WEBVTT

00:00:03.160 --> 00:00:04.990 - First good afternoon, everyone,

00:00:04.990 --> 00:00:09.540 and I hope you somehow managed to enjoy your winter break

00:00:09.540 --> 00:00:11.360 you in this special time.

00:00:11.360 --> 00:00:16.310 And this is our first talk, seminar talk this semester,

00:00:16.310 --> 00:00:18.600 and we have invited Dr. Eugene Katsevich

00:00:18.600 --> 00:00:22.300 from Wharton School at UPenn.

00:00:22.300 --> 00:00:26.340 And he's going to present something really exciting,

00:00:26.340 --> 00:00:30.460 I know his original work on statistical analysis

00:00:32.284 --> 00:00:34.440 single cell CRISPR screening.

00:00:34.440 --> 00:00:39.440 And I will hand it over to Eugene from now, from here.

00:00:39.810 --> 00:00:42.840 And, but if Eugene wanted to start or wait one

00:00:42.840 --> 00:00:45.253 or two minutes to start, it's up to you.

00:00:46.190 --> 00:00:49.433 - Yeah maybe, I mean, yeah, I don't know.

00:00:50.910 --> 00:00:53.200 If people will filter in, maybe I'll wait another minute

00:00:53.200 --> 00:00:56.590 or two, 'cause I think, I feel like the first part

00:00:56.590 --> 00:00:58.180 of the talk is very important.

00:00:58.180 --> 00:01:01.040 So I think if people missed the first part of the talk,

00:01:01.040 --> 00:01:03.920 then it'll be maybe hard to follow along later.

00:01:03.920 --> 00:01:08.920 So I'm happy to wait just another minute or two.

00:01:09.440 --> 00:01:12.450 I understand perfectly that it's a strange time

00:01:12.450 --> 00:01:15.180 for everyone, so for all those who were able

00:01:15.180 --> 00:01:17.310 to make it today, I really appreciate

00:01:17.310 --> 00:01:20.223 your adjusting the schedule.

00:01:22.270 --> 00:01:25.390 Also maybe one remark I can make is that,

00:01:25.390 --> 00:01:27.400 since it is a smaller audience,

00:01:27.400 --> 00:01:30.790 I think we can make this seminar just about

00:01:30.790 --> 00:01:32.420 as interactive as you want.

00:01:32.420 --> 00:01:36.960 So you should definitely feel free to stop me

00:01:36.960 --> 00:01:37.923 at any point.

00:01:38.870 --> 00:01:40.370 I don't know how many of you are familiar

00:01:40.370 --> 00:01:43.440 with the CRISPR screen stuff I'm gonna talk about,

00:01:43.440 --> 00:01:47.533 but I'm very happy to just make it very interactive.

00:01:54.285 --> 00:01:58.367 I will maybe start sharing my screen

00:01:58.367 --> 00:02:00.040 and maybe I'll start launching

00:02:00.040 --> 00:02:04.023 into some of the introductory things.

00:02:12.730 --> 00:02:13.900 So...

00:02:16.270 --> 00:02:18.123 Oh wow, wait, is this the...

00:02:20.660 --> 00:02:23.910 I greatly apologize.

00:02:23.910 --> 00:02:28.300 Clearly, the label on my slides is wrong.

00:02:28.300 --> 00:02:30.520 I have updated my slides since then,

00:02:30.520 --> 00:02:32.730 but I think the title page has not been updated,

00:02:32.730 --> 00:02:34.503 that's extremely embarrassing.

00:02:39.807 --> 00:02:43.360 Well, maybe then I should skip past this slide very quickly.

00:02:43.360 --> 00:02:47.260 So hello everyone, thank you so much

00:02:47.260 --> 00:02:48.723 for making it to my talk.

00:02:49.760 --> 00:02:52.440 Today, I'll be talking about some Statistical Analysis Tools

00:02:52.440 --> 00:02:54.590 for Single Cell CRISPR Screens.

00:02:54.590 --> 00:02:56.240 So the most important thing to take away

00:02:56.240 --> 00:02:58.890 from this slide are my collaborators here.

00:02:58.890 --> 00:03:02.100 So Tim Barry is a grad student

00:03:02.100 --> 00:03:03.723 of mine who was actually at CMU.

00:03:05.340 --> 00:03:07.670 I am jointly advising him with Kathryne Roeder

00:03:07.670 --> 00:03:12.670 also at CMU, who is used to be my postdoc advisor.

00:03:15.060 --> 00:03:17.163 So I'll skip quickly to the next slide.

00:03:20.000 --> 00:03:22.060 So here's the motivation.

00:03:22.060 --> 00:03:25.160 And by the way, if anyone has joined recently,

00:03:25.160 --> 00:03:27.853 please just stop me at any point.

00:03:28.990 --> 00:03:29.980 So here's the motivation.

00:03:29.980 --> 00:03:31.490 So we have done lots

00:03:31.490 --> 00:03:34.540 and lots of genome wide association studies to date.

00:03:34.540 --> 00:03:37.100 So we have a lot of little markers

00:03:37.100 --> 00:03:41.430 along the genome that we think are associated with diseases.

00:03:41.430 --> 00:03:43.190 And so the question is what's the next step?

00:03:43.190 --> 00:03:45.960 Like how do we actually translate these

00:03:45.960 --> 00:03:48.750 into insights into diseases?

00:03:48.750 --> 00:03:50.800 And hopefully later on things like,

00:03:50.800 --> 00:03:52.040 therapeutics and so on.

00:03:52.040 --> 00:03:55.680 So what we need to do is we need to understand how

00:03:55.680 --> 00:03:57.690 like basically the mechanisms

00:03:57.690 --> 00:04:00.870 by what mechanism are these associations actually resulting

00:04:00.870 --> 00:04:02.640 in an increased disease risk.

00:04:02.640 --> 00:04:04.710 So here's a typical situation here

00:04:04.710 --> 00:04:07.760 as our genome and here's a disease association

00:04:07.760 --> 00:04:10.660 and frequently these disease associations

00:04:10.660 --> 00:04:13.330 they might not take place within genes.

00:04:13.330 --> 00:04:16.103 And so that makes them pretty hard to interpret.

00:04:17.500 --> 00:04:22.500 So what's hypothesized to be the case here is that instead

00:04:23.560 --> 00:04:28.560 of disrupting genes directly, these variants

00:04:29.300 --> 00:04:33.330 are disrupting regulatory elements such as enhancers.

00:04:33.330 --> 00:04:36.090 So let's just like briefly here review

00:04:37.503 --> 00:04:40.800 that an enhancer is a region of the genome.

00:04:40.800 --> 00:04:42.330 That could be a certain distance

00:04:42.330 --> 00:04:45.160 from the gene that actually folds

00:04:45.160 --> 00:04:46.970 in three-dimensional space to come

00:04:46.970 --> 00:04:51.970 in close proximity to the promoter of the gene.

00:04:52.250 --> 00:04:54.930 And essentially the enhancers job is to recruit a lot

00:04:54.930 --> 00:04:56.610 of the machinery that actually is going to lead

00:04:56.610 --> 00:04:58.450 to the expression of this gene.

00:04:58.450 --> 00:05:00.450 So if you disrupt the enhancer

00:05:00.450 --> 00:05:02.150 then this will disrupt the recruitment

00:05:02.150 --> 00:05:04.420 of all of these different transcription factors

00:05:04.420 --> 00:05:08.763 which will then end up causing some trouble.

00:05:09.700 --> 00:05:14.080 And so it's this sort of like, for example, in this case

00:05:14.920 --> 00:05:16.610 let's say that this disease association

00:05:16.610 --> 00:05:20.850 as is disrupting enhanced or one, well, this might suggest

00:05:20.850 --> 00:05:24.370 if enhancer one is regulating gene two, that

00:05:24.370 --> 00:05:28.150 the disease mechanism is actually proceeding essentially

00:05:28.150 --> 00:05:30.653 or being mediated by the expression of gene too.

00:05:32.120 --> 00:05:37.120 And so this would be a very great

00:05:37.760 --> 00:05:40.600 and clean way of interpreting GWAS hits.

00:05:40.600 --> 00:05:44.320 But the problem is that we don't actually know

00:05:44.320 --> 00:05:45.960 or we have a very hazy sense

00:05:45.960 --> 00:05:49.640 of which enhancers actually regulate which genes.

00:05:49.640 --> 00:05:52.100 So this is kind of a difficult problem

00:05:52.100 --> 00:05:53.690 for a few different reasons.

00:05:53.690 --> 00:05:56.760 The first reason is that there's a potentially

00:05:56.760 --> 00:06:00.160 many to many mapping between enhancers in genes.

00:06:00.160 --> 00:06:03.430 So in enhancer it can regulate multiple genes

00:06:03.430 --> 00:06:07.540 and a single gene can be regulated by multiple enhancers.

00:06:07.540 --> 00:06:10.270 So the other thing is that any answers don't even

00:06:10.270 --> 00:06:13.320 need to be all too close to the genes that they regulate.

00:06:13.320 --> 00:06:17.690 There could be situations like we saw here where

00:06:17.690 --> 00:06:19.970 the regulation can skip the adjacent gene

00:06:19.970 --> 00:06:21.610 and go to the next one.

00:06:21.610 --> 00:06:24.150 And so in general regulations can

00:06:24.150 --> 00:06:27.930 are thought to happen within about a megabase distance

00:06:29.860 --> 00:06:32.870 in terms of the linear distance in the genome.

00:06:32.870 --> 00:06:36.390 So this is a hard problem, and it's basically

00:06:36.390 --> 00:06:38.330 the motivating problem for this talk

00:06:38.330 --> 00:06:40.650 which enhancers regulate which genes.

00:06:40.650 --> 00:06:42.070 This is a sort

00:06:42.070 --> 00:06:45.033 of a very fundamental and important problem in genomics.

00:06:46.800 --> 00:06:51.800 So in today's talk, I'm going to first talk about

00:06:52.720 --> 00:06:56.200 a new assay called a single cell CRISPR screen

00:06:56.200 --> 00:06:59.213 that allows us to get at this question,

00:07:02.780 --> 00:07:05.520 then I'm gonna talk about the challenges

00:07:05.520 --> 00:07:07.690 that previous methods have encountered

00:07:07.690 --> 00:07:10.160 in analyzing these single cell CRISPR screen

00:07:10.160 --> 00:07:14.350 datasets, never propose a new methodology based

00:07:14.350 --> 00:07:16.393 on this idea of conditional resampling.

00:07:17.650 --> 00:07:20.320 And then I will show you how this works

00:07:20.320 --> 00:07:22.583 on real data and close with the discussion.

00:07:25.450 --> 00:07:28.210 So let me first introduce the biological assay here

00:07:28.210 --> 00:07:31.140 which is called the Single Cell CRISPR screen.

00:07:31.140 --> 00:07:34.070 So actually backing up a second,

00:07:34.070 --> 00:07:35.230 this is a very important problem

00:07:35.230 --> 00:07:36.820 and people have considered it before.

00:07:36.820 --> 00:07:40.003 So how do people typically approach gene-enhancer mapping?

00:07:41.230 --> 00:07:46.100 I think the most common approach is what I call here

00:07:46.100 --> 00:07:48.220 an indirect observational approach.

00:07:48.220 --> 00:07:49.600 And there are many of these.

00:07:49.600 --> 00:07:50.690 So what this picture is,

00:07:50.690 --> 00:07:53.640 is a basically a more detailed picture

00:07:53.640 --> 00:07:56.840 of what happens when an enhancer or a pictured here comes

00:07:56.840 --> 00:08:00.080 into contact with the promoter of a gene.

00:08:00.080 --> 00:08:01.320 There are lots of kind

00:08:01.320 --> 00:08:05.510 of indirect signals of this regulation.

00:08:05.510 --> 00:08:08.027 Obviously you have just the actual expression

00:08:08.027 --> 00:08:12.310 of the gene, but you'll have the confirmation

00:08:12.310 --> 00:08:16.310 of the chromatin in the vicinity of the promoter

00:08:16.310 --> 00:08:18.060 and in the enhancer

00:08:18.060 --> 00:08:21.890 you have basically transcription factor binding data.

00:08:21.890 --> 00:08:25.320 And all of these data are essentially indirect ways

00:08:25.320 --> 00:08:28.490 of trying to make a conclusion

00:08:28.490 --> 00:08:31.220 about which enhancers might be regulating which genes.

00:08:31.220 --> 00:08:33.147 So for example, using high C data

00:08:33.147 --> 00:08:36.520 if you find an enhancer to be a 3D contact

00:08:36.520 --> 00:08:39.235 with the promoter, then this could be a single signal

00:08:39.235 --> 00:08:41.973 that there is some regulation going on.

00:08:43.350 --> 00:08:45.030 The issue is that these approaches have not

00:08:45.030 --> 00:08:47.420 proved very reliable at the end of the day.

00:08:47.420 --> 00:08:49.040 These are observational approaches,

00:08:49.040 --> 00:08:51.450 and basically even if you have

00:08:52.370 --> 00:08:55.593 contact in 3D space, this is not necessarily a signal.

00:08:56.560 --> 00:08:58.540 This doesn't necessarily mean that regulation

00:08:58.540 --> 00:08:59.500 is actually occurring,

00:08:59.500 --> 00:09:03.630 and so essentially we haven't gotten all too far

00:09:03.630 --> 00:09:05.490 with these indirect approaches.

00:09:05.490 --> 00:09:07.020 So the exciting thing is

00:09:07.020 --> 00:09:12.020 that recently with the development of CRISPR technology

00:09:12.700 --> 00:09:17.070 we can now actually go in and instead of observationally

00:09:17.070 --> 00:09:19.700 just essentially take a look inside a cell.

00:09:19.700 --> 00:09:23.690 We can actually go in and make modifications where we

00:09:23.690 --> 00:09:27.900 for example, knockouts enhancers using the system

00:09:27.900 --> 00:09:29.770 called CRISPR Interference.

00:09:29.770 --> 00:09:30.820 And then we try to look

00:09:30.820 --> 00:09:33.053 at what the results are for gene expression.

00:09:33.990 --> 00:09:37.630 So this shows you a little cartoon

00:09:37.630 --> 00:09:40.070 of the CRISPR interference system.

00:09:40.070 --> 00:09:41.510 And so the way that it works is

00:09:41.510 --> 00:09:46.510 that you have this CAS nine protein whose job is to attach

00:09:48.270 --> 00:09:50.530 to a certain segment of DNA.

00:09:50.530 --> 00:09:52.530 And the specific segment of DNA it attaches

00:09:52.530 --> 00:09:56.860 to is specified by this guide, or I do.

00:09:56.860 --> 00:09:58.550 And so in this way,

00:09:58.550 --> 00:10:00.720 the attachment can be highly specific

00:10:00.720 --> 00:10:03.730 to the sequence of the enhancer.

00:10:03.730 --> 00:10:06.950 And then this for CRISPR Interference

00:10:06.950 --> 00:10:08.890 the CAS nine brings along with it

00:10:08.890 --> 00:10:13.200 all of these repressive elements that essentially knock

00:10:13.200 --> 00:10:17.460 out this enhancer, meaning they prevent the enhancement

00:10:17.460 --> 00:10:20.513 from actually helping to regulate this gene.

00:10:21.690 --> 00:10:24.260 And so the idea, so firstly

00:10:24.260 --> 00:10:27.880 this is a promising solution because it allows us to

00:10:27.880 --> 00:10:30.330 interrogate these regulatory relationships

00:10:30.330 --> 00:10:31.680 in a much more direct way

00:10:31.680 --> 00:10:34.890 than we've been able to do until recently.

00:10:34.890 --> 00:10:38.620 And so the overall idea is that,

00:10:38.620 --> 00:10:40.700 it's the idea of simple disrupt enhancers

00:10:40.700 --> 00:10:43.640 and see which genes expression drops.

00:10:43.640 --> 00:10:46.450 And so just as a cartoon here, let's say we knock

00:10:46.450 --> 00:10:48.670 out this enhancer, then we would expect

00:10:48.670 --> 00:10:53.023 to see the gene that regulates to be down-regulated.

00:10:53.870 --> 00:10:54.770 And then we can think

00:10:54.770 --> 00:10:58.200 about designing perturbations for multiple enhancers.

00:10:58.200 --> 00:11:00.460 And so if you perturb this enhancer

00:11:00.460 --> 00:11:04.193 then maybe you'll see a response in these two genes.

00:11:07.120 --> 00:11:09.460 - Very naive question, just to make sure I

00:11:09.460 --> 00:11:13.607 didn't misunderstand notion here is enhancer always

00:11:13.607 --> 00:11:16.423 upregulating gene kind of regulate?

00:11:17.950 --> 00:11:20.020 - I think enhancers specifically

00:11:20.020 --> 00:11:22.480 are thought to upregulate genes.

00:11:22.480 --> 00:11:24.800 However, it's a good question because there are other kinds

00:11:24.800 --> 00:11:28.450 of elements that are, can actually be silencers for example.

00:11:28.450 --> 00:11:31.550 And so that's just another example of a kind

00:11:31.550 --> 00:11:33.464 of a regulatory element.

00:11:33.464 --> 00:11:35.880 So the effect could go in either direction

00:11:35.880 --> 00:11:38.160 and this talk I'll primarily talk about enhancers

00:11:38.160 --> 00:11:41.610 but really everything I say goes through for other kinds

00:11:41.610 --> 00:11:44.060 of regulatory elements.

00:11:44.060 --> 00:11:44.910 - Thanks.

00:11:44.910 --> 00:11:47.283 - Yeah, very good question.

00:11:50.150 --> 00:11:54.290 So now the actual assay

00:11:54.290 --> 00:11:58.730 That allows you to do this out of large scale.

00:11:58.730 --> 00:12:00.130 So the scale is the question here

00:12:00.130 --> 00:12:02.780 because you can do CRISPR experiments where

00:12:02.780 --> 00:12:05.310 you essentially like knock out one enhancer

00:12:05.310 --> 00:12:07.870 in a whole batch of cells, and then,

00:12:07.870 --> 00:12:09.870 maybe go enhancer by enhancer

00:12:09.870 --> 00:12:12.630 and this ends up not being a very scalable approach.

00:12:12.630 --> 00:12:15.300 So there has been proposed

00:12:16.358 --> 00:12:18.930 this new asset called the single cell CRISPR screen

00:12:18.930 --> 00:12:21.550 in which you basically pool a whole bunch

00:12:21.550 --> 00:12:23.350 of perturbations together,

00:12:23.350 --> 00:12:26.330 and then the readout that you get is single cell

00:12:26.330 --> 00:12:29.640 RNA sequencing, which allows you to also basically look

00:12:29.640 --> 00:12:30.750 at the impact of all

00:12:30.750 --> 00:12:32.480 of those different enhancement perturbations

00:12:32.480 --> 00:12:34.540 on the entire transcriptome.

00:12:34.540 --> 00:12:37.570 And so in the slide, I'm gonna give you a brief overview

00:12:37.570 --> 00:12:40.320 of how these screens work.

00:12:40.320 --> 00:12:41.600 So first way you do is you start

00:12:41.600 --> 00:12:44.260 with a library of CRISPR perturbations.

00:12:44.260 --> 00:12:47.583 So you just, let's say maybe you take,

00:12:49.227 --> 00:12:51.760 10,000 enhancers across the genome

00:12:51.760 --> 00:12:54.613 and then you basically design CRISPR guide.

00:12:54.613 --> 00:12:56.893 RNAs targeting each of those enhancers.

00:12:57.860 --> 00:13:00.190 Once you have a library of these perturbations

00:13:00.190 --> 00:13:03.130 you then infect a big pool

00:13:03.130 --> 00:13:05.910 of cells with all of these perturbations.

00:13:05.910 --> 00:13:07.640 And so what's important to note here is

00:13:07.640 --> 00:13:12.640 that essentially these perturbations get randomly integrated

00:13:13.360 --> 00:13:17.280 into the different cells they're delivered through a

00:13:17.280 --> 00:13:18.734 like a virus system

00:13:18.734 --> 00:13:21.880 the details aren't very important, but the importance is

00:13:21.880 --> 00:13:23.780 that these perturbations get integrated

00:13:23.780 --> 00:13:25.720 into cells essentially at random.

00:13:25.720 --> 00:13:27.690 And so each cell gets its own collection

00:13:27.690 --> 00:13:30.040 of CRISPR perturbations.

00:13:30.040 --> 00:13:34.760 So now in order to basically actually read out what happened

00:13:34.760 --> 00:13:37.640 in our experiment, we use single cell RNA sequencing.

00:13:37.640 --> 00:13:39.800 And as a result of the sequencing experiment

00:13:39.800 --> 00:13:44.070 we get two pieces of information, firstly, by the way

00:13:44.070 --> 00:13:46.270 two pieces of information for every step.

00:13:46.270 --> 00:13:47.400 So for every cell

00:13:47.400 --> 00:13:49.900 we first measure the perturbations that are present.

00:13:49.900 --> 00:13:52.240 So which of these guide or nays did we detect,

00:13:52.240 --> 00:13:53.120 and then secondly

00:13:53.120 --> 00:13:57.000 the gene expression for the whole transcriptome.

00:13:57.000 --> 00:13:59.300 So this is essentially our data here.

00:13:59.300 --> 00:14:01.000 And then once we have this data

00:14:01.000 --> 00:14:05.460 we can now do the analysis component, which really ends

00:14:05.460 --> 00:14:08.360 up being a kind of differential expression analysis.

00:14:08.360 --> 00:14:11.970 So consider a particular gene-enhancer pair.

00:14:11.970 --> 00:14:15.410 So what we can do is we could take all of the cells

00:14:15.410 --> 00:14:17.370 and we can break them up into two groups.

00:14:17.370 --> 00:14:20.360 Those cells for which that enhancer was knocked out

00:14:20.360 --> 00:14:23.330 which are in orange here, and those cells

00:14:23.330 --> 00:14:25.620 for which that enhancer was not knocked out.

00:14:25.620 --> 00:14:29.360 We can then split, essentially look

00:14:29.360 --> 00:14:32.500 at the expression of the gene of interest

00:14:32.500 --> 00:14:34.680 and see whether there's a systematic difference

00:14:34.680 --> 00:14:36.500 between the expression of this gene

00:14:36.500 --> 00:14:39.000 and these two populations of cells.

00:14:39.000 --> 00:14:42.740 So, and then if there is a significant difference

00:14:42.740 --> 00:14:44.850 then we can make a conclusion that that particular

00:14:44.850 --> 00:14:47.363 enhancer is regulating that particular gene.

00:14:48.440 --> 00:14:51.700 So it seems quite simple on first glance,

00:14:51.700 --> 00:14:55.240 but this analysis part actually turns

00:14:55.240 --> 00:14:59.280 out to be a challenging statistical problem.

00:14:59.280 --> 00:15:01.830 And so the analysis

00:15:01.830 --> 00:15:05.623 of these screens is actually the subject of this talk.

00:15:06.870 --> 00:15:09.800 Okay so, maybe one more slide

00:15:09.800 --> 00:15:12.300 and then I'll stop and see if people have questions.

00:15:12.300 --> 00:15:14.550 So just to make it a little bit more concrete

00:15:15.730 --> 00:15:18.560 there's a kind of a large data set that might be one

00:15:18.560 --> 00:15:21.770 of the largest out there right now by Gasperini at all.

00:15:21.770 --> 00:15:24.230 It was published in cell last year.

00:15:24.230 --> 00:15:27.280 Oh wow, I guess two years ago now to 2019,

00:15:27.280 --> 00:15:31.050 and so they were working with 200,000 K five 62 cells

00:15:31.050 --> 00:15:33.910 and they were looking at 6,000 candidate enhancers.

00:15:33.910 --> 00:15:35.460 And so they're looking at, I mean

00:15:35.460 --> 00:15:37.670 essentially the whole transcriptome, at least the part

00:15:37.670 --> 00:15:41.090 of it that has any expression in the cell type.

00:15:41.090 --> 00:15:45.460 And they identified 85,000 enhancer gene pairs

00:15:45.460 --> 00:15:49.600 that they essentially thought were plausible

00:15:49.600 --> 00:15:53.090 to have some regulation and in their experiment

00:15:53.090 --> 00:15:57.790 they had 28 per patients on average per cell.

00:15:57.790 --> 00:16:00.770 And so the way that this data would look is, think

00:16:00.770 --> 00:16:04.710 about the rows as being the cells and then the columns.

00:16:04.710 --> 00:16:05.960 So you have two groups of columns.

00:16:05.960 --> 00:16:07.910 Firstly, you have the gene expressions,

00:16:07.910 --> 00:16:10.440 and so since these are single cell data

00:16:10.440 --> 00:16:12.610 we have these highly discreet counts

00:16:12.610 --> 00:16:17.403 of reeds or UMRs for every gene.

00:16:18.470 --> 00:16:20.900 And then also we have the second bit of information

00:16:20.900 --> 00:16:22.920 which is a binary matrix, which tells you

00:16:22.920 --> 00:16:27.180 which cells received, which perturbations.

00:16:27.180 --> 00:16:29.490 So in general, in this presentation, I'll talk

00:16:29.490 --> 00:16:34.353 I'll denote gene expression by Y and perturbations by X.

00:16:35.630 --> 00:16:37.720 And so there's also a third and very important piece

00:16:37.720 --> 00:16:41.510 of information, which are technical factors per cell.

00:16:41.510 --> 00:16:43.410 Perhaps the main one that I'll talk

00:16:43.410 --> 00:16:46.160 about today is the sequencing depth.

00:16:46.160 --> 00:16:47.670 So this is just the total number

00:16:47.670 --> 00:16:52.670 of reads or UMRs I measured from this cell.

00:16:52.690 --> 00:16:56.650 And so this basically just varies randomly across cells

00:16:56.650 --> 00:16:57.950 just as an artifact of your experiment.

00:16:57.950 --> 00:16:59.760 There are other technical factors

00:16:59.760 --> 00:17:01.993 like batch and so on and so forth.

00:17:03.200 --> 00:17:05.510 Okay, so this brings me to the end

00:17:05.510 --> 00:17:07.600 of the first section where I tell you

00:17:07.600 --> 00:17:10.540 about the data and the asset.

00:17:10.540 --> 00:17:13.930 So are there any questions before I move on

00:17:13.930 --> 00:17:18.653 to talking more about the analysis of these types of data.

00:17:25.400 --> 00:17:27.100 I'm assuming there are no questions

00:17:27.100 --> 00:17:30.213 but do feel free to stop me if there are.

00:17:34.028 --> 00:17:36.900 So as I said, this actually turns out to be kind of

00:17:36.900 --> 00:17:40.270 like an annoyingly challenging statistical problem.

00:17:40.270 --> 00:17:43.210 And so to illustrate this to you, let me first

00:17:43.210 --> 00:17:45.180 give you a sense of what analysis methods there

00:17:45.180 --> 00:17:47.080 are out there.

00:17:47.080 --> 00:17:51.920 I should say, by the way that given the sort of the novelty

00:17:51.920 --> 00:17:55.220 of this assay, there hasn't been a lot of work in terms

00:17:55.220 --> 00:17:57.800 of designing methods specifically designed

00:17:58.920 --> 00:18:00.610 for this kind of data.

00:18:00.610 --> 00:18:04.260 So most of the existing analysis methods are basically

00:18:04.260 --> 00:18:06.780 proposed by the same people who are

00:18:06.780 --> 00:18:09.163 producing the single cell CRISPR screen data.

00:18:10.460 --> 00:18:13.580 So by the way, so in this slide

00:18:13.580 --> 00:18:16.870 I'm going to it actually for the remainder of the talk

00:18:16.870 --> 00:18:20.480 I'm actually going to essentially focus our attention

00:18:20.480 --> 00:18:24.610 on a certain gene and a certain enhancer

00:18:24.610 --> 00:18:26.830 and just consider the problem

00:18:26.830 --> 00:18:30.180 and figuring out whether that enhancer regulates that gene.

00:18:30.180 --> 00:18:33.750 And so I'm gonna use YI, to denote the expression

00:18:33.750 --> 00:18:37.970 of that gene and cell I XI as the binary indicator

00:18:37.970 --> 00:18:41.020 for whether that enhancer was perturbed in that cell

00:18:41.020 --> 00:18:46.020 and ZI the vector of these extra technical co-variants.

00:18:47.590 --> 00:18:51.810 So With that notation out of the way,

00:18:51.810 --> 00:18:56.810 the first kind of popular method for analyzing these data

00:18:57.830 --> 00:18:59.730 is negative binomial regression.

00:18:59.730 --> 00:19:03.540 For those of you familiar with bulk RNA-seq differential

00:19:03.540 --> 00:19:07.820 expression analysis, this is similar to the DESeq2

00:19:07.820 --> 00:19:11.290 methodology where you just run a negative binomial

00:19:11.290 --> 00:19:16.290 regression of the gene expression, Y on a linear combination

00:19:16.830 --> 00:19:20.110 of the perturbation indicator, as well as all

00:19:20.110 --> 00:19:21.833 of your technical co-variants.

00:19:23.490 --> 00:19:27.830 And so Negative Binomial is a common model for these sort of

00:19:27.830 --> 00:19:30.710 over dispersed count data that you encounter

00:19:30.710 --> 00:19:33.213 in RNA sequencing data.

00:19:35.040 --> 00:19:38.330 Okay, next, there is a rank based approach.

00:19:38.330 --> 00:19:43.080 So this is non-parametric where it's actually much simpler.

00:19:43.080 --> 00:19:47.870 You just, you cross tabulate yourselves by two criteria.

00:19:47.870 --> 00:19:50.590 First, you see whether they have the perturbation or not.

00:19:50.590 --> 00:19:55.000 And second, you see whether they have essentially higher

00:19:55.000 --> 00:19:57.230 than median expression on this gene or lower

00:19:57.230 --> 00:19:59.090 than median expression on this gene.

00:19:59.090 --> 00:20:03.283 And then you do a two by two table test for independence.

00:20:04.900 --> 00:20:08.460 And finally there are also permutation based approaches

00:20:08.460 --> 00:20:12.230 where the idea is to take some test statistic

00:20:12.230 --> 00:20:16.330 and then calibrate it under the null distribution

00:20:16.330 --> 00:20:19.210 by permuting this column right here

00:20:19.210 --> 00:20:22.843 the assignments of the perturbations to the cells.

00:20:24.290 --> 00:20:27.800 So yes, that, I guess that's, what's written here.

00:20:27.800 --> 00:20:32.800 So okay, there's like maybe all these methods sound

00:20:35.510 --> 00:20:39.150 reasonable at first, but the more you actually look

00:20:39.150 --> 00:20:40.440 at the existing literature

00:20:40.440 --> 00:20:42.650 the more there are various scattered signs

00:20:42.650 --> 00:20:47.340 like none of these methods are like really doing the trick.

00:20:47.340 --> 00:20:48.880 And so here are

00:20:49.880 --> 00:20:53.500 the methods that I described on the previous slide.

00:20:53.500 --> 00:20:54.620 I don't know if I named them

00:20:54.620 --> 00:20:57.060 but so virtual FACS is the rank based one

00:20:57.060 --> 00:21:00.590 and scMAGeCK is the one of the permutation based ones.

00:21:00.590 --> 00:21:05.261 And so you look at plots actually from

00:21:05.261 --> 00:21:09.550 the original papers themselves who propose these methods

00:21:09.550 --> 00:21:12.550 and you see some signs of miscalibration.

00:21:12.550 --> 00:21:15.760 And so like, for example, I'm gonna be talking mostly

00:21:15.760 --> 00:21:18.110 about this data and to a lesser extent

00:21:18.110 --> 00:21:22.640 about this data in my talk, but so looking here

00:21:22.640 --> 00:21:24.300 so I guess perhaps I should first talk

00:21:24.300 --> 00:21:27.230 about the concept of a Negative Control Perturbation.

00:21:27.230 --> 00:21:29.760 So a Negative Control Perturbation is a guide

00:21:29.760 --> 00:21:33.410 or but it's actually not designed to

00:21:33.410 --> 00:21:36.570 target any particular sequence along the genome.

00:21:36.570 --> 00:21:39.560 So you don't expect cells that are infected

00:21:39.560 --> 00:21:41.920 with a negative control perturbation to look any different

00:21:41.920 --> 00:21:45.980 from cells that have no perturbation.

00:21:45.980 --> 00:21:49.920 And so in this Gasperini data

00:21:49.920 --> 00:21:53.010 they have 50 different negative control guide RNAs,

00:21:53.010 --> 00:21:57.300 and so what they did is they basically plotted a QQ plot

00:21:57.300 --> 00:22:00.210 of all of the negative control guide RNAs,

00:22:00.210 --> 00:22:05.210 paired with all of the genes and the genome,

00:22:05.730 --> 00:22:08.800 and what they found is and perhaps on this QQ plot

00:22:08.800 --> 00:22:12.490 this doesn't look like a severe inflation from uniformity

00:22:12.490 --> 00:22:17.000 but it's important to keep in mind the scale of this Y axis.

00:22:17.000 --> 00:22:21.240 And so essentially this amounts

00:22:21.240 --> 00:22:24.210 to a massive amount of deviation

00:22:24.210 --> 00:22:27.050 from the uniform distribution in those P-values.

00:22:27.050 --> 00:22:30.520 So in other words, negative control,

00:22:30.520 --> 00:22:33.300 gene-enhancer pairs are looking incredibly

00:22:33.300 --> 00:22:35.563 significant according to this analysis.

00:22:37.110 --> 00:22:40.430 So in this particular analysis

00:22:40.430 --> 00:22:42.356 they essentially found

00:22:42.356 --> 00:22:46.750 the same thing here it's portrayed as a Manhattan plot

00:22:46.750 --> 00:22:49.570 but you see a lot

00:22:49.570 --> 00:22:52.720 of things reaching significance when right only

00:22:52.720 --> 00:22:57.720 the circle points are those that essentially were replicated

00:22:58.090 --> 00:23:02.050 in a bulk RNA sequencing experiment.

00:23:02.050 --> 00:23:07.050 And then this one finally looks like they perturbed

00:23:09.615 --> 00:23:13.150 lots of different enhancers and essentially looked

00:23:13.150 --> 00:23:15.540 at the effect on this one particular gene.

00:23:15.540 --> 00:23:19.360 And essentially what they found is that essentially all

00:23:19.360 --> 00:23:21.840 of the enhancers that they tested appeared to

00:23:21.840 --> 00:23:24.530 actually be per, like, have an effect on the expression

00:23:24.530 --> 00:23:28.612 of this gene, when in fact this is biologically imposible.

00:23:28.612 --> 00:23:31.750 So this is clearly an issue.

00:23:31.750 --> 00:23:36.110 Now, these original papers clearly knew

00:23:36.110 --> 00:23:39.060 that there was an issue, and so for each of the papers

00:23:39.060 --> 00:23:40.260 they kind of have a little bit

00:23:40.260 --> 00:23:44.640 of an ad hoc fix in order to basically correct their P-value

00:23:44.640 --> 00:23:48.223 of distributions, so that they look a little bit more,

00:23:50.090 --> 00:23:52.100 closer to being calibrated.

00:23:52.100 --> 00:23:54.900 And so I'm, I think for the sake of time

00:23:54.900 --> 00:23:57.880 I'm probably not going to get into exactly how

00:23:57.880 --> 00:24:01.740 they propose to fix their P-value distributions.

00:24:01.740 --> 00:24:05.640 What I will say is that we looked in detail

00:24:05.640 --> 00:24:08.110 especially at the strategy that they use here

00:24:08.110 --> 00:24:09.960 and to a lesser extent at the strategies.

00:24:09.960 --> 00:24:10.910 Well, actually I think here

00:24:10.910 --> 00:24:13.720 they basically said just not to apply their method

00:24:13.720 --> 00:24:18.000 to data where there's too high, essentially

00:24:18.000 --> 00:24:20.650 to where they're too many perturbations per cell.

00:24:20.650 --> 00:24:23.650 So in this case, they just said, don't apply this method.

00:24:23.650 --> 00:24:25.920 We looked into the kinds of fixes that they proposed

00:24:25.920 --> 00:24:28.050 in these two papers, and they essentially

00:24:28.050 --> 00:24:31.040 they don't quite work in the way that you would expect.

00:24:31.040 --> 00:24:33.360 And so what we thought is that,

00:24:33.360 --> 00:24:36.810 what we'd like to do is kind of look a little deeper

00:24:36.810 --> 00:24:38.850 into this problem and try to ask ourselves

00:24:38.850 --> 00:24:40.900 why are we seeing all of these issues?

00:24:40.900 --> 00:24:43.740 Why do people keep running into these miscalibration issues

00:24:43.740 --> 00:24:48.740 and let's try to basically address those underlying issues.

00:24:50.210 --> 00:24:53.020 So we thought about it a little bit

00:24:53.020 --> 00:24:55.490 and we thought about challenges

00:24:55.490 --> 00:24:58.620 for both parametric and non-parametric methods.

00:24:58.620 --> 00:25:00.940 So for parametric methods

00:25:00.940 --> 00:25:04.850 this actually shouldn't really come as a surprise probably

00:25:04.850 --> 00:25:07.850 to most people here, gene expression is known to

00:25:07.850 --> 00:25:10.770 be pretty hard to model in single cells.

00:25:10.770 --> 00:25:14.660 So of course we have these essentially highly discreet

00:25:15.740 --> 00:25:19.670 lots of zeros counts that are over dispersed

00:25:19.670 --> 00:25:23.360 perhaps more importantly, given how sparse the data are.

00:25:23.360 --> 00:25:25.580 It's actually pretty hard to get a good estimate

00:25:25.580 --> 00:25:27.570 of that dispersion parameter.

00:25:27.570 --> 00:25:29.680 And so there's currently no standard way

00:25:29.680 --> 00:25:32.020 of estimating that dispersion parameter

00:25:32.020 --> 00:25:35.360 and basically every paper, comes up

00:25:35.360 --> 00:25:37.413 with their own way of doing this.

00:25:39.950 --> 00:25:41.080 They're even just debates

00:25:41.080 --> 00:25:43.750 about what parametric models are appropriate for these data,

00:25:43.750 --> 00:25:46.040 should they be zero inflated,

00:25:46.040 --> 00:25:50.150 should they not be, and some genes have even been observed

00:25:50.150 --> 00:25:52.240 to have bi-modal expression patterns.

00:25:52.240 --> 00:25:54.840 So essentially all of these things are telling us

00:25:54.840 --> 00:25:56.550 that it's kind of hard to shoe horn

00:25:56.550 --> 00:25:58.280 single cell gene expression,

00:25:58.280 --> 00:26:01.260 into a nice, neat parametric model.

00:26:01.260 --> 00:26:04.440 So obviously if you have missed specification of your model

00:26:04.440 --> 00:26:06.530 such as a bad estimate for a dispersion perimeter

00:26:06.530 --> 00:26:09.261 that very well could cause miscalibration

00:26:09.261 --> 00:26:10.523 of the kind that we saw.

00:26:13.050 --> 00:26:16.310 So next we can think about non-parametric methods.

00:26:16.310 --> 00:26:19.340 So maybe, obviously if these data

00:26:19.340 --> 00:26:20.780 are hard to model parametrically

00:26:20.780 --> 00:26:23.803 maybe the non-parametric methods are going to save us.

00:26:24.730 --> 00:26:26.780 But the observation that we made that I think is

00:26:26.780 --> 00:26:29.080 quite important is that these technical factors

00:26:29.080 --> 00:26:31.960 that I mentioned before, like sequencing depth,

00:26:31.960 --> 00:26:34.700 they impact not only the expressions of genes

00:26:34.700 --> 00:26:38.370 but also the detection of these CRISPR guider in is.

00:26:38.370 --> 00:26:41.020 So I might have led you to believe

00:26:41.020 --> 00:26:43.280 in one of my early slides that we can basically

00:26:43.280 --> 00:26:45.680 perfectly measure which cell contains

00:26:45.680 --> 00:26:50.033 which CRISPR perturbations, but this is actually not true.

00:26:50.970 --> 00:26:53.720 So single cell RNA sequencing

00:26:53.720 --> 00:26:58.590 it's essentially just like this kind of a sampling process.

00:26:58.590 --> 00:27:02.657 And so the more reads you sample from a cell

00:27:02.657 --> 00:27:05.450 the more likely you are to detect a guide RNAs.

00:27:05.450 --> 00:27:09.650 And so we just essentially looked at, for example,

00:27:09.650 --> 00:27:13.380 this is for one of the datasets and we just made

00:27:13.380 --> 00:27:16.860 a scatterplot of the total number of guide RNAs detected

00:27:16.860 --> 00:27:19.860 per cell versus the total number of UMI.

00:27:19.860 --> 00:27:21.430 So this is the sequencing depth

00:27:21.430 --> 00:27:23.960 and we found this extremely clear

00:27:23.960 --> 00:27:25.450 I guess I'm not showing you the P-value

00:27:25.450 --> 00:27:28.510 but this P-value was like absurdly significant

00:27:28.510 --> 00:27:30.730 to just basically confirm that

00:27:30.730 --> 00:27:33.260 if you have more sequencing depth in a cell,

00:27:33.260 --> 00:27:36.003 you're going to find more guide our news in that cell.

00:27:36.870 --> 00:27:39.540 And so the issue with this is

00:27:39.540 --> 00:27:43.200 that we basically have a confounding problem on our hands.

00:27:43.200 --> 00:27:47.520 So think about this graphical model that's illustrating

00:27:47.520 --> 00:27:48.570 what's going on

00:27:48.570 --> 00:27:51.550 in a single cell CRISPR screen experiment in

00:27:51.550 --> 00:27:55.960 this gray box is kind of the underlying biological reality.

00:27:55.960 --> 00:27:58.408 Let's say we have this presence of this guide RNA

00:27:58.408 --> 00:28:01.901 and the expression of this gene and the guide RNA is

00:28:01.901 --> 00:28:05.470 or the, yeah, I guess the, the CRISPR knockdown

00:28:05.470 --> 00:28:08.190 of the enhancer is either affecting gene expression

00:28:08.190 --> 00:28:12.670 or it is not, but we read it out.

00:28:12.670 --> 00:28:17.610 Some essentially imprecise the measurement

00:28:17.610 --> 00:28:18.840 of the guide RNA presence.

00:28:18.840 --> 00:28:20.210 We also read out

00:28:20.210 --> 00:28:23.100 and imprecise measurement of the gene expression.

00:28:23.100 --> 00:28:27.190 And what's most important is that the technical factors such

00:28:27.190 --> 00:28:29.630 as sequencing depth, they're actually impacting both

00:28:29.630 --> 00:28:32.690 of these measurements, they're coming from the same cell.

00:28:32.690 --> 00:28:36.980 And so even if there is no association between the guide RNA

00:28:36.980 --> 00:28:41.610 and the gene, if you just basically naively look

00:28:41.610 --> 00:28:44.650 at the association between the measured guide RNA presence

00:28:44.650 --> 00:28:46.140 and the measured gene expression

00:28:46.140 --> 00:28:50.210 you're going to find some association.

00:28:50.210 --> 00:28:52.640 And so this is clearly an issue.

00:28:52.640 --> 00:28:55.060 And so essentially in order to correct

00:28:55.060 --> 00:28:56.900 for this confounding effect, it's very important

00:28:56.900 --> 00:29:01.380 to test instead of just testing independence between

00:29:01.380 --> 00:29:04.360 the perturbation and the expression.

00:29:04.360 --> 00:29:06.660 We want to test conditional independence, where

00:29:06.660 --> 00:29:10.050 we're conditioning on all of these technical factors.

00:29:10.050 --> 00:29:13.670 And so this shows you why non-parametric methods tend to

00:29:13.670 --> 00:29:18.240 suffer is because when you do things like permute your data

00:29:18.240 --> 00:29:21.110 or rank your data, there's this underlying assumption

00:29:21.110 --> 00:29:23.640 that all of the cells are exchangeable and you're

00:29:23.640 --> 00:29:26.750 using that exchange ability to build your inference on.

00:29:26.750 --> 00:29:30.010 And so when you do those tests, they're implicitly

00:29:30.010 --> 00:29:34.300 actually testing just the direct independence

00:29:34.300 --> 00:29:35.990 the unconditional independence.

00:29:35.990 --> 00:29:39.420 And so this sort of inflation we saw

00:29:39.420 --> 00:29:42.770 in the non-parametric methods be explained by this

00:29:42.770 --> 00:29:43.903 Source of confounding.

00:29:46.660 --> 00:29:49.700 So that's actually it for that part of my talk

00:29:50.720 --> 00:29:52.700 any questions about the existing methods

00:29:52.700 --> 00:29:54.040 and the analysis challenges

00:29:54.040 --> 00:29:57.480 and why there's a need to think about new methodology

00:29:57.480 --> 00:29:59.883 for this for this problem.

00:30:05.583 --> 00:30:07.333 Okay, I will move on.

00:30:09.510 --> 00:30:12.920 So this is the part of the talk where I'm going to

00:30:12.920 --> 00:30:17.173 propose a new analysis method for this kind of data.

00:30:19.120 --> 00:30:22.120 And so the key kind of idea we're gonna use is

00:30:22.120 --> 00:30:27.087 conditional resampling, which is proposed by not us.

00:30:29.820 --> 00:30:34.400 So the idea of the conditional randomization test

00:30:34.400 --> 00:30:37.070 well, it's actually, depending on how you look at it

00:30:37.070 --> 00:30:39.500 it's quite an old idea and it has some connections to

00:30:39.500 --> 00:30:44.500 causal inference, but it was proposed also incandescent all.

00:30:45.380 --> 00:30:48.320 And essentially the setup is that you want to

00:30:48.320 --> 00:30:51.750 test conditional independence and you're under

00:30:51.750 --> 00:30:54.580 the assumption that you have a decent estimate

00:30:54.580 --> 00:30:57.090 of the distribution of X given Z.

00:30:57.090 --> 00:30:59.750 So remember X is the perturbation.

00:30:59.750 --> 00:31:01.790 Y is the expression and Z are the,

00:31:01.790 --> 00:31:03.720 essentially the confounders.

00:31:03.720 --> 00:31:06.920 So one way of thinking about it from a causal inference

00:31:06.920 --> 00:31:11.920 standpoint is let's say we know the propensity score,

00:31:11.920 --> 00:31:14.530 can we test whether there's a causal relationship

00:31:14.530 --> 00:31:19.530 between X and Y sort of controlling for these Confounders?

00:31:19.750 --> 00:31:24.180 So the idea of the conditional randomization test

00:31:24.180 --> 00:31:26.000 is the following.

00:31:26.000 --> 00:31:30.980 First, you take any test statistic T of your data,

00:31:30.980 --> 00:31:34.250 and in order to calibrate this test statistic

00:31:34.250 --> 00:31:38.860 under the null hypothesis, instead of doing a permutation

00:31:38.860 --> 00:31:40.850 we're gonna do a slightly more sophisticated

00:31:40.850 --> 00:31:45.090 resampling operation, where we're going to go through,

00:31:45.090 --> 00:31:50.090 and for every cell, we are going to resample whether

00:31:50.480 --> 00:31:54.840 or not it received the given perturbation, but conditionally

00:31:54.840 --> 00:31:58.920 on the specific technical factors that were in that cell.

00:31:58.920 --> 00:32:02.650 And here we're using crucially the information that we have

00:32:02.650 --> 00:32:06.510 a handle on what this sort of propensity score is.

00:32:06.510 --> 00:32:10.030 And then we're just going to recompute the test

00:32:10.030 --> 00:32:13.720 the same test statistic on the resample data.

00:32:13.720 --> 00:32:16.570 And then we're just gonna define the a P-value

00:32:16.570 --> 00:32:20.310 in the usual way for a resampling based procedure.

00:32:20.310 --> 00:32:22.810 So one way of thinking about it is

00:32:22.810 --> 00:32:26.650 that it's kind of like a permutation test, but it's one

00:32:26.650 --> 00:32:30.725 in which the reassignments of the guide RNAs

00:32:30.725 --> 00:32:35.725 to the cells is one that respects

00:32:36.778 --> 00:32:40.460 the confounding that there is

00:32:40.460 --> 00:32:43.433 in the data instead of treating all the cells exchangeable.

00:32:44.570 --> 00:32:49.570 So this is great because the CRT adjust

00:32:51.010 --> 00:32:55.100 for confounders basically by construction and importantly

00:32:55.100 --> 00:32:58.640 it avoids assumptions on the gene expression distribution.

00:32:58.640 --> 00:33:01.420 And in fact, provably, the P-value you get

00:33:01.420 --> 00:33:06.033 out of the CRT is valid, even if essentially,

00:33:07.290 --> 00:33:10.833 even if the test statistic T is, anything you want.

00:33:12.865 --> 00:33:15.017 So in the sense that kind of addresses

00:33:15.017 --> 00:33:19.350 the confounding issues, like basically the Achilles heel

00:33:19.350 --> 00:33:22.770 of the non-parametric methods, but avoiding assumptions

00:33:22.770 --> 00:33:24.530 on the gene expression distribution

00:33:24.530 --> 00:33:27.200 as sort of was the pitfall of the parametric methods.

00:33:27.200 --> 00:33:30.430 And it kind of seems to be doing something

00:33:30.430 --> 00:33:32.970 that's avoiding both of those issues.

00:33:32.970 --> 00:33:36.840 Now, of course, there's a, trade-off in the

00:33:36.840 --> 00:33:40.230 CRT does require you to have some estimate

00:33:40.230 --> 00:33:42.130 of this propensity score.

00:33:42.130 --> 00:33:47.130 So, and then secondly, the CRT is computationally expensive

00:33:47.520 --> 00:33:50.870 if you consider, or if you compare it to like

00:33:50.870 --> 00:33:53.150 just like a parametric regression here

00:33:53.150 --> 00:33:54.650 we're doing a parametric regression

00:33:54.650 --> 00:33:57.240 but we're doing it lots of times.

00:33:57.240 --> 00:33:59.763 And so how do we get around some of these issues?

00:34:01.360 --> 00:34:04.600 So, and in particular, how do we actually go

00:34:04.600 --> 00:34:08.450 about applying this idea to single cell CRISPR screens?

00:34:08.450 --> 00:34:12.640 And so, firstly, do we understand this distribution

00:34:12.640 --> 00:34:17.070 of the probability of observing a guide or in a

00:34:17.070 --> 00:34:19.520 given a set of technical factors?

00:34:19.520 --> 00:34:24.520 So what we're going to do in this particular method,

00:34:25.070 --> 00:34:26.830 well, first we're gonna observe that it's

00:34:26.830 --> 00:34:30.630 this is kind of a simpler phenomenon than gene expression

00:34:30.630 --> 00:34:33.030 like guide our nays are not really, like subject

00:34:33.030 --> 00:34:36.580 to all of the complicated regulatory patterns of genes.

00:34:36.580 --> 00:34:39.730 And secondly, kind of under the hood,

00:34:39.730 --> 00:34:44.730 the actual assortments of guide our nays

00:34:44.869 --> 00:34:48.420 to cells is, you know, like fairly well modeled.

00:34:48.420 --> 00:34:51.940 It's just basically like in that sense

00:34:51.940 --> 00:34:53.300 the cells are pretty exchangeable.

00:34:53.300 --> 00:34:54.720 What's not exchangeable it just basically

00:34:54.720 --> 00:34:56.480 this measurement process.

00:34:56.480 --> 00:34:58.910 So this is just kind of a simpler object

00:34:58.910 --> 00:35:02.770 in the specific case of single cell CRISPR screens.

00:35:02.770 --> 00:35:03.940 So we can try to bring

00:35:03.940 --> 00:35:07.850 to bear various knowledge to try to get a good sense

00:35:07.850 --> 00:35:09.650 of this in this case,

00:35:09.650 --> 00:35:12.260 we're just gonna sort of do the easiest thing possible

00:35:12.260 --> 00:35:15.293 and we're gonna fit it using an logistic regression.

00:35:16.550 --> 00:35:18.980 The second thing we're going to do is think

00:35:18.980 --> 00:35:21.140 about what test statistic to use.

00:35:21.140 --> 00:35:26.140 So I had the separate paper about essentially the power

00:35:27.370 --> 00:35:28.840 of the conditioner randomization tests.

00:35:28.840 --> 00:35:33.000 And what we found is that the closer the test statistic is

00:35:33.000 --> 00:35:37.610 to the true conditional distribution of Y given X, Z

00:35:37.610 --> 00:35:39.890 I guess I should say the true likelihood,

00:35:39.890 --> 00:35:41.150 the better the power will be.

00:35:41.150 --> 00:35:44.620 And so in that sense, what we wanna do is we

00:35:44.620 --> 00:35:49.350 wanna leverage existing models that people have used such

00:35:49.350 --> 00:35:51.490 as negative binomial regression.

00:35:51.490 --> 00:35:54.340 It's not going to matter whether the model is true or not

00:35:54.340 --> 00:35:58.840 for the sake of type one error control, but we hope

00:35:58.840 --> 00:36:01.140 that we can do a better job in terms of power

00:36:02.470 --> 00:36:05.993 by trying to get a good model for this.

00:36:07.350 --> 00:36:10.090 And finally, how do we mitigate the computational cost?

00:36:10.090 --> 00:36:12.230 And so we had a few ideas for this as well.

00:36:12.230 --> 00:36:15.140 So one of them is called the distilled CRT.

00:36:15.140 --> 00:36:18.710 And so I'll if time permits, which might or might not

00:36:18.710 --> 00:36:20.470 I'll give you a few more details

00:36:20.470 --> 00:36:24.280 about how you can use this to have a much faster

00:36:25.260 --> 00:36:28.370 for every resample to be quick.

00:36:28.370 --> 00:36:31.720 And then we're also going to use this hack, essentially

00:36:31.720 --> 00:36:35.510 that what we found is that the resampling distribution

00:36:35.510 --> 00:36:40.220 it actually kind of looks pretty reasonable.

00:36:40.220 --> 00:36:42.840 It kind of looks like a normal, but it's sort

00:36:42.840 --> 00:36:45.500 of how some extra skew and maybe some extra heavy tails.

00:36:45.500 --> 00:36:47.560 And so what we're gonna do is we're going to

00:36:47.560 --> 00:36:51.720 fit a skew T distribution to the essentially

00:36:51.720 --> 00:36:54.630 the empirical distribution of the resample test statistics.

00:36:54.630 --> 00:36:57.610 And in that way, we can get more accurate P-values

00:36:57.610 --> 00:37:00.520 without doing as many recent samples.

00:37:00.520 --> 00:37:02.610 And so putting together all of these pieces

00:37:02.610 --> 00:37:05.580 we get this method, which we call Sceptre

00:37:05.580 --> 00:37:08.400 or single cell perturbation screen analysis

00:37:08.400 --> 00:37:10.103 via conditional resampling.

00:37:11.010 --> 00:37:12.920 And so essentially what we do is what I said

00:37:12.920 --> 00:37:14.650 on the previous slide.

00:37:14.650 --> 00:37:18.840 We first use a logistic regression to fit a probability

00:37:18.840 --> 00:37:21.983 for every cell that we would find a perturbation there.

00:37:22.970 --> 00:37:25.500 And then we're gonna use these perturbation probabilities

00:37:25.500 --> 00:37:28.003 and resample this particular column.

00:37:28.960 --> 00:37:32.470 And so we now we have a whole bunch of resample datasets.

00:37:32.470 --> 00:37:34.860 Now we're going to use a negative binomial regression

00:37:34.860 --> 00:37:37.750 or more precisely a distilled negative binomial regression

00:37:37.750 --> 00:37:41.570 for speed, to get the test statistic

00:37:41.570 --> 00:37:42.810 for both the original data.

00:37:42.810 --> 00:37:45.593 And for all of these re resample datasets.

00:37:46.500 --> 00:37:47.710 Then we're gonna put together all

00:37:47.710 --> 00:37:51.070 of these recycled test statistics into this gray histogram.

00:37:51.070 --> 00:37:54.110 And again, we're gonna fit this magenta curve

00:37:54.110 --> 00:37:56.630 which is the skew T distribution

00:37:56.630 --> 00:37:59.020 which seems to fit pretty well in most cases.

00:37:59.020 --> 00:38:01.540 And then we're gonna compare the original test statistic

00:38:01.540 --> 00:38:06.540 against this skew T distribution and get a P-value that way.

00:38:06.760 --> 00:38:09.890 And so this is represented by the shaded region here.

00:38:09.890 --> 00:38:14.060 And I think what's noteworthy is to compare this fitted

00:38:14.060 --> 00:38:15.370 and all No distribution

00:38:15.370 --> 00:38:19.070 to this standard normal No distribution.

00:38:19.070 --> 00:38:20.240 I guess I should have said here

00:38:20.240 --> 00:38:24.020 that the actual test statistics are a Z values extracted

00:38:24.020 --> 00:38:25.810 from the negative binomial regression.

00:38:25.810 --> 00:38:29.740 So if your model were true, the Z values

00:38:29.740 --> 00:38:33.510 under the No would follow a standard normal distribution.

00:38:33.510 --> 00:38:36.550 And so what we find is that when we resample we

00:38:36.550 --> 00:38:39.700 get something that's not the standard normal distribution.

00:38:39.700 --> 00:38:42.270 And so in the sense you can view it as,

00:38:42.270 --> 00:38:47.270 a sort of measure of the departure sort of from,

00:38:48.250 --> 00:38:50.930 or sort of the lack of model fit that went

00:38:50.930 --> 00:38:52.963 into this negative binomial regression.

00:38:54.480 --> 00:38:56.640 So another way of putting this is that

00:38:56.640 --> 00:38:58.960 you can imagine that if you did happen

00:38:58.960 --> 00:39:02.690 to correctly specify your negative binomial regression model

00:39:02.690 --> 00:39:05.950 then you would sort of be getting back the same P-value

00:39:05.950 --> 00:39:07.870 that you would have gotten otherwise.

00:39:07.870 --> 00:39:08.703 So in that sense

00:39:08.703 --> 00:39:10.610 we're not really reinventing the wheel here

00:39:10.610 --> 00:39:13.490 if you do have a good parametric model, but if you don't

00:39:13.490 --> 00:39:16.390 then we can correct for it using this resampling strategy.

00:39:17.890 --> 00:39:19.240 So I guess this is an important slide

00:39:19.240 --> 00:39:23.140 so maybe I will stay here for a little bit and ask

00:39:23.140 --> 00:39:27.963 if anyone has questions about how our methodology works.

00:39:31.290 --> 00:39:33.200 - Hi, I have a bunker question.

00:39:33.200 --> 00:39:35.870 So have you tried to hurdle model to deal

00:39:35.870 --> 00:39:39.610 with this kind of full data is the cause

00:39:39.610 --> 00:39:41.903 of the weird distribution of the data?

00:39:44.650 --> 00:39:47.900 - Oh, so let's see.

00:39:47.900 --> 00:39:52.050 You mean to model the, essentially to model the gene

00:39:52.050 --> 00:39:56.900 expressions or do you mean to model the CRISPR perturbations

00:39:58.355 --> 00:40:02.910 - From this page,

00:40:02.910 --> 00:40:04.830 so first step you use a logistic regression

00:40:04.830 --> 00:40:08.220 and then you use a nickname by knowing that binomial.

00:40:08.220 --> 00:40:12.940 So it's like a two step models, but to hurdle model

00:40:12.940 --> 00:40:16.413 they combine them together to deal with the overall dataset.

00:40:17.740 --> 00:40:21.010 - I see, I will admit that I'm not familiar with those

00:40:21.010 --> 00:40:24.100 models but I will definitely take a look

00:40:24.100 --> 00:40:26.350 at those and see if they might be applicable.

00:40:27.540 --> 00:40:32.540 Yeah, I guess like in this sense

00:40:33.220 --> 00:40:36.620 the approach that I've proposed here is pretty flexible.

00:40:36.620 --> 00:40:37.453 I mean, really

00:40:37.453 --> 00:40:41.900 what makes this approach work well is as long

00:40:41.900 --> 00:40:43.980 as you have a decent approximation

00:40:43.980 --> 00:40:45.950 to these probation probabilities

00:40:45.950 --> 00:40:48.110 we're thinking about them as propensity scores.

00:40:48.110 --> 00:40:52.480 So aside from that but

00:40:52.480 --> 00:40:54.730 because really what's standing behind this as the generality

00:40:54.730 --> 00:40:56.270 of the conditional randomization test where

00:40:56.270 --> 00:40:58.200 you can basically use any test statistic you want.

00:40:58.200 --> 00:41:02.300 And so, definitely the method is flexible

00:41:02.300 --> 00:41:05.080 and can incorporate different choices,

00:41:05.080 --> 00:41:06.830 like the one that you've mentioned,

00:41:07.867 --> 00:41:09.710 But we haven't tried it we haven't, we haven't tried it.

00:41:09.710 --> 00:41:11.360 I'm not familiar with this model.

00:41:12.200 --> 00:41:13.050 Thank you though.

00:41:15.040 --> 00:41:19.683 Anyone else have any questions about the methodology?

00:41:24.090 --> 00:41:27.340 Okay, perhaps I'll okay.

00:41:27.340 --> 00:41:31.250 So yes, so this is kind of like a separate thing

00:41:31.250 --> 00:41:32.830 which I will not get

00:41:32.830 --> 00:41:35.590 into details of for the sake of time, but we had

00:41:35.590 --> 00:41:39.060 the separate paper whose focus was just basically,

00:41:39.060 --> 00:41:41.810 the conditional randomization test is a cool test

00:41:41.810 --> 00:41:43.290 but everyone knows it's slow.

00:41:43.290 --> 00:41:46.160 So how can we essentially accelerate it

00:41:46.160 --> 00:41:49.030 while retaining a lot of its power advantages?

00:41:49.030 --> 00:41:51.410 And so what we found is that

00:41:51.410 --> 00:41:54.260 if you just ever so slightly modified the test statistic

00:41:54.260 --> 00:41:58.716 by sort of regressing Y first on the confounders,

00:41:58.716 --> 00:42:01.780 and then on X, instead

00:42:01.780 --> 00:42:04.030 of regressing it on both at the same time

00:42:04.030 --> 00:42:07.690 what we found is that this ends up being much, much faster

00:42:07.690 --> 00:42:10.050 because only the second step needs to be repeated

00:42:10.050 --> 00:42:13.950 upon resampling, and the second step is much cheaper.

00:42:13.950 --> 00:42:18.770 So what we did is that we, in the context of sector

00:42:18.770 --> 00:42:19.690 we built on this

00:42:19.690 --> 00:42:23.430 by accelerating the resampling steps even further

00:42:23.430 --> 00:42:25.110 by leveraging the sparsity

00:42:25.110 --> 00:42:28.140 of the CRISPR perturbation vector X.

00:42:28.140 --> 00:42:31.360 And so perhaps the most important part is that the cost

00:42:31.360 --> 00:42:34.400 of the CRT for one gene-enhancer pair went

00:42:34.400 --> 00:42:37.850 down from 25 minutes down to 20 seconds

00:42:37.850 --> 00:42:40.450 as a result of these computational accelerations.

00:42:40.450 --> 00:42:42.750 And so for reference a single negative binomial

00:42:42.750 --> 00:42:45.040 regression took three seconds.

00:42:45.040 --> 00:42:46.130 So it's still,

00:42:46.130 --> 00:42:50.200 we're a factor of six or seven, more expensive than the

00:42:50.200 --> 00:42:52.510 just the sort of vanilla single regression

00:42:52.510 --> 00:42:56.380 but it's definitely, I think sort of within,

00:42:56.380 --> 00:42:58.070 definitely within an order of magnitude

00:42:58.070 --> 00:43:01.983 and hopefully as you can tell a much better statistically.

00:43:03.160 --> 00:43:08.160 So I will show you a few, so this is a simulation.

00:43:11.970 --> 00:43:14.320 I'm not gonna go through it in detail, but the idea is

00:43:14.320 --> 00:43:18.610 that what we're demonstrating here is that you can give

00:43:18.610 --> 00:43:22.340 Sceptre essentially negative binomial models

00:43:22.340 --> 00:43:25.310 that are miss specified in different ways.

00:43:25.310 --> 00:43:27.360 You can, give it a dispersion

00:43:27.360 --> 00:43:29.710 that's too large, a dispersion that's too small

00:43:29.710 --> 00:43:32.210 or maybe the true model does have zero inflation

00:43:32.210 --> 00:43:33.660 but we're not accounting for it.

00:43:33.660 --> 00:43:35.930 And what we find is that Sceptre essentially

00:43:35.930 --> 00:43:38.883 is well calibrated, regardless,

00:43:40.080 --> 00:43:42.780 whereas if you just essentially took

00:43:42.780 --> 00:43:46.920 the like the wrong dispersion estimates at face value

00:43:46.920 --> 00:43:48.870 you would encounter problems.

00:43:48.870 --> 00:43:50.860 And this SE magic approach

00:43:50.860 --> 00:43:53.820 which basically is a permutation approach.

00:43:53.820 --> 00:43:56.240 It's just sort of not doing a great job accounting

00:43:56.240 --> 00:43:58.690 for the confounding it, so we see this inflation.

00:44:00.990 --> 00:44:03.020 So perhaps more excitingly

00:44:03.020 --> 00:44:07.640 I'd like to show you an application to real data.

00:44:07.640 --> 00:44:10.900 So I guess this is the, so firstly

00:44:10.900 --> 00:44:13.050 we wanna make sure method is actually calibrated.

00:44:13.050 --> 00:44:15.610 So if you remember the initial observation was

00:44:15.610 --> 00:44:18.000 in a lot of these methods, aren't calibrated.

00:44:18.000 --> 00:44:20.240 So because I'm running a little short on time

00:44:20.240 --> 00:44:22.380 let's kind of maybe ignore this panel here

00:44:22.380 --> 00:44:24.280 and focus our attention here.

00:44:24.280 --> 00:44:28.240 So this is the Gasperini data that I introduced before.

00:44:28.240 --> 00:44:33.240 And so this red line here is actually the QQ plot you saw

00:44:33.750 --> 00:44:35.440 on one of my first slides

00:44:35.440 --> 00:44:39.290 of all of those negative control gene-enhancer pairs.

00:44:39.290 --> 00:44:40.500 It looks different here because

00:44:40.500 --> 00:44:42.580 the scale is I've sort of cut off the scale

00:44:42.580 --> 00:44:44.730 so we can actually visualize it.

00:44:44.730 --> 00:44:47.423 So we see a quite significant departure.

00:44:48.300 --> 00:44:51.032 What we actually did is we thought, okay, maybe

00:44:51.032 --> 00:44:53.660 they have a bad estimate of the dispersion

00:44:53.660 --> 00:44:55.210 but maybe we can use some more

00:44:55.210 --> 00:45:00.210 like state-of-the-art single cell sort of methods

00:45:00.480 --> 00:45:02.800 to improve our estimate of the dispersion.

00:45:02.800 --> 00:45:04.260 And so maybe we don't need to go

00:45:04.260 --> 00:45:06.150 to all the effort of doing the resampling.

00:45:06.150 --> 00:45:07.460 And so what we found is that

00:45:07.460 --> 00:45:10.120 when we use a state-of-the-art dispersion estimate

00:45:10.120 --> 00:45:12.643 we still have very substantial miscalibration.

00:45:14.060 --> 00:45:15.260 This is, I think, just a Testament

00:45:15.260 --> 00:45:17.540 to the fact that it's just hard to estimate that perimeter

00:45:17.540 --> 00:45:20.640 because there's not all that much data to estimate it.

00:45:20.640 --> 00:45:23.810 And then by comparison, we built Sceptre

00:45:23.810 --> 00:45:27.030 from the same exact negative binomial model

00:45:27.030 --> 00:45:28.890 which is this improved one,

00:45:28.890 --> 00:45:32.370 and we found that the negative control P-values

00:45:32.370 --> 00:45:35.750 are I think, excellently calibrated.

00:45:35.750 --> 00:45:39.700 So this shows you, again, the benefit

00:45:39.700 --> 00:45:42.350 of this different way of calibrating your test statistic

00:45:42.350 --> 00:45:45.583 and not relying on the parametric model for gene expression.

00:45:47.250 --> 00:45:50.040 So this figure just shows a few of the other methods

00:45:50.040 --> 00:45:52.973 but for the sake of time, I'm going to move on.

00:45:54.960 --> 00:45:57.860 This is looking at positive control data.

00:45:57.860 --> 00:46:01.420 So this basically is like, trying to get a sense of power.

00:46:01.420 --> 00:46:03.600 And so, again, maybe if we restrict our attention

00:46:03.600 --> 00:46:06.320 to this left panel here, what we found is that

00:46:06.320 --> 00:46:09.660 if we just plot the, our P-values

00:46:09.660 --> 00:46:12.180 versus the P-values, by the way, maybe I should say

00:46:12.180 --> 00:46:13.640 what is a positive control.

00:46:13.640 --> 00:46:14.680 A positive control

00:46:14.680 --> 00:46:18.200 in this case is a CRISPR perturbation that instead

00:46:18.200 --> 00:46:21.770 of targeting and enhancer is targeting the transcription

00:46:21.770 --> 00:46:24.740 start sites of a gene.

00:46:24.740 --> 00:46:29.250 And so essentially, like we don't need any extra biology

00:46:29.250 --> 00:46:32.120 to know that, if you target a transcription start site

00:46:32.120 --> 00:46:34.310 that's really going to knock out the gene.

00:46:34.310 --> 00:46:36.960 And so you can still try to do your association test and see

00:46:36.960 --> 00:46:40.140 if you've picked up those positive control associations.

00:46:40.140 --> 00:46:41.280 And so what we find is that

00:46:41.280 --> 00:46:44.270 actually Sceptre not only is better calibrated

00:46:44.270 --> 00:46:47.730 but it also tends to have more significant P-values

00:46:47.730 --> 00:46:49.010 on those positive controls.

00:46:49.010 --> 00:46:52.690 So it apparently is boosting both the sensitivity

00:46:52.690 --> 00:46:55.753 and the specificity of this association tests.

00:46:56.623 --> 00:46:59.811 - Eugene here are the original empirical P-value is this

00:46:59.811 --> 00:47:03.004 from the negative binomial test.

00:47:03.004 --> 00:47:06.837 So after we did the conditional recommendation

00:47:08.410 --> 00:47:11.390 if you actually have better P-values

00:47:11.390 --> 00:47:13.180 for the positive control pairs.

00:47:13.180 --> 00:47:18.180 - Yes, so you would expect, you would expect it's like

00:47:21.590 --> 00:47:23.490 aren't we just making the P-value is just,

00:47:23.490 --> 00:47:24.610 like less significant

00:47:24.610 --> 00:47:26.590 in a way to just help with the calibration.

00:47:26.590 --> 00:47:28.730 So how can it be boosting power?

00:47:28.730 --> 00:47:33.730 But I like the degree of inflation sort of varies

00:47:34.440 --> 00:47:37.290 like essentially it's not like, and what we'll see this

00:47:37.290 --> 00:47:40.170 I think on the next slide as well, essentially

00:47:40.170 --> 00:47:43.380 we're not like, sort of what sector is doing is not

00:47:43.380 --> 00:47:45.710 like a monotone transformation of things.

00:47:45.710 --> 00:47:50.183 It kind of there's not actually just maybe to illustrate it.

00:47:50.183 --> 00:47:55.183 I think, this is just an example where essentially what

00:47:59.270 --> 00:48:01.010 we would have gotten from the sort

00:48:01.010 --> 00:48:03.790 of the vanilla negative binomial analysis is the area

00:48:03.790 --> 00:48:07.480 under this dotted or dashed curve here.

00:48:07.480 --> 00:48:11.470 And so Sceptre could, well, basically whoops sorry,

00:48:11.470 --> 00:48:14.580 it could have a, like a lighter tail as it has in this case.

00:48:14.580 --> 00:48:19.580 And so it could sort of either make the P-values

00:48:19.890 --> 00:48:21.900 on the more significant or less significant.

00:48:21.900 --> 00:48:23.750 It's correcting the miscalibration

00:48:23.750 --> 00:48:26.410 but not necessarily in a way that's like conservative.

00:48:26.410 --> 00:48:30.253 And so this is encouraging.

00:48:31.620 --> 00:48:34.540 Yeah, that's a good question though.

00:48:34.540 --> 00:48:36.700 - I guess that depends on

00:48:36.700 --> 00:48:39.713 the confounding you included in the model.

00:48:40.750 --> 00:48:45.750 So then I would expect it well, re reduce the significance

00:48:47.470 --> 00:48:49.720 but if you include other co-founding

00:48:49.720 --> 00:48:54.720 that's mostly contributing to the noise level probably.

00:48:55.100 --> 00:48:58.900 - Yeah, sure, so I think I'm right.

00:48:58.900 --> 00:49:01.890 Yeah, let me think we are, let me see

00:49:02.830 --> 00:49:04.540 I think in this case, we're correcting

00:49:04.540 --> 00:49:08.220 for approximately the same confounders here.

00:49:08.220 --> 00:49:09.910 So they already had some confounders

00:49:09.910 --> 00:49:10.743 that they were correcting for

00:49:10.743 --> 00:49:11.790 in the original negative binomial.

00:49:11.790 --> 00:49:13.910 So in that sense, it's a little bit more

00:49:13.910 --> 00:49:15.540 of maybe an apples to apples comparison.

00:49:15.540 --> 00:49:19.690 It's just a question of how do you calibrate

00:49:19.690 --> 00:49:20.790 that test statistic that is

00:49:20.790 --> 00:49:23.400 trying to correct for the confounders

00:49:23.400 --> 00:49:24.930 but I think what you're getting at

00:49:24.930 --> 00:49:26.890 I do think it can go either way.

00:49:26.890 --> 00:49:29.193 It's not obvious that Sceptre would make a P-value

00:49:29.193 --> 00:49:31.720 or they're more or less significant.

00:49:31.720 --> 00:49:34.340 I think I will say just as a small detail here

00:49:34.340 --> 00:49:38.280 that in addition to the negative binomial regression

00:49:38.280 --> 00:49:40.410 this P-value, it says,

00:49:40.410 --> 00:49:42.210 there's this strange word empirical here.

00:49:42.210 --> 00:49:43.400 What it means is that

00:49:43.400 --> 00:49:46.540 they've kind of also applied their fixed that they had

00:49:46.540 --> 00:49:48.448 because they realized that they had the miscalibration

00:49:48.448 --> 00:49:49.940 and then they kind of like smashed all

00:49:49.940 --> 00:49:52.086 of their P-values sort of,

00:49:52.086 --> 00:49:55.300 so these are sort of like, so in that sense

00:49:55.300 --> 00:49:56.730 it's not an apples to apples comparison

00:49:56.730 --> 00:49:58.470 but what we're doing is we're comparing

00:49:58.470 --> 00:50:00.100 to the P-values that were actually used

00:50:00.100 --> 00:50:02.096 for the analysis in this, in this paper.

00:50:02.096 --> 00:50:05.830 So maybe that makes it even harder to compare, but yes.

00:50:05.830 --> 00:50:08.430 So take this plot with a grain of salt, if you will.

00:50:09.900 --> 00:50:13.920 Perhaps I think the most exciting part is

00:50:13.920 --> 00:50:18.020 actually applying this to new gene-enhancer pairs

00:50:18.020 --> 00:50:20.770 where we don't know necessarily what the answer is.

00:50:20.770 --> 00:50:23.528 And so this plot just shows you

00:50:23.528 --> 00:50:26.500 we're just plotting it's actually, I guess

00:50:26.500 --> 00:50:29.740 similar to this plot we saw here

00:50:29.740 --> 00:50:32.820 except now we're looking at the candidate enhancers.

00:50:32.820 --> 00:50:35.950 And so essentially the different colors.

00:50:35.950 --> 00:50:38.490 So firstly, this also just shows you

00:50:38.490 --> 00:50:42.260 that this is very much not a monotonic transformation.

00:50:42.260 --> 00:50:47.160 Like you really can like, if you look into this quadrant

00:50:47.160 --> 00:50:50.750 this is an example where the original P-value was very

00:50:50.750 --> 00:50:53.690 not significant, but according to Sceptre

00:50:53.690 --> 00:50:56.883 it can be very significant and vice versa.

00:50:58.050 --> 00:51:00.740 So essentially I've just kind of highlighted

00:51:00.740 --> 00:51:03.140 those gene-enhancer pairs that were,

00:51:03.140 --> 00:51:05.420 found by one method and not the other.

00:51:05.420 --> 00:51:08.610 And so the upshot is that there's a total

00:51:08.610 --> 00:51:12.380 of about, roughly 500 or so found.

00:51:12.380 --> 00:51:15.350 Well, I guess after found 563

00:51:15.350 --> 00:51:17.860 of those 200 were new in the sense

00:51:17.860 --> 00:51:20.520 that they were not found by the original analysis.

00:51:20.520 --> 00:51:23.720 And then 107 were found by the original analysis

00:51:23.720 --> 00:51:25.860 but were not found by us.

00:51:25.860 --> 00:51:28.300 And we have strong reasons to believe

00:51:28.300 --> 00:51:30.100 that these could be false positives based

00:51:30.100 --> 00:51:33.913 on exactly the sorts of miscalibration that I presented.

00:51:35.060 --> 00:51:38.390 We did look at a few specific new discoveries here

00:51:38.390 --> 00:51:43.390 and found that they were corroborated by EQTL data.

00:51:43.400 --> 00:51:44.870 And for those of you who are familiar

00:51:44.870 --> 00:51:48.410 enhancer RNA correlation data, since I'm running low

00:51:48.410 --> 00:51:51.190 on time, I don't have time to explain this to you

00:51:51.190 --> 00:51:52.560 but these are all P-values

00:51:52.560 --> 00:51:56.313 of association based on orthogonal functional assets.

00:51:57.800 --> 00:52:00.770 Also, we found that our discoveries were more enriched

00:52:00.770 --> 00:52:03.600 for biological signals in a few different ways.

00:52:03.600 --> 00:52:06.020 One of them is that, and again,

00:52:06.020 --> 00:52:08.100 I'm sort of maybe going a little bit

00:52:08.100 --> 00:52:10.810 more quickly here 'cause I'm about to run out of time

00:52:10.810 --> 00:52:13.070 but there are these things called topologically

00:52:13.070 --> 00:52:16.070 associating domains, which are basically regions

00:52:16.070 --> 00:52:18.560 in the genome within which most

00:52:18.560 --> 00:52:21.680 of these regulatory interactions are thought to occur.

00:52:21.680 --> 00:52:24.970 And so what we find is that a greater fraction

00:52:24.970 --> 00:52:27.450 of the gene-enhancer pairs we found compared

00:52:27.450 --> 00:52:29.710 to the original analysis did lie

00:52:29.710 --> 00:52:32.380 in the same top logically associating domain.

00:52:32.380 --> 00:52:34.530 So in this case, 74% versus

00:52:34.530 --> 00:52:37.040 the 71% found in the original analysis.

00:52:37.040 --> 00:52:39.290 So in this sense, I mean, it's just kind of

00:52:39.290 --> 00:52:43.890 like a first order sense of biological plausibility.

00:52:43.890 --> 00:52:45.890 I think people are starting to think

00:52:45.890 --> 00:52:47.590 that there are interactions that are sort of

00:52:47.590 --> 00:52:49.160 outside of tabs as well.

00:52:49.160 --> 00:52:51.760 So I don't think this is a signal that,

00:52:51.760 --> 00:52:54.670 26% of these things are false discoveries

00:52:54.670 --> 00:52:59.360 but we definitely do expect, a high degree

00:52:59.360 --> 00:53:03.633 of enrichment for within tad interactions.

00:53:05.060 --> 00:53:06.640 Also if you do look

00:53:06.640 --> 00:53:08.950 at some of these more circumstantial pieces

00:53:08.950 --> 00:53:13.950 of evidence for regulations, such as things

00:53:13.970 --> 00:53:18.950 like transcription factor binding or histone modifications

00:53:18.950 --> 00:53:22.430 so we can use CHiP-seq to essentially assess

00:53:23.427 --> 00:53:27.150 for any given what

00:53:27.150 --> 00:53:32.150 whether there is these kind of signatures of regulation.

00:53:32.830 --> 00:53:35.430 And so what we found is that we did a little bit

00:53:35.430 --> 00:53:37.640 of an enrichment analysis where we looked at all

00:53:37.640 --> 00:53:40.240 of those enhancers that were found to be paired

00:53:40.240 --> 00:53:43.450 to genes by sector versus the original method

00:53:43.450 --> 00:53:45.980 and looked to what extent they were enriched

00:53:45.980 --> 00:53:49.270 for these other signatures

00:53:49.270 --> 00:53:51.730 these CHiP-seq based signatures of regulation.

00:53:51.730 --> 00:53:53.260 And what we found is that

00:53:53.260 --> 00:53:57.580 across eight of these CHiP-seq targets, and by the way

00:53:57.580 --> 00:53:59.450 these eight are not selected.

00:53:59.450 --> 00:54:02.930 These actually were the exact eight CHiP-seq targets

00:54:02.930 --> 00:54:05.830 that they examined in the original paper,

00:54:05.830 --> 00:54:08.450 we found greater enrichment.

00:54:08.450 --> 00:54:12.550 So in this sense, also the enhancers being

00:54:12.550 --> 00:54:14.810 picked up by Sceptre are just more biologically

00:54:14.810 --> 00:54:19.170 plausible using these orthogonal kinds of assets.

00:54:19.170 --> 00:54:21.490 So I find this very exciting

00:54:21.490 --> 00:54:24.550 and I'm just gonna maybe make a few remarks

00:54:24.550 --> 00:54:25.770 and hopefully there's just a little bit

00:54:25.770 --> 00:54:27.300 of time for questions.

00:54:27.300 --> 00:54:30.380 I will also be around for a few minutes after the seminar.

00:54:30.380 --> 00:54:32.950 If anyone wants to stick around and ask me questions

00:54:32.950 --> 00:54:34.770 you also might have your next thing to go to.

00:54:34.770 --> 00:54:36.173 So I understand if not.

00:54:37.310 --> 00:54:40.570 But maybe the summary is that, mapping gene-enhancer

00:54:40.570 --> 00:54:42.890 regulatory relationships is very important.

00:54:42.890 --> 00:54:47.460 If we wanna translate GWAS hits into disease insights.

00:54:47.460 --> 00:54:50.190 And there's been this very exciting new technology

00:54:50.190 --> 00:54:52.970 that allows us to answer that question.

00:54:52.970 --> 00:54:56.010 This technology was proposed very recently,

00:54:56.010 --> 00:54:59.610 and so there aren't that many methods out there

00:54:59.610 --> 00:55:01.800 to analyze these kinds of data.

00:55:01.800 --> 00:55:05.200 And so what we did with Sceptre is we leveraged recent

00:55:05.200 --> 00:55:08.250 methological advances in statistics to overcome the primary

00:55:08.250 --> 00:55:11.440 limitations of the parametric and non-parametric analysis

00:55:11.440 --> 00:55:13.360 methods that were available.

00:55:13.360 --> 00:55:17.170 And finally, we applied it to the largest existing

00:55:17.170 --> 00:55:18.660 data set of this kind.

00:55:18.660 --> 00:55:21.970 And what we get is a greater number of more biologically

00:55:21.970 --> 00:55:24.530 meaningful regulatory relationships.

00:55:24.530 --> 00:55:27.720 So I had a few other discussion slides, maybe I'll just

00:55:27.720 --> 00:55:30.660 read the title to you without getting into the details

00:55:30.660 --> 00:55:33.570 but this is a rapidly developing technology.

00:55:33.570 --> 00:55:36.930 And we do foresee that sector will be applicable

00:55:36.930 --> 00:55:40.020 to future iterations of the technology.

00:55:40.020 --> 00:55:42.230 So that's promising.

00:55:42.230 --> 00:55:45.800 And secondly, this is more like the beginning

00:55:45.800 --> 00:55:47.260 of the road than the end of the road.

00:55:47.260 --> 00:55:49.323 There are lots of remaining challenges,

00:55:50.550 --> 00:55:52.710 this includes looking for interactions

00:55:52.710 --> 00:55:56.160 among enhancers, things like dealing

00:55:56.160 --> 00:55:59.600 with multiple guidances, how are you in the same enhancer,

00:55:59.600 --> 00:56:02.430 they're just basically like a whole, I would say, playground

00:56:02.430 --> 00:56:05.603 of statistical problems that have yet to be addressed.

00:56:06.540 --> 00:56:11.290 So maybe finally, if you'd like to learn more

00:56:11.290 --> 00:56:13.390 we have a pre-printed on bio archive.

00:56:13.390 --> 00:56:15.890 I wanna acknowledge my co-authors again.

00:56:15.890 --> 00:56:20.240 And finally, so Tim has worked very well hard

00:56:20.240 --> 00:56:24.030 on putting, making this an art package so

00:56:24.030 --> 00:56:26.110 you can find out on GitHub

00:56:26.110 --> 00:56:28.940 and I'm very happy to take questions now

00:56:28.940 --> 00:56:32.230 but if you have any burning questions that come

00:56:32.230 --> 00:56:34.960 to you 30 minutes after my talk

00:56:34.960 --> 00:56:37.390 please feel free to email me at this address.

00:56:37.390 --> 00:56:40.050 So thank you, and I should have said at the top, thank you

00:56:40.050 --> 00:56:41.873 Lexi for the invitation.

00:56:41.873 --> 00:56:44.990 - Thank you for agreeing to present your work here.

00:56:44.990 --> 00:56:46.660 It's really a nice talk.

00:56:46.660 --> 00:56:47.510 - Yeah Thank you.

00:56:48.610 --> 00:56:52.370 - So I have some, maybe less related question

00:56:52.370 --> 00:56:55.770 to your current work, but maybe interesting to consider.

00:56:55.770 --> 00:56:57.440 I am not sure.

00:56:57.440 --> 00:56:59.400 Have you looked at the correlation structure

00:56:59.400 --> 00:57:02.500 between the X matrix?

00:57:02.500 --> 00:57:06.570 - Yeah, so essentially my sense is that gets

00:57:06.570 --> 00:57:11.400 like a factor model where you have all

00:57:11.400 --> 00:57:16.120 of these sort of confounders that are inducing correlation

00:57:16.120 --> 00:57:21.120 among all the axis, but essentially like once you account

00:57:21.403 --> 00:57:25.170 for that confounding, it's independent.

00:57:25.170 --> 00:57:28.043 - I see (indistinct) correlation.

00:57:29.240 --> 00:57:33.170 - So it's fairly small correlation and essentially

00:57:33.170 --> 00:57:35.870 the reason for, and this is very different from

00:57:35.870 --> 00:57:37.850 for example, genome-wide association studies.

00:57:37.850 --> 00:57:38.683 So it's like, Oh

00:57:38.683 --> 00:57:40.270 is there some analog of Lincoln's this equilibrium.

00:57:40.270 --> 00:57:43.487 And the key difference here is that

00:57:43.487 --> 00:57:46.240 it's essentially a design experiments.

00:57:46.240 --> 00:57:48.820 So even though you're not controlling exactly

00:57:48.820 --> 00:57:51.190 which cells receive what perturbations you are

00:57:51.190 --> 00:57:53.600 basically assigning them at random.

00:57:53.600 --> 00:57:54.880 So if it worked

00:57:54.880 --> 00:57:58.030 for this sort of pesky measurement mechanism business

00:57:58.030 --> 00:58:00.920 it would be an unconfounded problem.

00:58:00.920 --> 00:58:05.480 But essentially, so the only correlations are coming

00:58:05.480 --> 00:58:07.860 from this measurement.

00:58:07.860 --> 00:58:10.040 Yes so that is a great question

00:58:10.040 --> 00:58:11.120 because you can ask, well

00:58:11.120 --> 00:58:13.060 how did I do the slight of hand run?

00:58:13.060 --> 00:58:14.960 Like slide three all of a sudden I was working

00:58:14.960 --> 00:58:16.580 with like one enhancer

00:58:16.580 --> 00:58:18.300 and where did all the rest of them go.

00:58:18.300 --> 00:58:21.430 And I think we're actually not losing all too much

00:58:21.430 --> 00:58:22.960 by doing this, especially

00:58:22.960 --> 00:58:25.430 since we are controlling for those technical factors.

00:58:25.430 --> 00:58:28.140 - Yeah thanks that makes sense to me.

00:58:28.140 --> 00:58:30.591 And another thing is maybe more, less than less

00:58:30.591 --> 00:58:34.350 statistical is how many confounding factors

00:58:34.350 --> 00:58:35.490 they are controlling

00:58:35.490 --> 00:58:39.159 and what are the important ones that you have identified?

00:58:39.159 --> 00:58:40.900 - Yeah, I mean, so in this case

00:58:40.900 --> 00:58:43.870 we're doing essentially we're following the lead of

00:58:44.880 --> 00:58:45.870 the original paper

00:58:45.870 --> 00:58:47.700 for which confounding factors with control for.

00:58:47.700 --> 00:58:50.520 So in addition to sequencing depth.

00:58:50.520 --> 00:58:51.970 Yeah, so they do have a batch of fact

00:58:51.970 --> 00:58:54.820 and there's also something called Percent Might've Country.

00:58:54.820 --> 00:58:58.210 So it's like what fraction of all the reads that you got

00:58:58.210 --> 00:59:01.760 in this particular cell came from mitochondrial DNA

00:59:01.760 --> 00:59:06.760 as opposed to, regular DNA, maybe a few others

00:59:08.560 --> 00:59:09.850 like just total number

00:59:09.850 --> 00:59:12.744 of genes expressed in the cell, things of this nature.

00:59:12.744 --> 00:59:14.550 So I think here we're correcting

00:59:14.550 --> 00:59:18.300 for about five, but you could think of other things

00:59:18.300 --> 00:59:23.300 like cell cycle, this is a pretty K five 62 is a pretty

00:59:25.210 --> 00:59:27.340 homogeneous cell line, but especially

00:59:27.340 --> 00:59:30.380 once you get to other kinds of, tissue samples

00:59:30.380 --> 00:59:33.040 you might need to think about, cell type

00:59:33.040 --> 00:59:35.140 and things of this nature.

00:59:35.140 --> 00:59:38.110 So I think there are lots to consider here,

00:59:38.110 --> 00:59:40.870 we used kind of five easy ones.

00:59:41.907 --> 00:59:42.913 - Okay, thanks.

00:59:44.440 --> 00:59:46.537 Any more questions for Eugene?

00:59:48.030 --> 00:59:50.790 Yeah, I think we are approximating

00:59:52.191 --> 00:59:55.480 the end of the talk, the seminar.

00:59:55.480 --> 00:59:58.340 So thanks again for your great talk.

00:59:58.340 --> 01:00:00.630 And if you have any further questions

01:00:00.630 --> 01:00:04.970 you can just send emails to Eugene offline.

01:00:04.970 --> 01:00:07.697 - Yes, yes, definitely don't hesitate to reach out.