validated.

00:45:17.040 --> 00:45:19.860 But the ultimate point is that

00:45:19.860 --> 00:45:24.540 these are important selected sites that may be relevant

00:45:24.540 --> 00:45:28.080 to the design of vaccines for influenza.

00:45:28.080 --> 00:45:29.930 So similarlY, we'd like to illuminate

00:45:30.913 --> 00:45:33.710 which sites might be changing rapidly

00:45:33.710 --> 00:45:36.083 and under positive selection in Coronavirus,

00:45:37.241 --> 00:45:38.913 not only during the human epidemic,

00:45:38.913 --> 00:45:40.930 but again during the zonotic zoonotic time period.

00:45:40.930 --> 00:45:42.670 And so now we're finally coming to the final

00:45:42.670 --> 00:45:45.530 part of my talk, which is what we're doing

00:45:45.530 --> 00:45:48.440 in terms of the model average estimation the mcos

00:45:48.440 --> 00:45:51.072 and natural selection in SARS coronavirus,

00:45:51.072 --> 00:45:52.553 one and SARS coronavirus two,

00:45:52.553 --> 00:45:53.400 Corona viruses during zoonosis.

00:45:53.400 --> 00:45:55.521 But the whole point here is really

00:45:55.521 --> 00:45:56.730 explain to you what I've done because the results I have

00:45:56.730 --> 00:46:00.696 as I said are I just have a few plots of some of the stuff

00:46:00.696 --> 00:46:02.559 longest selection we were able to check

00:46:02.559 --> 00:46:04.619 because we have to process through a lot more data

00:46:04.619 --> 00:46:06.679 before we get a more in depth look at the lesser

00:46:06.679 --> 00:46:10.130 selected sites that are on these genes.

00:46:10.130 --> 00:46:13.400 And so we looked at this for the for Coronavirus.

00:46:13.400 --> 00:46:17.110 This is just a Coronavirus, Getty image that Yale

00:46:17.110 --> 00:46:20.453 has used looking at Coronavirus.

00:46:21.450 --> 00:46:23.010 And again, as I mentioned,

00:46:23.010 --> 00:46:26.170 we're looking at these two sites of where COVID-19

00:46:26.170 --> 00:46:30.100 emergence occurred, and where SARS emergence occurred.

00:46:30.100 --> 00:46:31.960 And the question is, are there changes

00:46:32.855 --> 00:46:34.010 that happen there that are specifically

00:46:34.010 --> 00:46:37.870 responsible perhaps for those zoonosis and the only results

00:46:37.870 --> 00:46:40.230 I have are just a few results again, highlighting some of

00:46:40.230 --> 00:46:42.340 the strongest selection we saw.

00:46:42.340 --> 00:46:44.190 This is actually a diagram of the spike

00:46:44.190 --> 00:46:46.880 protein which if you've heard much about COVID-19

00:46:46.880 --> 00:46:49.430 molecular biology, you probably have heard about the spike

00:46:50.361 --> 00:46:52.412 protein, it's what sticks out from the virus.

00:46:52.412 --> 00:46:55.530 It's what grabs onto the AC receptor,

00:46:55.530 --> 00:46:58.330 and essentially is what most vaccines

00:46:58.330 --> 00:47:01.360 that one might design for the virus would target.

00:47:01.360 --> 00:47:04.400 And the point is that the recombination binding

00:47:04.400 --> 00:47:07.127 domain, which has gotten a lot of press already turns out

00:47:07.127 --> 00:47:07.960 to have the selected sites.

00:47:07.960 --> 00:47:11.540 You can see them here, here, here and here.

00:47:11.540 --> 00:47:12.567 These are sites that are selected,

00:47:12.567 --> 00:47:13.400 meaning they're changing rapidly

00:47:13.400 --> 00:47:16.750 during the pre zoonotic phase.

00:47:16.750 --> 00:47:19.350 So these are sites that are changing, not in humans,

00:47:20.410 --> 00:47:21.620 but in the bats in the pangolins.

00:47:21.620 --> 00:47:24.580 And whatever other animals that this virus

00:47:24.580 --> 00:47:27.487 is spreading among, or has been spreading among

00:47:27.487 --> 00:47:28.680 before the zoonosis to humans.

00:47:28.680 --> 00:47:29.888 So then the question is, are similar sites under

00:47:29.888 --> 00:47:30.721 selection during zoonosis?

00:47:30.721 --> 00:47:35.560 And during post zoonosis?

00:47:35.560 --> 00:47:37.610 And the answer right now is yes,

00:47:37.610 --> 00:47:38.720 it seems kind of similar,

00:47:38.720 --> 00:47:40.060 although we don't get the same sites.

00:47:40.060 --> 00:47:42.149 So we have to do a little bit

00:47:42.149 --> 00:47:43.830 more molecular, you know, staring at this and understanding

00:47:43.830 --> 00:47:46.313 it because these results are literally

00:47:46.313 --> 00:47:47.676 I got these results today, actually.

00:47:47.676 --> 00:47:50.260 So we have to sort of do more of this

00:47:51.165 --> 00:47:52.630 and we actually can actually look at more depth

00:47:53.508 --> 00:47:54.530 and get more sites with other approaches

00:47:54.530 --> 00:47:57.290 that we haven't implemented at this moment.

00:47:57.290 --> 00:47:58.123 But during near zoonosis what you see is again,

00:47:58.123 --> 00:48:03.020 the selected sites which are in bright red

00:48:06.387 --> 00:48:08.267 are also on the sort of the visible side

00:48:08.267 --> 00:48:10.350 of the recombination binding domain

00:48:12.796 --> 00:48:17.380 of the spike protein, which is the tip

00:48:17.380 --> 00:48:21.363 the outside portion of this gene.

00:48:22.742 --> 00:48:24.100 Lastly, if we look post-zoonosis that's in

00:48:24.100 --> 00:48:26.400 the evolution of humans, we again see that

00:48:26.400 --> 00:48:30.043 the selected sites are sites that are at this tip region.

00:48:32.585 --> 00:48:34.615 Again, none of this is terribly surprising.

00:48:34.615 --> 00:48:36.378 The interesting thing is that it kind of indicates

00:48:36.378 --> 00:48:37.700 that the zoonosis it kind of indicates consistency.

00:48:37.700 --> 00:48:40.061 Again, there's a lot more to do before

00:48:40.061 --> 00:48:41.547 we can conclude anything like this,

00:48:41.547 --> 00:48:43.610 but the idea we have right now indicates

00:48:43.610 --> 00:48:46.250 a good deal of consistency between the selection

00:48:46.250 --> 00:48:50.570 that's ongoing in humans during zoonosis and pre zoonosis.

00:48:50.570 --> 00:48:52.960 And what that implies is that this may

00:48:53.865 --> 00:48:55.520 well have been as I said, very briefly,

00:48:55.520 --> 00:48:58.930 during this talk an instance where there's a virus

00:48:59.950 --> 00:49:01.020 just circulating around in bats and penguins

00:49:01.020 --> 00:49:03.580 that could have caused this disease at any time,

00:49:03.580 --> 00:49:06.560 it's just a matter of whether or not we actually

00:49:06.560 --> 00:49:10.990 have exposure to, to those organisms

00:49:10.990 --> 00:49:13.590 that allows the transmission to happen.

00:49:13.590 --> 00:49:15.540 Consistent with this, I'll just mention

00:49:17.058 --> 00:49:18.352 a couple like verbal points,

00:49:18.352 --> 00:49:20.447 which is that all the evidence that we have indicates

00:49:20.447 --> 00:49:23.150 that this virus spread extremely quickly

00:49:23.150 --> 00:49:26.010 from the moment that it zoonosis into humans.

00:49:26.010 --> 00:49:28.190 And in fact, in most cases of zoonosis,

00:49:28.190 --> 00:49:29.440 we find that that's true,

00:49:30.839 --> 00:49:32.510 which is somewhat counterintuitive.

00:49:32.510 --> 00:49:34.157 Obviously, it hasn't adapted to humans,

00:49:34.157 --> 00:49:37.003 it has adapted to the amount of mammalian immune system.

00:49:37.003 --> 00:49:38.893 And so to the extent that our immune system is not

00:49:38.893 --> 00:49:40.730 tremendously different from that of bats or pangolins,

00:49:40.730 --> 00:49:43.670 it may be not surprising that it can infect us.

00:49:43.670 --> 00:49:46.619 But one of the things that is true is that

00:49:46.619 --> 00:49:47.780 if it did not spread very quickly,

00:49:47.780 --> 00:49:50.720 very easily from the very moment it transmitted to someone,

00:49:50.720 --> 00:49:52.330 it would probably lead to a dead end.

00:49:52.330 --> 00:49:54.810 In other words, if you don't have

00:49:54.810 --> 00:49:57.163 an ability to transmit and spread just from the get go,

00:49:57.163 --> 00:49:59.630 the first person who gets infected

00:49:59.630 --> 00:50:02.140 is very unlikely to transmit it to someone else.

00:50:02.140 --> 00:50:04.330 So it sort of has to be well pre adapted

00:50:04.330 --> 00:50:07.120 for a zoonotic event to actually spread in humans.

00:50:07.120 --> 00:50:09.273 Now there's, we need more zoonotic events,

00:50:10.816 --> 00:50:11.649 God forbid that it actually happens,

00:50:13.440 --> 00:50:15.064 to really get a better picture of that.

00:50:15.064 --> 00:50:15.897 But the general result and the scientific

00:50:15.897 --> 00:50:18.091 literature does seem to show that zoonosis happens.

00:50:18.091 --> 00:50:22.360 the disease's already well set to cause problems.

00:50:22.360 --> 00:50:23.770 And the examples that we don't have where

00:50:23.770 --> 00:50:25.340 it happens like that, like MERS

00:50:26.886 --> 00:50:28.786 or like, well, MERS is a good example.

00:50:29.869 --> 00:50:31.031 It's a really deadly disease,

00:50:31.031 --> 00:50:31.980 but it doesn't transmit well among humans.

00:50:31.980 --> 00:50:34.720 And so that's an example where maybe it's transmitting

00:50:34.720 --> 00:50:37.210 to humans, but it's not transmitting among humans.

00:50:37.210 --> 00:50:38.960 And it's very hard for that disease

00:50:40.067 --> 00:50:42.017 to catch on within the human population

00:50:43.194 --> 00:50:45.229 and do human transmission as opposed to zoonotic events.

00:50:45.229 --> 00:50:46.592 And that's because it doesn't transmit

00:50:46.592 --> 00:50:48.342 and it doesn't usually evolve that ability

00:50:48.342 --> 00:50:50.650 to transmit over the short time that

00:50:50.650 --> 00:50:53.280 that individuals might get infected.

00:50:53.280 --> 00:50:56.880 when when they get it usually from camels.

00:50:56.880 --> 00:50:59.000 Okay, so I've showed you those examples.

00:50:59.000 --> 00:51:01.780 I just wanna to mention what else we're gonna be doing.

00:51:01.780 --> 00:51:03.780 So I what I just showed you was actually

00:51:04.668 --> 00:51:06.420 the sort of SARS coronavirus to some sites

00:51:06.420 --> 00:51:07.990 that are under selection in search

00:51:07.990 --> 00:51:09.570 for Coronavirus two genes.

00:51:09.570 --> 00:51:12.031 This is the S gene right here.

00:51:12.031 --> 00:51:12.864 That's the spike gene.

00:51:12.864 --> 00:51:14.710 We're gonna be looking at that in SARS coronavirus,

00:51:14.710 --> 00:51:17.530 one and two, we're also going to be looking

00:51:17.530 --> 00:51:21.660 at other genes in the genomes.

00:51:21.660 --> 00:51:22.960 These have other functions.

00:51:22.960 --> 00:51:26.142 The M gene, for instance, is a membrane gene.

00:51:26.142 --> 00:51:27.990 So it might be relevant to and the gene

00:51:27.990 --> 00:51:32.290 as well might be relevant to vaccine generation.

00:51:32.290 --> 00:51:34.610 Like if we could generate a vaccine that targeted

00:51:34.610 --> 00:51:37.560 those, maybe they would be unable to change at the same

00:51:41.249 --> 00:51:44.045 pace that spike protein would they might be more conserved.

00:51:44.045 --> 00:51:44.878 And that might be one approach towards developing a vaccine.

00:51:46.312 --> 00:51:47.145 That would be a longer term vaccine because one thing we

00:51:48.726 --> 00:51:50.193 have to worry about, of course with this Coronavirus,

00:51:53.186 --> 00:51:55.378 is and I have other research that we're doing on

00:51:55.378 --> 00:51:57.275 this question, which I'd love to talk about if anyone's

00:51:57.275 --> 00:51:58.771 curious, but you can estimate

00:51:58.771 --> 00:52:00.152 what the actual waning immunity of it is,

00:52:00.152 --> 00:52:00.985 even though we don't have data on that by Looking

00:52:03.422 --> 00:52:05.180 at other related species and using the phylogeny

00:52:05.180 --> 00:52:07.970 to understand how the how the waning immunity

00:52:07.970 --> 00:52:09.380 has evolved across the species

00:52:09.380 --> 00:52:11.230 and what the projected or most likely

00:52:12.158 --> 00:52:13.463 waning immunity of SARS coronavirus is,

00:52:14.600 --> 00:52:16.403 and it's, it tends to be it looks like

00:52:16.403 --> 00:52:17.746 it's around 80 weeks or so.

00:52:17.746 --> 00:52:20.815 So if we get about 8 weeks of waiting a period

00:52:20.815 --> 00:52:22.120 of immunity from this, that's not that

00:52:22.120 --> 00:52:24.750 much in terms of every two years or so we're gonna have

00:52:24.750 --> 00:52:27.540 Coronavirus coming around and in terms of we're going to

00:52:27.540 --> 00:52:29.340 be susceptible again to Coronavirus.

00:52:30.287 --> 00:52:31.120 Not that we're going to get it every two years.

00:52:33.436 --> 00:52:36.245 And what that would mean is that

00:52:36.245 --> 00:52:38.088 it's likely to persist as a circulating virus.

00:52:38.088 --> 00:52:39.839 And if it remains as deadly as it is that's a serious issue.

00:52:39.839 --> 00:52:41.544 So we're gonna really want to buy a vaccine.

00:52:41.544 --> 00:52:43.460 And we're not necessarily going to wanna have another flu

00:52:44.334 --> 00:52:45.213 vaccine that we have to get every year.

00:52:48.661 --> 00:52:50.632 So what we really want to do is target

00:52:50.632 --> 00:52:52.570 some genes that may be under more constraint

00:52:52.570 --> 00:52:55.630 then the recombination binding protein gene, the spike gene.

00:52:56.508 --> 00:52:58.280 So anyway, so the point is looking at multiple genes for

00:52:59.738 --> 00:53:01.410 trying to understand where conservative regions are where

00:53:02.809 --> 00:53:03.873 regions that are under selection are important.

00:53:05.224 --> 00:53:06.848 And we'll be doing that.

00:53:06.848 --> 00:53:10.625 And hopefully some of those results will

00:53:10.625 --> 00:53:14.507 help to guide the kind of generation of vaccines,

00:53:14.507 --> 00:53:16.374 and also the generation of therapeutics,

00:53:16.374 --> 00:53:18.642 because sites that are under

00:53:18.642 --> 00:53:19.866 selection are functional.

00:53:19.866 --> 00:53:20.892 So if you actually design a therapeutic

00:53:20.892 --> 00:53:22.418 that interferes with the sites that are under selection

00:53:22.418 --> 00:53:24.513 sort of in an opposite way, from vaccines, vaccines,

00:53:24.513 --> 00:53:26.041 we really want to target something that just doesn't change.

00:53:26.041 --> 00:53:27.058 With therapeutics, we may want to target

00:53:27.058 --> 00:53:29.586 the changing regions, if we can design something

00:53:29.586 --> 00:53:31.385 that generically does, because those changing

00:53:31.385 --> 00:53:32.314 regions are functional.

00:53:32.314 --> 00:53:33.147 In other words, those sites at the end of the spike protein

00:53:33.147 --> 00:53:35.440 are clearly ones that do bind the ACE gene.

00:53:35.440 --> 00:53:36.990 It's just that they're flexible

00:53:37.939 --> 00:53:39.383 about what they are in order to bind it.

00:53:41.975 --> 00:53:43.240 So we need to include

00:53:43.240 --> 00:53:46.047 all of those changing sites, if we wanna dissolve develop

00:53:46.047 --> 00:53:50.190 a therapeutic that for instance, would somehow interfering

00:53:50.190 --> 00:53:53.459 with the binding of Ace to receptors from the spike genes.

00:53:53.459 --> 00:53:56.223 So thank you very much for listening to the ongoing work

00:53:56.223 --> 00:53:59.025 we're doing on COVID-19.

00:53:59.025 --> 00:54:03.124 I would love to entertain any questions that you have.

00:54:03.124 --> 00:54:04.888 Let me just take one moment to acknowledge

00:54:04.888 --> 00:54:09.427 some of the people that I should acknowledge in this work,

00:54:09.427 --> 00:54:11.421 I already showed you a picture of John John who was earlier

00:54:11.421 --> 00:54:13.289 the the picture and the associated with the Mac ml approach

00:54:13.289 --> 00:54:15.317 that we developed many years ago 10 years ago basically

00:54:15.317 --> 00:54:17.635 Yinfei Wu has been taking the lead on this project.

00:54:17.635 --> 00:54:19.027 She's a master student.

00:54:19.027 --> 00:54:21.277 Yano os Wang was an assistant was in visiting

00:54:22.423 --> 00:54:24.204 Assistant Professor Stephen Gaugham,

00:54:24.204 --> 00:54:25.602 is in the Evie department

00:54:25.602 --> 00:54:27.587 has been helping out with this analysis.

00:54:27.587 --> 00:54:29.740 Haley Hassler is in my lab, has been helping out

00:54:29.740 --> 00:54:32.290 with phylogenetics Jayveer Singh is an undergrad

00:54:32.290 --> 00:54:35.030 who's been doing some of the research work

00:54:35.030 --> 00:54:37.188 some of the actually literature research

00:54:37.188 --> 00:54:38.540 that has helped us to contextualize

00:54:38.540 --> 00:54:40.910 the work we're doing Mofeed Najib

00:54:40.910 --> 00:54:43.760 produced those diagrams of the spike protein

00:54:43.760 --> 00:54:45.790 with the sites that we have identified

00:54:45.790 --> 00:54:47.323 as under selection so far,

00:54:48.380 --> 00:54:52.400 Zheng Wang is a long term collaborator of mine who works

00:54:53.683 --> 00:54:55.530 on nearly all the phylogenetic projects

00:54:55.530 --> 00:54:58.670 that I do, who's works with me.

00:54:58.670 --> 00:55:02.070 And then Alex Thornburg is A long term collaborator of mine,

00:55:02.070 --> 00:55:05.870 now in North Carolina.

00:55:05.870 --> 00:55:07.950 He was while he's currently at the North Carolina

00:55:07.950 --> 00:55:11.390 Museum of sciences, but he works on a lot of phylogenetic

00:55:11.390 --> 00:55:13.100 projects with me as well.

00:55:13.100 --> 00:55:15.610 And by the way, all of this, fortunately

00:55:15.610 --> 00:55:19.120 was recently awarded one of the NSF rapid grants

00:55:19.120 --> 00:55:20.060 to do this research.

00:55:20.060 --> 00:55:21.900 So we're very pleased to have funding to

00:55:21.900 --> 00:55:25.068 continue to work on this as time goes on, which is good

00:55:25.068 --> 00:55:26.530 because it's taking quite a lot of work

00:55:27.426 --> 00:55:28.283 to do the sequence wrangling.

00:55:29.286 --> 00:55:30.119 And the analyses themselves.

00:55:30.119 --> 00:55:32.190 As I mentioned, they're computationally intensive.

00:55:32.190 --> 00:55:34.660 So Alex and I were the PI's on that particular

00:55:35.721 --> 00:55:36.620 grant from the NSF.

00:55:36.620 --> 00:55:38.870 So we're excited to continue to do that work.

00:55:40.596 --> 00:55:41.901 And with that, I think I would

00:55:41.901 --> 00:55:42.773 like to entertain any questions you might have.

00:55:45.045 --> 00:55:46.745 - Thank you, Jeff, this was great.

00:55:47.617 --> 00:55:49.200 I'm sure we have a lot of questions

00:55:49.200 --> 00:55:50.563 who gets first?

00:55:54.490 --> 00:55:56.490 Again, you can type the questions on the

00:55:58.961 --> 00:56:00.794 chat box or just mute.

00:56:12.968 --> 00:56:14.100 - I have a quick question.

00:56:14.100 --> 00:56:15.764 - Okay.

00:56:15.764 --> 00:56:19.560 - You mentioned or you touched a bit on this before,

00:56:19.560 --> 00:56:23.600 but how would this compare to cite wise estimates

00:56:23.600 --> 00:56:25.500 of omega that you would get from Pamel

00:56:27.840 --> 00:56:28.673 or similar program?

00:56:28.673 --> 00:56:31.738 - So I'm sorry, I sort of was rushing at the end,

00:56:31.738 --> 00:56:34.792 I didn't explain that, in fact, I'm using pamel for some,

00:56:34.792 --> 00:56:36.169 So I'm using Pamela

00:56:36.169 --> 00:56:38.657 for the pre zoonosis analysis, and for the post zoonosis

00:56:39.615 --> 00:56:42.893 analysis, because as I mentioned during the talk,

00:56:43.734 --> 00:56:45.664 if you have a large phylogeny

00:56:45.664 --> 00:56:47.623 with multiple branches, et cetera, et cetera,

00:56:49.376 --> 00:56:50.209 where you can look over that entire phylogeny then you

00:56:51.360 --> 00:56:52.363 can get multiple changes at individual sites,

00:56:53.233 --> 00:56:55.130 which is what pamel actually uses to infer selection, right?

00:56:55.130 --> 00:56:57.170 You have to have the site change not just once

00:56:57.170 --> 00:56:59.393 but twice or three times.

00:57:01.713 --> 00:57:02.546 And then it says all that's under selection because

00:57:06.683 --> 00:57:10.350 it keeps changing again and again and again.

00:57:11.571 --> 00:57:12.959 So, so Pamela allows you to do that

00:57:12.959 --> 00:57:15.354 if you have this sort of deep time

00:57:15.354 --> 00:57:17.232 or large amount of time and multiple lineages that you're

00:57:17.232 --> 00:57:19.275 looking at, the master of approach that I'm using, enables

00:57:19.275 --> 00:57:22.170 you to do that on just a single lineage without needing

00:57:22.170 --> 00:57:23.203 multiple changes, I mean, multiple changes

00:57:23.203 --> 00:57:24.578 on a single language you can't even detect

00:57:24.578 --> 00:57:25.668 because it just looks like one change

00:57:25.668 --> 00:57:28.135 if you have the ancestral sequence, which is what we do

00:57:28.135 --> 00:57:30.634 ancestral data summation, get the ancestral sequence.

00:57:30.634 --> 00:57:33.227 And if you have the descendant sequence, a changes

00:57:33.227 --> 00:57:34.714 to T, you don't know if it changed to A to G to C to T again

00:57:34.714 --> 00:57:36.315 or if it just changed a to T, you have no idea you can

00:57:36.315 --> 00:57:38.047 just say it changed once.

00:57:38.047 --> 00:57:39.753 And so there's no real way to run pants,

00:57:39.753 --> 00:57:41.159 there is a way but it's really it's statistically

00:57:41.159 --> 00:57:41.992 really underpowered terrible thing

00:57:41.992 --> 00:57:44.164 to do to try to run pamel on a single lineage

00:57:44.164 --> 00:57:46.731 and figure out whether something's under selection.

00:57:46.731 --> 00:57:49.320 The advantage of this approach is because it

00:57:49.320 --> 00:57:51.382 can use that polymorphism data, the data of like what's

00:57:51.382 --> 00:57:54.072 just circulating in within populations as a metric for how

00:57:54.072 --> 00:57:55.888 much mutation is occurring.

00:57:55.888 --> 00:57:59.390 You can essentially divide out by that

00:57:59.390 --> 00:58:02.680 and then again, because we're integrating over all

00:58:03.544 --> 00:58:05.850 these models of how these things change, we're essentially

00:58:06.879 --> 00:58:08.930 borrowing information from neighboring sites for what their

00:58:10.488 --> 00:58:12.837 rates of change are, et cetera et cetera

00:58:12.837 --> 00:58:13.670 to estimate what the possible amount

00:58:14.770 --> 00:58:16.122 of selection is on all these sites.

00:58:16.122 --> 00:58:19.263 So by using the polymorphism data, and by doing this model

00:58:19.263 --> 00:58:21.445 averaging approach, we're actually able

00:58:21.445 --> 00:58:23.100 to take individual lineages and estimate

00:58:23.100 --> 00:58:25.050 the selection on them.

00:58:25.050 --> 00:58:28.880 And that's what we're doing in the near zonosis analysis

00:58:28.880 --> 00:58:30.730 that I showed you in the middle here.

00:58:32.610 --> 00:58:33.443 So there are different ways of doing the analysis.

00:58:34.924 --> 00:58:37.174 And it's necessitated by the fact that we just have this

00:58:37.174 --> 00:58:39.146 one lineage and there's no way it won't be a single lineage

00:58:39.146 --> 00:58:41.884 in any dataset we look at because for zoonosis,

00:58:41.884 --> 00:58:43.950 we're going to have human sequences,

00:58:43.950 --> 00:58:44.783 we're gonna have some animal sequences,

00:58:44.783 --> 00:58:47.722 we're not going to know we're not going

00:58:47.722 --> 00:58:50.010 to have any information about the actual zoonosis.

00:58:50.010 --> 00:58:51.600 Even if we knew the first human,

00:58:51.600 --> 00:58:54.011 we could just take that as an estimate.

00:58:54.011 --> 00:58:55.680 We still probably need some data here.

00:58:55.680 --> 00:58:57.970 Maybe you could have the first human

00:58:57.970 --> 00:58:59.910 and the first animal that you got it from.

00:58:59.910 --> 00:59:00.960 That just doesn't exist.

00:59:00.960 --> 00:59:03.500 We don't have that data for any zoonosis.

00:59:03.500 --> 00:59:06.690 How would we would never be there at the moment.

00:59:06.690 --> 00:59:08.710 So we have to assume that there's a number

00:59:08.710 --> 00:59:10.400 of transmissions among humans

00:59:10.400 --> 00:59:13.164 and a number of transmissions among animals

00:59:13.164 --> 00:59:14.090 during that near zoonotic period.

00:59:14.090 --> 00:59:15.600 And it's just a single lineage.

00:59:15.600 --> 00:59:17.513 So we can't really run pamel on that,

00:59:19.061 --> 00:59:21.095 in summary, because pamel requires multiple

00:59:21.095 --> 00:59:22.330 changes multiple lineages to have power

00:59:23.201 --> 00:59:24.730 to actually infer evolutionary change.

00:59:24.730 --> 00:59:26.640 MASS-PRF fortunatelY, can do that,

00:59:26.640 --> 00:59:28.450 because you can look on single lineages.

00:59:28.450 --> 00:59:31.270 So you can use MK tests as well on single lineage

00:59:32.533 --> 00:59:34.063 is basically designed to look at single lineages.

00:59:35.544 --> 00:59:36.523 But the problem with MK tests, as I mentioned,

00:59:37.371 --> 00:59:38.813 is that they're assuming the entire

00:59:38.813 --> 00:59:39.910 gene is under selection, which means it doesn't give you

00:59:41.071 --> 00:59:43.120 the scope or understanding about recombination

00:59:44.044 --> 00:59:46.088 binding gene sites under selection or something like that.

00:59:46.088 --> 00:59:47.440 It often will just give you a result of the genes not under

00:59:47.440 --> 00:59:49.023 selection, which is not true.

00:59:51.386 --> 00:59:52.219 - Does that answer your question?

00:59:53.599 --> 00:59:54.673 - Yes.

00:59:54.673 --> 00:59:55.506 - Great.

00:59:59.966 --> 01:00:01.799 - Any other questions?

01:00:03.691 --> 01:00:04.980 - I have one more if no one else wants to.

01:00:04.980 --> 01:00:06.690 - Sure, go ahead.

01:00:06.690 --> 01:00:10.480 - So in B cells, we have mechanisms

01:00:10.480 --> 01:00:12.560 that have mutation that specifically

01:00:12.560 --> 01:00:16.637 bias towards replacement mutations.

01:00:16.637 --> 01:00:18.350 So in the absence of selection,

01:00:18.350 --> 01:00:21.050 the mutation mechanisms actually cause

01:00:21.050 --> 01:00:22.533 an Omega greater than one.

01:00:24.270 --> 01:00:27.690 would this have any way of correcting for that?

01:00:27.690 --> 01:00:30.796 - So the tricky part is, and I don't know how it might,

01:00:30.796 --> 01:00:33.062 the tricky part is not so much running the software,

01:00:33.062 --> 01:00:37.310 which you could certainly do on that.

01:00:37.310 --> 01:00:38.900 The tricky part would be identifying

01:00:38.900 --> 01:00:43.000 what polymorphism is, in the case of those cells.

01:00:43.000 --> 01:00:47.000 So if you could identify sets of cells that are undergoing

01:00:47.000 --> 01:00:50.718 the mutation but aren't under selection in some way, then

01:00:50.718 --> 01:00:54.360 you could use that as the proxy for the way we use it here

01:00:54.360 --> 01:00:57.140 is polymorphism within population polymorphism,

01:00:57.140 --> 01:00:58.290 and then estimate that.

01:00:59.176 --> 01:01:01.235 I just don't know whether you have a way of

01:01:01.235 --> 01:01:02.068 doing Doing that if you want to discuss

01:01:02.917 --> 01:01:04.795 it with me, we could.

01:01:04.795 --> 01:01:06.803 That's sort of always the key for detecting selection.

01:01:09.279 --> 01:01:11.089 And it's, you know, many of you may be familiar that I work

01:01:11.089 --> 01:01:13.463 on cancer and some of the work that I do.

01:01:13.463 --> 01:01:14.546 It's the same

01:01:17.573 --> 01:01:20.593 problem that I'm working on there all the time, I'm trying

01:01:20.593 --> 01:01:23.196 to understand what the baseline mutation rates of cancer

01:01:23.196 --> 01:01:25.181 in cancer and somatic evolution of cells are.

01:01:25.181 --> 01:01:27.355 Because if I understand the baseline rates

01:01:27.355 --> 01:01:28.963 , how often those things change,

01:01:28.963 --> 01:01:29.878 just the mutation alone,

01:01:29.878 --> 01:01:31.722 then I can always estimate selection.

01:01:31.722 --> 01:01:34.292 And that's the thing we almost always want to

01:01:34.292 --> 01:01:37.258 know about in the analog analysis of sequence data.

01:01:37.258 --> 01:01:42.217 So, again, it's all about figuring out if there's some piece

01:01:42.217 --> 01:01:45.790 of the data that can be used to estimate that polymorphism

01:01:45.790 --> 01:01:47.863 and an approach like this, the benefit of an approach like

01:01:47.863 --> 01:01:50.126 this would be, you know, maybe you can estimate that for

01:01:50.126 --> 01:01:51.799 some portions of the gene, but not others, you know, maybe

01:01:51.799 --> 01:01:53.583 then there's a way that you could use this sort of model

01:01:53.583 --> 01:01:55.030 averaging approach to get at the underlying rate that it's

01:01:55.986 --> 01:01:56.819 happening, even if you can't estimate

01:01:58.111 --> 01:01:58.944 for that particular site, for instance.

01:02:00.284 --> 01:02:02.314 So I think the Might be potential to do it,

01:02:02.314 --> 01:02:04.408 but it just depends, you know, about on whether

01:02:04.408 --> 01:02:07.430 there's a critical, you know, set of data in what you're

01:02:08.990 --> 01:02:11.624 looking at which I haven't spent much time

01:02:11.624 --> 01:02:13.218 looking at back in the day.

01:02:13.218 --> 01:02:14.987 So I wouldn't know whether there's some way

01:02:14.987 --> 01:02:18.630 of baseline getting that baseline polymorphism or baseline

01:02:18.630 --> 01:02:21.633 mutation rate, which essentially amounts to the same thing.

01:02:22.545 --> 01:02:25.559 It just depends on whether, you know, you're assuming the

01:02:25.559 --> 01:02:28.901 population is sort of has, you know,

01:02:28.901 --> 01:02:31.231 it's just whether you're looking at at a population level,

01:02:31.231 --> 01:02:32.560 or you have some sort of covariance matrix

01:02:33.653 --> 01:02:35.063 to better understand the mutation rates itself.

01:02:36.180 --> 01:02:37.513 - I think there is a similar population B cells,

01:02:37.513 --> 01:02:41.233 - Great, so I encourage you to look into that.

01:02:44.150 --> 01:02:46.570 - Jeff, I have a quick question.

01:02:46.570 --> 01:02:49.600 I'm not too familiar with genome sequencing.

01:02:49.600 --> 01:02:52.510 But I think the Clustering Problem,

01:02:52.510 --> 01:02:55.330 the issue and the solution you have

01:02:55.330 --> 01:02:58.030 can be applied to many types of data.

01:02:58.030 --> 01:02:59.370 So I'm kind of confused.

01:02:59.370 --> 01:03:01.830 So you start In the diagram where you describe

01:03:01.830 --> 01:03:05.610 the different steps, you said that you first pick the most

01:03:05.610 --> 01:03:06.855 likely cluster and then you essentially

01:03:06.855 --> 01:03:09.305 keep splitting the clusters, right?

01:03:09.305 --> 01:03:11.551 How do you get the first clusters? Like

01:03:11.551 --> 01:03:16.168 there is some randomness in how you split the first?

01:03:16.168 --> 01:03:18.746 - Oh, so I sorry, I apologize.

01:03:18.746 --> 01:03:22.350 I didn't explain it in enough detail.

01:03:22.350 --> 01:03:24.380 The reason why it's so computationally intensive

01:03:24.380 --> 01:03:26.668 is we look at all possible.

01:03:26.668 --> 01:03:28.910 all possible exhaustedly.

01:03:28.910 --> 01:03:31.330 Now, I actually spent a year of my life trying

01:03:31.330 --> 01:03:34.070 to find a way to develop a Bayesian approach

01:03:34.070 --> 01:03:35.870 or some approach that would allow me

01:03:38.006 --> 01:03:39.880 to not look at all possible, you know, like to

01:03:39.880 --> 01:03:40.713 make this because because if you could do that,

01:03:40.713 --> 01:03:45.094 this would be a great way for doing tons of different things

01:03:45.094 --> 01:03:47.094 on very large data sets, right, large, like,

01:03:47.094 --> 01:03:50.200 and what amazed me is, I found that

01:03:50.200 --> 01:03:53.445 it was just an impenetrable problem.

01:03:53.445 --> 01:03:55.770 If I didn't look at every possible model.

01:03:55.770 --> 01:03:59.840 I could not get it to work I couldn't prove that

01:03:59.840 --> 01:04:02.563 That's Through like, I don't have any proof, that's true.

01:04:03.652 --> 01:04:05.183 And I would encourage anyone who really wants to dive

01:04:05.183 --> 01:04:06.016 in there, go ahead.

01:04:06.016 --> 01:04:06.970 But I'll warn you that I spent a year

01:04:06.970 --> 01:04:09.184 banging my head against that problem.

01:04:09.184 --> 01:04:10.275 And when I didn't

01:04:10.275 --> 01:04:11.882 exhaustively search all the models, I could not, I always

01:04:11.882 --> 01:04:15.534 caused these biases, like there was no way to sample them.

01:04:15.534 --> 01:04:17.217 I even have ways of sampling the models

01:04:17.217 --> 01:04:19.493 according to their probability.

01:04:23.767 --> 01:04:27.517 But even that causes a bias because sometimes

01:04:30.526 --> 01:04:31.359 there's a large number.

01:04:31.359 --> 01:04:33.693 So if you look at the, if you think

01:04:33.693 --> 01:04:35.415 about the set of models, it's a very large set of models.

01:04:35.415 --> 01:04:37.915 And there isn't actually a huge amount

01:04:37.915 --> 01:04:41.839 of likelihood differences between these models.

01:04:41.839 --> 01:04:43.256 That's the thing.

01:04:44.596 --> 01:04:49.497 So when you don't exhaustively sample the models,

01:04:49.497 --> 01:04:53.464 if you just sample some of the most likely models,

01:04:53.464 --> 01:04:55.728 you actually are sampling just

01:04:55.728 --> 01:04:57.137 one corner of the space.

01:04:57.137 --> 01:04:59.487 And it's possible for a bunch of

01:04:59.487 --> 01:05:00.320 not quite so likely models, but reasonable models

01:05:00.320 --> 01:05:02.747 that are not in that corner to sort of be actually

01:05:02.747 --> 01:05:03.830 highly influential on the model average.

01:05:03.830 --> 01:05:04.663 And so the bottom line is like sampling

01:05:04.663 --> 01:05:06.471 by trying to pick in the you know, most likely space doesn't

01:05:06.471 --> 01:05:07.430 work sampling by picking randomly doesn't work.

01:05:07.430 --> 01:05:08.939 And I could go into more detail about it.

01:05:08.939 --> 01:05:10.400 But it turned out that I couldn't do it

01:05:10.400 --> 01:05:11.641 any way other than exhaustive sampling.

01:05:11.641 --> 01:05:13.512 So, I say that Sorry, I missed that mistake.

01:05:13.512 --> 01:05:16.130 I couldn't do it by any biased approach

01:05:16.130 --> 01:05:18.152 towards that exhaustive handling

01:05:18.152 --> 01:05:19.413 the approach that I'm showing you right here.

01:05:20.546 --> 01:05:21.986 Actually, there are two ways of doing it.

01:05:21.986 --> 01:05:23.220 One is to sample stochastically,

01:05:23.220 --> 01:05:27.180 according to likelihood, and the other is to sample exactly

01:05:27.180 --> 01:05:30.210 across all exhausted sampling significantly works.

01:05:30.210 --> 01:05:32.662 In fact, it's implemented in the approach that I

01:05:32.662 --> 01:05:35.243 was just showing, I'm sorry, I just sort of jumped too fast

01:05:35.243 --> 01:05:36.877 to say what I was saying.

01:05:36.877 --> 01:05:38.169 So sampling stochastically works

01:05:38.169 --> 01:05:39.700 and sampling exhaustively work sampling stochastically is

01:05:39.700 --> 01:05:41.652 still very computationally intensive.

01:05:41.652 --> 01:05:44.204 But there's no I couldn't

01:05:44.204 --> 01:05:46.990 find any way to sort of, you know, important sample or do

01:05:48.264 --> 01:05:49.633 some sort of approach that would allow me to get a smaller

01:05:49.633 --> 01:05:52.616 set of models, which would then if we could do that,

01:05:52.616 --> 01:05:55.070 that could be really important,

01:05:55.070 --> 01:05:57.194 because then you could do this

01:05:57.194 --> 01:05:58.630 on more than like 2000 site,

01:05:58.630 --> 01:06:00.110 it's somewhere around 2000 sites.

01:06:00.110 --> 01:06:02.310 So you start running into real problems with

01:06:03.505 --> 01:06:04.850 just too much computing computation time

01:06:06.384 --> 01:06:07.228 to make it worthwhile.

01:06:07.228 --> 01:06:09.583 So we could extend this to 10,000 100,000, you know,

01:06:10.874 --> 01:06:12.990 potentially really, really large numbers of sites,

01:06:12.990 --> 01:06:15.650 and really, really sparse sets of sites.

01:06:15.650 --> 01:06:17.640 If only we could find a way

01:06:19.342 --> 01:06:22.142 to bias the sampling towards models that are more likely

01:06:24.040 --> 01:06:25.637 without causing biases in the results.

01:06:25.637 --> 01:06:26.470 I couldn't find any way to do.

01:06:27.370 --> 01:06:30.360 - This seems very much related to tree based

01:06:30.360 --> 01:06:34.360 methods where essentially you've got, like split the space

01:06:35.600 --> 01:06:38.073 and then you model of geology models,

01:06:38.966 --> 01:06:40.650 like the random forest, for example,

01:06:40.650 --> 01:06:43.213 or is very much related to that right.

01:06:45.447 --> 01:06:47.460 - Yeah, I have to say I was now familiar

01:06:47.460 --> 01:06:48.830 with those approaches.

01:06:48.830 --> 01:06:52.351 But when I was completely unfamiliar with it, yeah, I sort

01:06:52.351 --> 01:06:53.690 of thought about it that way.

01:06:53.690 --> 01:06:55.680 But you're absolutely right.

01:06:55.680 --> 01:06:57.250 Yeah, I guess the difference but here

01:06:57.250 --> 01:06:59.757 you have a sequence like one sequence,

01:06:59.757 --> 01:07:01.114 tghere you have a space.

01:07:01.114 --> 01:07:02.418 So you just split in

01:07:02.418 --> 01:07:04.888 different dimensions, but it is really good.

01:07:04.888 --> 01:07:09.888 - And I can mention, just to speculate,

01:07:10.170 --> 01:07:12.100 I'm kind of interested in a number of

01:07:13.390 --> 01:07:14.383 other ways of applying this.

01:07:15.349 --> 01:07:17.210 So for instance, if the one I've been thinking about

01:07:18.257 --> 01:07:19.754 and actually worked on a little

01:07:19.754 --> 01:07:20.739 bit haven't gotten very far with, but it's like,

01:07:20.739 --> 01:07:22.070 when you're dealing with event spaces over time,

01:07:22.070 --> 01:07:24.390 like if you have days, and you have individuals like,

01:07:24.390 --> 01:07:26.690 prominent us in public health,

01:07:26.690 --> 01:07:29.110 like individuals who are undergoing events

01:07:29.110 --> 01:07:31.180 you end up with a very sparse matrix of events.

01:07:31.180 --> 01:07:36.180 And so we use these approaches like survival plots

01:07:37.895 --> 01:07:40.096 all these approaches that we use to sort of understand

01:07:40.096 --> 01:07:40.929 how these rare events are happening,

01:07:42.161 --> 01:07:43.611 and how people are changing over this,

01:07:43.611 --> 01:07:45.100 that event space is actually really sparse.

01:07:45.100 --> 01:07:46.970 But it's kind of a matrix.

01:07:46.970 --> 01:07:48.380 And you could do this in two dimensions,

01:07:48.380 --> 01:07:49.360 not just one, right?

01:07:49.360 --> 01:07:51.590 So you could model average across two dimensions,

01:07:51.590 --> 01:07:53.472 and then you could get something

01:07:53.472 --> 01:07:55.030 that the thing that really appeals to me about that is that

01:07:55.030 --> 01:07:58.393 again, it's really this approach is really,

01:08:00.360 --> 01:08:04.427 it only builds up from the this binomial event

01:08:04.427 --> 01:08:08.540 No, no event, stuff, a picture that's very continuous over

01:08:08.540 --> 01:08:10.660 over the space and involves no assumptions

01:08:10.660 --> 01:08:12.310 about distribution whatsoever.

01:08:12.310 --> 01:08:14.180 So I'm just wondering if there aren't instances

01:08:14.180 --> 01:08:16.170 where, you know, we could come up

01:08:17.046 --> 01:08:18.500 with a better understanding of what's going on

01:08:18.500 --> 01:08:20.270 with individuals in a matrix such as

01:08:20.270 --> 01:08:22.090 that by using this approach.

01:08:22.090 --> 01:08:23.300 And it's an approach that is

01:08:23.300 --> 01:08:26.380 that still works even with these sparse spaces, because

01:08:26.380 --> 01:08:28.930 you can model average over these tremendously large number

01:08:28.930 --> 01:08:31.170 of models that all have fairly likely fairly

01:08:32.919 --> 01:08:33.752 equal likelihood to get a result.

01:08:34.883 --> 01:08:36.605 So I don't know that's just a sort of a

01:08:36.605 --> 01:08:37.603 speculation that there might be some interesting approaches

01:08:37.603 --> 01:08:41.031 , ways to approach those problems using this kind of kind

01:08:41.031 --> 01:08:43.903 of model averaging technique.

01:08:46.360 --> 01:08:48.870 - Great, I think we should wrap up.

01:08:48.870 --> 01:08:52.200 Thank you, Jeff, for this great presentation was great.

01:08:52.200 --> 01:08:54.843 And thank you all for joining today.

01:08:56.604 --> 01:08:57.930 See you next next seminar

01:08:57.930 --> 01:09:01.283 is gonna be I think, July 14.

01:09:01.283 --> 01:09:05.430 So we'll send out invites.

01:09:05.430 --> 01:09:07.331 All right, thank you, Jeff.

01:09:07.331 --> 01:09:08.223 Thank you all, bye, bye.