First good afternoon, everyone, and I hope you somehow managed to enjoy your winter break.

And this is our first talk, seminar talk this semester, and we have invited Dr. Eugene Katsevich from Wharton School at UPenn. And he’s going to present something really exciting, I know his original work on statistical analysis single cell CRISPR screening. And I will hand it over to Eugene from now, from here.

If people will filter in, maybe I’ll wait another minute or two, 'cause I think, I feel like the first part of the talk is very important. So I think if people missed the first part of the talk, then it’ll be maybe hard to follow along later. I’m happy to wait just another minute or two. I understand perfectly that it’s a strange time for everyone, so for all those who were able to make it today, I really appreciate your adjusting the schedule.

Also maybe one remark I can make is that, since it is a smaller audience, we can make this seminar just about as interactive as you want. So you should definitely feel free to stop me at any point. I don’t know how many of you are familiar
0:01:40.37 –> 0:01:43.44 with the CRISPR screen stuff I’m gonna talk about,
0:01:43.44 –> 0:01:47.533 but I’m very happy to just make it very interactive.
0:01:54.285 –> 0:01:58.367 I will maybe start sharing my screen
0:01:58.367 –> 0:02:00.04 and maybe I’ll start launching
0:02:00.04 –> 0:02:04.023 into some of the introductory things.
0:02:12.73 –> 0:02:13.9 So…
0:02:16.27 –> 0:02:18.123 Oh wow, wait, is this the...
0:02:20.66 –> 0:02:23.91 I greatly apologize.
0:02:23.91 –> 0:02:28.3 Clearly, the label on my slides is wrong.
0:02:28.3 –> 0:02:30.52 I have updated my slides since then,
0:02:30.52 –> 0:02:32.73 but I think the title page has not been updated,
0:02:32.73 –> 0:02:34.503 that’s extremely embarrassing.
0:02:39.807 –> 0:02:43.36 Well, maybe then I should skip past this slide very quickly.
0:02:43.36 –> 0:02:47.26 So hello everyone, thank you so much
0:02:47.26 –> 0:02:48.723 for making it to my talk.
0:02:49.76 –> 0:02:52.44 Today, I’ll be talking about some Statistical Analysis Tools
0:02:52.44 –> 0:02:54.59 for Single Cell CRISPR Screens.
0:02:54.59 –> 0:02:56.24 So the most important thing to take away
0:02:56.24 –> 0:02:58.89 from this slide are my collaborators here.
0:02:58.89 –> 0:03:02.1 So Tim Barry is a grad student
0:03:02.1 –> 0:03:03.723 of mine who was actually at CMU.
0:03:05.34 –> 0:03:07.67 I am jointly advising him with Kathryne Roeder
0:03:07.67 –> 0:03:12.67 also at CMU, who is used to be my postdoc advisor.
0:03:15.06 –> 0:03:17.163 So I’ll skip quickly to the next slide.
0:03:20 –> 0:03:22.06 So here’s the motivation.
0:03:22.06 –> 0:03:25.16 And by the way, if anyone has joined recently,
0:03:25.16 –> 0:03:27.853 please just stop me at any point.
0:03:28.99 –> 0:03:29.98 So here’s the motivation.
0:03:29.98 –> 0:03:31.49 So we have done lots
0:03:31.49 -> 0:03:34.54 and lots of genome wide association studies to date.
0:03:34.54 -> 0:03:37.1 So we have a lot of little markers
0:03:37.1 -> 0:03:41.43 along the genome that we think are associated with
diseases.
0:03:41.43 -> 0:03:43.19 And so the question is what’s the next step?
0:03:43.19 -> 0:03:45.96 Like how do we actually translate these
0:03:45.96 -> 0:03:48.75 into insights into diseases?
0:03:48.75 -> 0:03:50.8 And hopefully later on things like,
0:03:50.8 -> 0:03:52.04 therapeutics and so on.
0:03:52.04 -> 0:03:55.68 So what we need to do is we need to understand how
0:03:55.68 -> 0:03:57.69 like basically the mechanisms
0:03:57.69 -> 0:04:00.87 by what mechanism are these associations actually
resulting
0:04:00.87 -> 0:04:02.64 in an increased disease risk.
0:04:02.64 -> 0:04:04.71 So here’s a typical situation here
0:04:04.71 -> 0:04:07.76 as our genome and here’s a disease association
0:04:07.76 -> 0:04:10.66 and frequently these disease associations
0:04:10.66 -> 0:04:13.33 they might not take place within genes.
0:04:13.33 -> 0:04:16.103 And so that makes them pretty hard to interpret.
0:04:17.5 -> 0:04:22.5 So what’s hypothesized to be the case here is that instead
0:04:23.56 -> 0:04:28.56 of disrupting genes directly, these variants
0:04:29.3 -> 0:04:33.33 are disrupting regulatory elements such as enhancers.
0:04:33.33 -> 0:04:36.09 So let’s just like briefly here review
0:04:37.503 -> 0:04:40.8 that an enhancer is a region of the genome.
0:04:40.8 -> 0:04:42.33 That could be a certain distance
0:04:42.33 -> 0:04:45.16 from the gene that actually folds
0:04:45.16 -> 0:04:46.97 in three-dimensional space to come
0:04:46.97 -> 0:04:51.97 in close proximity to the promoter of the gene.
0:04:52.25 -> 0:04:54.93 And essentially the enhancers job is to recruit a lot
0:04:54.93 -> 0:04:56.61 of the machinery that actually is going to lead
0:04:56.61 -> 0:04:58.45 to the expression of this gene.
So if you disrupt the enhancer then this will disrupt the recruitment of all of these different transcription factors which will then end up causing some trouble. And so it’s this sort of like, for example, in this case let’s say that this disease association as is disrupting enhanced or one, well, this might suggest if enhancer one is regulating gene two, that the disease mechanism is actually proceeding essentially or being mediated by the expression of gene too. And so this would be a very great and clean way of interpreting GWAS hits. But the problem is that we don’t actually know or we have a very hazy sense of which enhancers actually regulate which genes. So this is kind of a difficult problem for a few different reasons. The first reason is that there’s a potentially many to many mapping between enhancers in genes. So in enhancer it can regulate multiple genes and a single gene can be regulated by multiple enhancers. So the other thing is that any answers don’t even need to be all too close to the genes that they regulate. There could be situations like we saw here where the regulation can skip the adjacent gene and go to the next one. And so in general regulations can be thought to happen within about a megabase distance in terms of the linear distance in the genome.
So this is a hard problem, and it’s basically the motivating problem for this talk which enhancers regulate which genes. This is a sort of a very fundamental and important problem in genomics.

So in today’s talk, I’m going to first talk about a new assay called a single cell CRISPR screen that allows us to get at this question, then I’m gonna talk about the challenges that previous methods have encountered in analyzing these single cell CRISPR screen datasets, never propose a new methodology based on this idea of conditional resampling.

And then I will show you how this works on real data and close with the discussion.

So let me first introduce the biological assay here which is called the Single Cell CRISPR screen.

So actually backing up a second, this is a very important problem and people have considered it before.

So how do people typically approach gene-enhancer mapping? I think the most common approach is what I call here an indirect observational approach.

And there are many of these. So what this picture is, is a basically a more detailed picture of what happens when an enhancer or a pictured here comes into contact with the promoter of a gene.

There are lots of kind
0:08:01.32 → 0:08:05.51 of indirect signals of this regulation.
0:08:05.51 → 0:08:08.027 Obviously you have just the actual expression
0:08:08.027 → 0:08:12.31 of the gene, but you’ll have the confirmation
0:08:12.31 → 0:08:16.31 of the chromatin in the vicinity of the promoter
0:08:16.31 → 0:08:18.06 and in the enhancer
0:08:18.06 → 0:08:21.89 you have basically transcription factor binding data.
0:08:21.89 → 0:08:25.32 And all of these data are essentially indirect ways
0:08:25.32 → 0:08:28.49 of trying to make a conclusion
0:08:28.49 → 0:08:31.22 about which enhancers might be regulating which
genes.
0:08:31.22 → 0:08:33.147 So for example, using high C data
0:08:33.147 → 0:08:36.52 if you find an enhancer to be a 3D contact
0:08:36.52 → 0:08:39.235 with the promoter, then this could be a single signal
0:08:39.235 → 0:08:41.973 that there is some regulation going on.
0:08:43.35 → 0:08:45.03 The issue is that these approaches have not
0:08:45.03 → 0:08:47.42 proved very reliable at the end of the day.
0:08:47.42 → 0:08:49.04 These are observational approaches,
0:08:49.04 → 0:08:51.45 and basically even if you have
0:08:52.37 → 0:08:55.593 contact in 3D space, this is not necessarily a signal.
0:08:55.593 → 0:08:58.54 This doesn’t necessarily mean that regulation
0:08:58.54 → 0:08:59.5 is actually occurring,
0:08:59.5 → 0:09:03.63 and so essentially we haven’t gotten all too far
0:09:03.63 → 0:09:05.49 with these indirect approaches.
0:09:05.49 → 0:09:07.02 So the exciting thing is
0:09:07.02 → 0:09:12.02 that recently with the development of CRISPR tech-
nology
0:09:12.02 → 0:09:17.07 we can now actually go in and instead of observationally
0:09:17.07 → 0:09:19.7 just essentially take a look inside a cell.
0:09:19.7 → 0:09:23.69 We can actually go in and make modifications where we
0:09:23.69 → 0:09:27.9 for example, knockouts enhancers using the system
0:09:27.9 → 0:09:29.77 called CRISPR Interference.
And then we try to look at what the results are for gene expression. So this shows you a little cartoon of the CRISPR interference system. And so the way that it works is that you have this CAS nine protein whose job is to attach to a certain segment of DNA. And the specific segment of DNA it attaches to is specified by this guide, or I do. And so in this way, the attachment can be highly specific to the sequence of the enhancer. And then this for CRISPR Interference the CAS nine brings along with it all of these repressive elements that essentially knock out this enhancer, meaning they prevent the enhancement from actually helping to regulate this gene. And so the idea, so firstly this is a promising solution because it allows us to interrogate these regulatory relationships in a much more direct way than we’ve been able to do until recently. And so the overall idea is that, it’s the idea of simple disrupt enhancers and see which genes expression drops. And so just as a cartoon here, let’s say we knock out this enhancer, then we would expect to see the gene that regulates to be down-regulated. And then we can think
about designing perturbations for multiple enhancers. And so if you perturb this enhancer then maybe you’ll see a response in these two genes. Very naive question, just to make sure I didn’t misunderstand notion here is enhancer always upregulating gene kind of regulate? I think enhancers specifically are thought to upregulate genes. However, it’s a good question because there are other kinds of elements that are, can actually be silencers for example. And so that’s just another example of a kind of regulatory element. So the effect could go in either direction and this talk I’ll primarily talk about enhancers but really everything I say goes through for other kinds of regulatory elements. Thanks. Yeah, very good question. So now the actual assay. That allows you to do this out of large scale. So the scale is the question here because you can do CRISPR experiments where you essentially like knock out one enhancer in a whole batch of cells, and then, maybe go enhancer by enhancer and this ends up not being a very scalable approach. So there has been proposed this new asset called the single cell CRISPR screen in which you basically pool a whole bunch of cells.
0:12:21.55 –> 0:12:23.35 of perturbations together,
0:12:23.35 –> 0:12:26.33 and then the readout that you get is single cell
0:12:26.33 –> 0:12:29.64 RNA sequencing, which allows you to also basically
look
0:12:29.64 –> 0:12:30.75 at the impact of all
0:12:30.75 –> 0:12:32.48 of those different enhancement perturbations
0:12:32.48 –> 0:12:34.54 on the entire transcriptome.
0:12:34.54 –> 0:12:37.57 And so in the slide, I’m gonna give you a brief overview
0:12:37.57 –> 0:12:40.32 of how these screens work.
0:12:40.32 –> 0:12:41.6 So first way you do is you start
0:12:41.6 –> 0:12:44.26 with a library of CRISPR perturbations.
0:12:44.26 –> 0:12:47.583 So you just, let’s say maybe you take,
0:12:49.227 –> 0:12:51.763 10,000 enhancers across the genome
0:12:51.76 –> 0:12:54.613 and then you basically design CRISPR guide.
0:12:54.613 –> 0:12:56.893 RNAs targeting each of those enhancers.
0:12:57.86 –> 0:13:00.19 Once you have a library of these perturbations
0:13:00.19 –> 0:13:03.13 you then infect a big pool
0:13:03.13 –> 0:13:05.91 of cells with all of these perturbations.
0:13:05.91 –> 0:13:07.64 And so what’s important to note here is
0:13:07.64 –> 0:13:12.64 that essentially these perturbations get randomly inte-
grated
0:13:13.36 –> 0:13:17.28 into the different cells they’re delivered through a
0:13:17.28 –> 0:13:18.734 like a virus system
0:13:18.734 –> 0:13:21.88 the details aren’t very important, but the importance is
0:13:21.88 –> 0:13:23.78 that these perturbations get integrated
0:13:23.78 –> 0:13:25.72 into cells essentially at random.
0:13:25.72 –> 0:13:27.69 And so each cell gets its own collection
0:13:27.69 –> 0:13:30.04 of CRISPR perturbations.
0:13:30.04 –> 0:13:34.76 So now in order to basically actually read out what happened
0:13:34.76 –> 0:13:37.64 in our experiment, we use single cell RNA sequencing.
And as a result of the sequencing experiment we get two pieces of information, firstly, by the way two pieces of information for every step. So for every cell we first measure the perturbations that are present. So which of these guide or nays did we detect, and then secondly the gene expression for the whole transcriptome. So this is essentially our data here. And then once we have this data we can now do the analysis component, which really ends up being a kind of differential expression analysis. So consider a particular gene-enhancer pair. So what we can do is we could take all of the cells and we can break them up into two groups. Those cells for which that enhancer was knocked out which are in orange here, and those cells for which that enhancer was not knocked out. We can then split, essentially look at the expression of the gene of interest and see whether there’s a systematic difference between the expression of this gene and these two populations of cells. So, and then if there is a significant difference then we can make a conclusion that that particular enhancer is regulating that particular gene. So it seems quite simple on first glance, but this analysis part actually turns out to be a challenging statistical problem. And so the analysis of these screens is actually the subject of this talk.
Okay so, maybe one more slide and then I'll stop and see if people have questions.

So just to make it a little bit more concrete, there's a kind of a large dataset that might be one of the largest out there right now by Gasperini at all. It was published in Cell last year.

Oh wow, I guess two years ago now to 2019, and so they were working with 200,000 K five 62 cells and they were looking at 6,000 candidate enhancers. And so they're looking at, I mean essentially the whole transcriptome, at least the part of it that has any expression in the cell type. And they identified 85,000 enhancer gene pairs that they essentially thought were plausible to have some regulation and in their experiment they had 28 per patients on average per cell.

And so the way that this data would look is, think about the rows as being the cells and then the columns. So you have two groups of columns. Firstly, you have the gene expressions, and so since these are single cell data, we have these highly discreet counts of reeds or UMRs for every gene.

And then also we have the second bit of information which tells you which cells received, which perturbations. So in general, in this presentation, I'll talk I'll denote gene expression by Y and perturbations by X.

And so there's also a third and very important piece of information, which are technical factors per cell.
Perhaps the main one that I’ll talk about today is the sequencing depth. So this is just the total number of reads or UMRs I measured from this cell. And so this basically just varies randomly across cells just as an artifact of your experiment. There are other technical factors like batch and so on and so forth. Okay, so this brings me to the end of the first section where I tell you about the data and the asset. So are there any questions before I move on to talking more about the analysis of these types of data. I’m assuming there are no questions but do feel free to stop me if there are. So as I said, this actually turns out to be kind of like an annoyingly challenging statistical problem. And so to illustrate this to you, let me first give you a sense of what analysis methods there are out there. So are there any questions before I move on? I should say, by the way that given the sort of the novelty of this assay, there hasn’t been a lot of work in terms of designing methods specifically designed for this kind of data. So most of the existing analysis methods are basically proposed by the same people who are producing the single cell CRISPR screen data. So by the way, so in this slide I’m going to actually for the remainder of the talk...
I’m actually going to essentially focus our attention on a certain gene and a certain enhancer and just consider the problem and figuring out whether that enhancer regulates that gene.

And so I’m gonna use $Y_i$ to denote the expression of that gene and cell $X_i$ as the binary indicator for whether that enhancer was perturbed in that cell and $Z_i$ the vector of these extra technical co-variants. So With that notation out of the way, the first kind of popular method for analyzing these data is negative binomial regression. For those of you familiar with bulk RNA-seq differential expression analysis, this is similar to the DESeq2 methodology where you just run a negative binomial regression of the gene expression, $Y$, on a linear combination of the perturbation indicator, as well as all of your technical co-variants. And so Negative Binomial is a common model for these sort of over dispersed count data that you encounter in RNA sequencing data. Okay, next, there is a rank based approach. So this is non-parametric where it’s actually much simpler. You just, you cross tabulate yourselves by two criteria. First, you see whether they have the perturbation or not. And second, you see whether they have essentially higher than median expression on this gene or lower than median expression on this gene.
And then you do a two by two table test for independence.

And finally there are also permutation based approaches where the idea is to take some test statistic and then calibrate it under the null distribution by permuting this column right here the assignments of the perturbations to the cells.

So yes, that, I guess that’s, what’s written here.

So okay, there’s like maybe all these methods sound reasonable at first, but the more you actually look at the existing literature the more there are various scattered signs like none of these methods are like really doing the trick.

And so here are the methods that I described on the previous slide.

So virtual FACS is the rank based one and scMAGECK is the one of the permutation based ones.

And so you look at plots actually from the original papers themselves who propose these methods and you see some signs of miscalibration.

And so like, for example, I’m gonna be talking mostly about this data and to a lesser extent about this data in my talk, but so looking here.

So a Negative Control Perturbation is a guide but it’s actually not designed to target any particular sequence along the genome.
So you don’t expect cells that are infected with a negative control perturbation to look any different from cells that have no perturbation. And so in this Gasperini data, they have 50 different negative control guide RNAs, and so what they did is they basically plotted a QQ plot paired with all of the genes and the genome, and what they found is perhaps on this QQ plot this doesn’t look like a severe inflation from uniformity but it’s important to keep in mind the scale of this Y axis. And so essentially this amounts to a massive amount of deviation from the uniform distribution in those P-values. So in other words, negative control, gene-enhancer pairs are looking incredibly significant according to this analysis. So in this particular analysis they essentially found the same thing here it’s portrayed as a Manhattan plot but you see a lot of things reaching significance when right only the circle points are those that essentially were replicated in a bulk RNA sequencing experiment. And then this one finally looks like they perturbed lots of different enhancers and essentially looked at the effect on this one particular gene. And essentially what they found is that essentially all of the enhancers that they tested appeared to
0:23:21.84  ->  0:23:24.53 actually be per, like, have an effect on the expression
0:23:24.53  ->  0:23:28.612 of this gene, when in fact this is biologically impossible.
0:23:28.612  ->  0:23:31.75 So this is clearly an issue.
0:23:31.75  ->  0:23:36.11 Now, these original papers clearly knew
0:23:36.11  ->  0:23:39.06 that there was an issue, and so for each of the papers
0:23:39.06  ->  0:23:40.26 they kind of have a little bit
0:23:40.26  ->  0:23:44.64 of an ad hoc fix in order to basically correct their
P-value
0:23:44.64  ->  0:23:48.223 of distributions, so that they look a little bit more,
0:23:50.09  ->  0:23:52.1 closer to being calibrated.
0:23:52.1  ->  0:23:54.9 And so I’m, I think for the sake of time
0:23:54.9  ->  0:23:57.88 I’m probably not going to get into exactly how
0:23:57.88  ->  0:24:01.74 they propose to fix their P-value distributions.
0:24:01.74  ->  0:24:05.64 What I will say is that we looked in detail
0:24:05.64  ->  0:24:08.11 especially at the strategy that they use here
0:24:08.11  ->  0:24:09.96 and to a lesser extent at the strategies.
0:24:09.96  ->  0:24:10.91 Well, actually I think here
0:24:10.91  ->  0:24:13.72 they basically said just not to apply their method
0:24:13.72  ->  0:24:18 to data where there’s too high, essentially
0:24:18  ->  0:24:20.65 to where they’re too many perturbations per cell.
0:24:20.65  ->  0:24:23.65 So in this case, they just said, don’t apply this method.
0:24:23.65  ->  0:24:25.92 We looked into the kinds of fixes that they proposed
0:24:25.92  ->  0:24:28.05 in these two papers, and they essentially
0:24:28.05  ->  0:24:31.04 they don’t quite work in the way that you would expect.
0:24:31.04  ->  0:24:33.36 And so what we thought is that,
0:24:33.36  ->  0:24:36.81 what we’d like to do is kind of look a little deeper
0:24:36.81  ->  0:24:38.85 into this problem and try to ask ourselves
0:24:38.85  ->  0:24:40.9 why are we seeing all of these issues?
0:24:40.9  ->  0:24:43.74 Why do people keep running into these miscalibration issues
0:24:43.74  ->  0:24:48.74 and let’s try to basically address those underlying is-

16
So we thought about it a little bit and we thought about challenges for both parametric and non-parametric methods. So for parametric methods this actually shouldn’t really come as a surprise probably to most people here, gene expression is known to be pretty hard to model in single cells. So of course we have these essentially highly discreet lots of zeros counts that are over dispersed perhaps more importantly, given how sparse the data are. It’s actually pretty hard to get a good estimate of that dispersion parameter. And so there’s currently no standard way of estimating that dispersion parameter and basically every paper, comes up with their own way of doing this. They’re even just debates about what parametric models are appropriate for these data, should they be zero inflated, should they not be, and some genes have even been observed to have bi-modal expression patterns. So essentially all of these things are telling us that it’s kind of hard to shoe horn single cell gene expression, into a nice, neat parametric model. So obviously if you have missed specification of your model such as a bad estimate for a dispersion perimeter
that very well could cause miscalibration of the kind that we saw. So next we can think about non-parametric methods. So maybe, obviously if these data are hard to model parametrically maybe the non-parametric methods are going to save us. But the observation that we made that I think is quite important is that these technical factors that I mentioned before, like sequencing depth, they impact not only the expressions of genes but also the detection of these CRISPR guider in is. So I might have led you to believe in one of my early slides that we can basically perfectly measure which cell contains which CRISPR perturbations, but this is actually not true. So single cell RNA sequencing it’s essentially just like this kind of a sampling process. And so the more reads you sample from a cell the more likely you are to detect a guide RNAs. And so we just essentially looked at, for example, this is for one of the datasets and we just made a scatterplot of the total number of guide RNAs detected per cell versus the total number of UMI. So this is the sequencing depth and we found this extremely clear I guess I’m not showing you the P-value but this P-value was like absurdly significant to just basically confirm that if you have more sequencing depth in a cell,
0:27:33.26 -> 0:27:36.003 you’re going to find more guide our news in that cell.
0:27:36.87 -> 0:27:39.54 And so the issue with this is
0:27:39.54 -> 0:27:43.2 that we basically have a confounding problem on our hands.
0:27:43.2 -> 0:27:47.52 So think about this graphical model that’s illustrating
0:27:47.52 -> 0:27:48.57 what’s going on
0:27:48.57 -> 0:27:51.55 in a single cell CRISPR screen experiment in
0:27:51.55 -> 0:27:55.96 this gray box is kind of the underlying biological reality.
0:27:55.96 -> 0:27:58.408 Let’s say we have this presence of this guide RNA
0:27:58.408 -> 0:28:01.901 and the expression of this gene and the guide RNA is
0:28:01.901 -> 0:28:05.47 or the, yeah, I guess the, the CRISPR knockdown
0:28:05.47 -> 0:28:08.19 of the enhancer is either affecting gene expression
0:28:08.19 -> 0:28:12.67 or it is not, but we read it out.
0:28:12.67 -> 0:28:17.61 Some essentially imprecise the measurement
0:28:17.61 -> 0:28:18.84 of the guide RNA presence.
0:28:18.84 -> 0:28:20.21 We also read out
0:28:20.21 -> 0:28:23.1 and imprecise measurement of the gene expression.
0:28:23.1 -> 0:28:27.19 And what’s most important is that the technical factors such
0:28:27.19 -> 0:28:29.63 as sequencing depth, they’re actually impacting both
0:28:29.63 -> 0:28:32.69 of these measurements, they’re coming from the same cell.
0:28:32.69 -> 0:28:36.98 And so even if there is no association between the guide RNA
0:28:36.98 -> 0:28:41.61 and the gene, if you just basically naively look
0:28:41.61 -> 0:28:44.65 at the association between the measured guide RNA presence
0:28:44.65 -> 0:28:46.14 and the measured gene expression
0:28:46.14 -> 0:28:50.21 you’re going to find some association.
0:28:50.21 -> 0:28:52.64 And so this is clearly an issue.
0:28:52.64 -> 0:28:55.06 And so essentially in order to correct
0:28:55.06 -> 0:28:56.9 for this confounding effect, it’s very important

19
to test instead of just testing independence between the perturbation and the expression. We want to test conditional independence, where we're conditioning on all of these technical factors. And so this shows you why non-parametric methods tend to suffer is because when you do things like permute your data or rank your data, there's this underlying assumption that all of the cells are exchangeable and you're using that exchange ability to build your inference on. And so when you do those tests, they're implicitly actually testing just the direct independence the unconditional independence. And so this sort of inflation we saw in the non-parametric methods be explained by this Source of confounding. So that's actually it for that part of my talk. Any questions about the existing methods and the analysis challenges and why there's a need to think about new methodology for this problem. Okay, I will move on. So this is the part of the talk where I'm going to propose a new analysis method for this kind of data. And so the key kind of idea we're gonna use is conditional resampling, which is proposed by not us. So the idea of the conditional randomization test well, it's actually, depending on how you look at it, it's quite an old idea and it has some connections to causal inference, but it was proposed also incandescent all.
And essentially the setup is that you want to test conditional independence and you’re under the assumption that you have a decent estimate of the distribution of $X$ given $Z$. So remember $X$ is the perturbation. $Y$ is the expression and $Z$ are the essentially the confounders. So one way of thinking about it from a causal inference standpoint is let’s say we know the propensity score, can we test whether there’s a causal relationship between $X$ and $Y$ sort of controlling for these Confounders? So the idea of the conditional randomization test is the following. First, you take any test statistic $T$ of your data, and in order to calibrate this test statistic under the null hypothesis, instead of doing a permutation we’re gonna do a slightly more sophisticated resampling operation, where we’re going to go through, and for every cell, we are going to resample whether or not it received the given perturbation, but conditionally on the specific technical factors that were in that cell. And here we’re using crucially the information that we have a handle on what this sort of propensity score is. And then we’re just going to recompute the test the same test statistic on the resample data. And then we’re just gonna define the a P-value in the usual way for a resampling based procedure. So one way of thinking about it is
that it’s kind of like a permutation test, but it’s one in which the reassignments of the guide RNAs to the cells is one that respects the confounding that there is in the data instead of treating all the cells exchangeable.

So this is great because the CRT adjust for confounders basically by construction and importantly it avoids assumptions on the gene expression distribution. And in fact, provably, the P-value you get out of the CRT is valid, even if essentially, even if the test statistic T is, anything you want. So in the sense that kind of addresses the confounding issues, like basically the Achilles heel of the non-parametric methods, but avoiding assumptions on the gene expression distribution.

And it kind of seems to be doing something that’s avoiding both of those issues. Now, of course, there’s a, trade-off in the CRT does require you to have some estimate of this propensity score. So, and then secondly, the CRT is computationally expensive if you consider, or if you compare it to like just like a parametric regression here we’re doing a parametric regression but we’re doing it lots of times. And so how do we get around some of these issues?
So, and in particular, how do we actually go about applying this idea to single cell CRISPR screens? And so, firstly, do we understand this distribution of the probability of observing a guide or in a given a set of technical factors? So what we’re going to do in this particular method, well, first we’re gonna observe that it’s this is kind of a simpler phenomenon than gene expression like guide our nays are not really, like subject to all of the complicated regulatory patterns of genes. And secondly, kind of under the hood, the actual assortments of guide our nays to cells is, you know, like fairly well modeled. It’s just basically like in that sense the cells are pretty exchangeable. What’s not exchangeable it just basically this measurement process. So this is just kind of a simpler object in the specific case of single cell CRISPR screens. So we can try to bring to bear various knowledge to try to get a good sense of this in this case, we’re just gonna sort of do the easiest thing possible and we’re gonna fit it using an logistic regression. The second thing we’re going to do is think about what test statistic to use. So I had the separate paper about essentially the power of the conditioner randomization tests. And what we found is that the closer the test statistic is to the true conditional distribution of Y given X, Z
I guess I should say the true likelihood, the better the power will be. And so in that sense, what we wanna do is we wanna leverage existing models that people have used such as negative binomial regression. It's not going to matter whether the model is true or not for the sake of type one error control, but we hope that we can do a better job in terms of power by trying to get a good model for this.

And finally, how do we mitigate the computational cost? And so we had a few ideas for this as well. So one of them is called the distilled CRT. So I'll if time permits, which might or might not I'll give you a few more details about how you can use this to have a much faster for every resample to be quick. And then we're also going to use this hack, essentially that what we found is that the resampling distribution actually kind of looks pretty reasonable. It kind of looks like a normal, but it’s sort of how some extra skew and maybe some extra heavy tails. And so what we’re gonna do is we're going to fit a skew T distribution to the essentially the empirical distribution of the resample test statistics. And in that way, we can get more accurate P-values without doing as many recent samples. And so putting together all of these pieces we get this method, which we call Sceptre.
0:37:05.58 –> 0:37:08.4 or single cell perturbation screen analysis
0:37:08.4 –> 0:37:10.103 via conditional resampling.
0:37:11.01 –> 0:37:12.92 And so essentially what we do is what I said
0:37:12.92 –> 0:37:14.65 on the previous slide.
0:37:14.65 –> 0:37:18.84 We first use a logistic regression to fit a probability
0:37:18.84 –> 0:37:21.983 for every cell that we would find a perturbation there.
0:37:22.97 –> 0:37:25.5 And then we’re gonna use these perturbation probabilities
0:37:25.5 –> 0:37:28.003 and resample this particular column.
0:37:28.96 –> 0:37:32.47 And so we now we have a whole bunch of resample datasets.
0:37:32.47 –> 0:37:34.86 Now we’re going to use a negative binomial regression
0:37:34.86 –> 0:37:37.75 or more precisely a distilled negative binomial regression
0:37:37.75 –> 0:37:41.57 for speed, to get the test statistic
0:37:41.57 –> 0:37:42.81 for both the original data.
0:37:42.81 –> 0:37:45.593 And for all of these re resample datasets.
0:37:46.5 –> 0:37:47.71 Then we’re gonna put together all
0:37:47.71 –> 0:37:51.07 of these recycled test statistics into this gray histogram.
0:37:51.07 –> 0:37:54.11 And again, we’re gonna fit this magenta curve
0:37:54.11 –> 0:37:56.63 which is the skew T distribution
0:37:56.63 –> 0:37:59.02 which seems to fit pretty well in most cases.
0:37:59.02 –> 0:38:01.54 And then we’re gonna compare the original test statistic
0:38:01.54 –> 0:38:06.54 against this skew T distribution and get a P-value that way.
0:38:06.76 –> 0:38:09.89 And so this is represented by the shaded region here.
0:38:09.89 –> 0:38:14.06 And I think what’s noteworthy is to compare this fitted
0:38:14.06 –> 0:38:15.37 and all No distribution
0:38:15.37 –> 0:38:19.07 to this standard normal No distribution.
0:38:19.07 –> 0:38:20.24 I guess I should have said here
0:38:20.24 –> 0:38:24.02 that the actual test statistics are a Z values extracted
0:38:24.02 –> 0:38:25.81 from the negative binomial regression.
So if your model were true, the Z values under the No would follow a standard normal distribution. And so what we find is that when we resample we get something that’s not the standard normal distribution. And so in the sense you can view it as, a sort of measure of the departure sort of from, or sort of the lack of model fit that went into this negative binomial regression. So another way of putting this is that you can imagine that if you did happen to correctly specify your negative binomial regression model then you would sort of be getting back the same P-value that you would have gotten otherwise. So in that sense we’re not really reinventing the wheel here if you do have a good parametric model, but if you don’t then we can correct for it using this resampling strategy. So I guess this is an important slide so maybe I will stay here for a little bit and ask if anyone has questions about how our methodology works.

Hi, I have a bunker question. So have you tried to hurdle model to deal with this kind of full data is the cause of the weird distribution of the data? Oh, so let’s see. You mean to model the, essentially to model the gene expressions or do you mean to model the CRISPR perturbations?
From this page, first step you use a logistic regression and then you use a nickname by knowing that binomial. So it’s like a two step models, but to hurdle model they combine them together to deal with the overall dataset.

I see, I will admit that I’m not familiar with those models but I will definitely take a look at those and see if they might be applicable. Yeah, I guess like in this sense the approach that I’ve proposed here is pretty flexible. I mean, really what makes this approach work well is as long as you have a decent approximation to these probation probabilities we’re thinking about them as propensity scores. So aside from that but because really what’s standing behind this as the generality of the conditional randomization test where you can basically use any test statistic you want. And so, definitely the method is flexible and can incorporate different choices, like the one that you’ve mentioned, But we haven’t tried it we haven’t, we haven’t tried it.

I’m not familiar with this model. Thank you though. Anyone else have any questions about the methodology?

Okay, perhaps I’ll okay. So yes, so this is kind of like a separate thing
0:41:31.25 –> 0:41:32.83 which I will not get
0:41:32.83 –> 0:41:35.59 into details of for the sake of time, but we had
0:41:35.59 –> 0:41:39.06 the separate paper whose focus was just basically,
0:41:39.06 –> 0:41:41.81 the conditional randomization test is a cool test
0:41:41.81 –> 0:41:43.29 but everyone knows it’s slow.
0:41:43.29 –> 0:41:46.16 So how can we essentially accelerate it
0:41:46.16 –> 0:41:49.03 while retaining a lot of its power advantages?
0:41:49.03 –> 0:41:51.41 And so what we found is that
0:41:51.41 –> 0:41:54.26 if you just ever so slightly modified the test statistic
0:41:54.26 –> 0:41:58.716 by sort of regressing Y first on the confounders,
0:41:58.716 –> 0:42:01.78 and then on X, instead
0:42:01.78 –> 0:42:04.03 of regressing it on both at the same time
0:42:04.03 –> 0:42:07.69 what we found is that this ends up being much, much
faster
0:42:07.69 –> 0:42:10.05 because only the second step needs to be repeated
0:42:10.05 –> 0:42:13.69 upon resampling, and the second step is much cheaper.
0:42:13.69 –> 0:42:18.77 So what we did is that we, in the context of sector
0:42:18.77 –> 0:42:21.56 by accelerating the resampling steps even further
0:42:21.56 –> 0:42:25.11 by leveraging the sparsity
0:42:25.11 –> 0:42:28.14 of the CRISPR perturbation vector X.
0:42:28.14 –> 0:42:31.36 And so perhaps the most important part is that the
cost
0:42:31.36 –> 0:42:34.4 of the CRT for one gene-enhancer pair went
0:42:34.4 –> 0:42:37.85 down from 25 minutes down to 20 seconds
0:42:37.85 –> 0:42:40.45 as a result of these computational accelerations.
0:42:40.45 –> 0:42:42.75 And so for reference a single negative binomial
0:42:42.75 –> 0:42:45.04 regression took three seconds.
0:42:45.04 –> 0:42:46.13 So it’s still,
0:42:46.13 –> 0:42:50.2 we’re a factor of six or seven, more expensive than the
0:42:50.2 –> 0:42:52.51 just the sort of vanilla single regression
but it’s definitely, I think sort of within,
definitely within an order of magnitude
and hopefully as you can tell a much better statistically.
So I will show you a few, so this is a simulation.
I’m not gonna go through it in detail, but the idea is
that what we’re demonstrating here is that you can
give Sceptre essentially negative binomial models
that are misspecified in different ways.
You can, give it a dispersion
that’s too large, a dispersion that’s too small
or maybe the true model does have zero inflation
but we’re not accounting for it.
Sceptre essentially is well calibrated, regardless,
whereas if you just essentially took
the like the wrong dispersion estimates at face value
you would encounter problems.
And this SE magic approach
which basically is a permutation approach.
It’s just sort of not doing a great job accounting
for the confounding it, so we see this inflation.
So perhaps more excitingly
I’d like to show you an application to real data.
So I guess this is the, so firstly
we wanna make sure method is actually calibrated.
So if you remember the initial observation was
in a lot of these methods, aren’t calibrated.
So because I’m running a little short on time
let’s kind of maybe ignore this panel here
and focus our attention here.

So this is the Gasperini data that I introduced before.

And so this red line here is actually the QQ plot you saw

on one of my first slides

of all of those negative control gene-enhancer pairs.

It looks different here because

the scale is I've sort of cut off the scale

so we can actually visualize it.

So we see a quite significant departure.

What we actually did is we thought, okay, maybe

ey they have a bad estimate of the dispersion

but maybe we can use some more

like state-of-the-art single cell sort of methods

to improve our estimate of the dispersion.

And so maybe we don’t need to go

to all the effort of doing the resampling.

And so what we found is that

when we use a state-of-the-art dispersion estimate

we still have very substantial miscalibration.

This is, I think, just a Testament

to the fact that it’s just hard to estimate that perimeter

because there’s not all that much data to estimate it.

And then by comparison, we built Sceptre

from the same exact negative binomial model

which is this improved one,

and we found that the negative control P-values

are I think, excellently calibrated.

So this shows you, again, the benefit

of this different way of calibrating your test statistic
and not relying on the parametric model for gene expression.

So this figure just shows a few of the other methods but for the sake of time, I'm going to move on.

This is looking at positive control data. So this basically is like, trying to get a sense of power. And so, again, maybe if we restrict our attention to this left panel here, what we found is that if we just plot the, our P-values versus the P-values, by the way, maybe I should say what is a positive control.

A positive control in this case is a CRISPR perturbation that instead of targeting and enhancer is targeting the transcription start sites of a gene. And so essentially, like we don’t need any extra biology to know that, if you target a transcription start site that’s really going to knock out the gene. And so you can still try to do your association test and see if you’ve picked up those positive control associations.

And so what we find is that actually Sceptre not only is better calibrated but it also tends to have more significant P-values on those positive controls.

So it apparently is boosting both the sensitivity and the specificity of this association tests. Eugene here are the original empirical P-value is this from the negative binomial test.

So after we did the conditional recommendation if you actually have better P-values
for the positive control pairs.

Yes, so you would expect, you would expect it’s like aren’t we just making the P-value is just, like less significant in a way to just help with the calibration.

So how can it be boosting power? But I like the degree of inflation sort of varies like essentially it’s not like, and what we’ll see this I think on the next slide as well, essentially we’re not like, sort of what sector is doing is not like a monotone transformation of things. It kind of there’s not actually just maybe to illustrate it.

I think, this is just an example where essentially what we would have gotten from the sort of the vanilla negative binomial analysis is the area under this dotted or dashed curve here. And so it could have a, like a lighter tail as it has in this case. And so it could sort of either make the P-values on the more significant or less significant.

It’s correcting the miscalibration but not necessarily in a way that’s like conservative. And so this is encouraging. Yeah, that’s a good question though.

- I guess that depends on the confounding you included in the model. So then I would expect it well, re reduce the significance but if you include other co-founding that’s mostly contributing to the noise level probably.
Yeah, sure, so I think I’m right.

Yeah, let me think we are, let me see

I think in this case, we’re correcting

for approximately the same confounders here.

So they already had some confounders

that they were correcting for

in the original negative binomial.

So in that sense, it’s a little bit more

of maybe an apples to apples comparison.

It’s just a question of how do you calibrate

that test statistic that is

trying to correct for the confounders

but I think what you’re getting at

I do think it can go either way.

It’s not obvious that Sceptre would make a P-value

or they’re more or less significant.

I think I will say just as a small detail here

that in addition to the negative binomial regression

this P-value, it says,

there’s this strange word empirical here.

What it means is that

they’ve kind of also applied their fixed that they had

because they realized that they had the miscalibration

and then they kind of like smashed all

of their P-values sort of,

so these are sort of like, so in that sense

it’s not an apples to apples comparison

but what we’re doing is we’re comparing

the P-values that were actually used

for the analysis in this, in this paper.
So maybe that makes it even harder to compare, but yes.

So take this plot with a grain of salt, if you will.

Perhaps I think the most exciting part is applying this to new gene-enhancer pairs where we don’t know necessarily what the answer is. And so this plot just shows you we’re just plotting it’s actually, I guess similar to this plot we saw here except now we’re looking at the candidate enhancers. And so essentially the different colors. So firstly, this also just shows you that this is very much not a monotonic transformation. Like you really can like, if you look into this quadrant this is an example where the original P-value was very not significant, but according to Sceptre it can be very significant and vice versa. So essentially I’ve just kind of highlighted those gene-enhancer pairs that were, found by one method and not the other. And so the upshot is that there’s a total of about, roughly 500 or so found. Well, I guess after found 563 of those 200 were new in the sense that they were not found by the original analysis. And then 107 were found by the original analysis but were not found by us. And we have strong reasons to believe that these could be false positives based on exactly the sorts of miscalibration that I presented. We did look at a few specific new discoveries here
0:51:38.39 → 0:51:43.39 and found that they were corroborated by EQTL data.
0:51:43.4 → 0:51:44.87 And for those of you who are familiar
0:51:44.87 → 0:51:48.41 enhancer RNA correlation data, since I’m running low
0:51:48.41 → 0:51:51.19 on time, I don’t have time to explain this to you
0:51:51.19 → 0:51:52.56 but these are all P-values
0:51:52.56 → 0:51:56.313 of association based on orthogonal functional assets.
0:51:57.8 → 0:52:00.77 Also, we found that our discoveries were more enriched
0:52:00.77 → 0:52:03.6 for biological signals in a few different ways.
0:52:03.6 → 0:52:06.02 One of them is that, and again,
0:52:06.02 → 0:52:08.1 I’m sort of maybe going a little bit
0:52:08.1 → 0:52:10.81 more quickly here ’cause I’m about to run out of time
0:52:10.81 → 0:52:13.07 but there are these things called topologically
0:52:13.07 → 0:52:16.07 associating domains, which are basically regions
0:52:16.07 → 0:52:18.56 in the genome within which most
0:52:18.56 → 0:52:21.68 of these regulatory interactions are thought to occur.
0:52:21.68 → 0:52:24.97 And so what we find is that a greater fraction
0:52:24.97 → 0:52:27.45 of the gene-enhancer pairs we found compared
0:52:27.45 → 0:52:29.71 to the original analysis did lie
0:52:29.71 → 0:52:32.38 in the same topologically associating domain.
0:52:32.38 → 0:52:34.53 So in this case, 74% versus
0:52:34.53 → 0:52:37.04 the 71% found in the original analysis.
0:52:37.04 → 0:52:39.29 So in this sense, I mean, it’s just kind of
0:52:39.29 → 0:52:43.89 like a first order sense of biological plausibility.
0:52:43.89 → 0:52:45.89 I think people are starting to think
0:52:45.89 → 0:52:47.59 that there are interactions that are sort of
0:52:47.59 → 0:52:49.16 outside of tabs as well.
0:52:49.16 → 0:52:51.76 So I don’t think this is a signal that,
0:52:51.76 → 0:52:54.67 26% of these things are false discoveries
0:52:54.67 → 0:52:59.36 but we definitely do expect, a high degree
0:52:59.36 → 0:53:03.633 of enrichment for within tad interactions.
0:53:05.06 → 0:53:06.64 Also if you do look
at some of these more circumstantial pieces of evidence for regulations, such as things like transcription factor binding or histone modifications so we can use CHiP-seq to essentially assess for any given what whether there is these kind of signatures of regulation. And so what we found is that we did a little bit of an enrichment analysis where we looked at all of those enhancers that were found to be paired to genes by sector versus the original method and looked to what extent they were enriched for these other signatures these CHiP-seq based signatures of regulation. And what we found is that across eight of these CHiP-seq targets, and by the way these eight are not selected. These actually were the exact eight CHiP-seq targets that they examined in the original paper, we found greater enrichment. So in this sense, also the enhancers being picked up by Sceptre are just more biologically plausible using these orthogonal kinds of assets. So I find this very exciting and I’m just gonna maybe make a few remarks and hopefully there’s just a little bit of time for questions. I will also be around for a few minutes after the seminar. If anyone wants to stick around and ask me questions you also might have your next thing to go to. So I understand if not.
But maybe the summary is that, mapping gene-enhancer regulatory relationships is very important. If we wanna translate GWAS hits into disease insights. And there’s been this very exciting new technology that allows us to answer that question. This technology was proposed very recently, and so there aren’t that many methods out there to analyze these kinds of data. And so what we did with Sceptre is we leveraged recent methological advances in statistics to overcome the primary limitations of the parametric and non-parametric analysis methods that were available. And finally, we applied it to the largest existing data set of this kind. And what we get is a greater number of more biologically meaningful regulatory relationships. So I had a few other discussion slides, maybe I’ll just read the title to you without getting into the details and this is a rapidly developing technology. And we do foresee that sector will be applicable to future iterations of the technology. So that’s promising. And secondly, this is more like the beginning of the road than the end of the road. There are lots of remaining challenges, this includes looking for interactions among enhancers, things like dealing
0:55:56.16 –> 0:55:59.6 with multiple guidances, how are you in the same enhancer,
0:55:59.6 –> 0:56:02.43 they’re just basically like a whole, I would say, playground
0:56:02.43 –> 0:56:05.603 of statistical problems that have yet to be addressed.
0:56:06.54 –> 0:56:11.29 So maybe finally, if you’d like to learn more
0:56:11.29 –> 0:56:13.39 we have a pre-printed on bio archive.
0:56:13.39 –> 0:56:15.89 I wanna acknowledge my co-authors again.
0:56:15.89 –> 0:56:20.24 And finally, so Tim has worked very well hard
0:56:20.24 –> 0:56:24.03 on putting, making this an art package so
0:56:24.03 –> 0:56:26.11 you can find out on GitHub
0:56:26.11 –> 0:56:28.94 and I’m very happy to take questions now
0:56:28.94 –> 0:56:32.23 but if you have any burning questions that come
0:56:32.23 –> 0:56:34.96 to you 30 minutes after my talk
0:56:34.96 –> 0:56:37.39 please feel free to email me at this address.
0:56:37.39 –> 0:56:40.05 So thank you, and I should have said at the top, thank you
0:56:40.05 –> 0:56:41.873 Lexi for the invitation.
0:56:41.873 –> 0:56:44.99 - Thank you for agreeing to present your work here.
0:56:44.99 –> 0:56:46.66 It’s really a nice talk.
0:56:46.66 –> 0:56:47.51 - Yeah Thank you.
0:56:48.61 –> 0:56:52.37 - So I have some, maybe less related question
0:56:52.37 –> 0:56:55.77 to your current work, but maybe interesting to consider.
0:56:55.77 –> 0:56:57.44 I am not sure.
0:56:57.44 –> 0:56:59.4 Have you looked at the correlation structure
0:56:59.4 –> 0:57:02.5 between the X matrix?
0:57:02.5 –> 0:57:06.57 - Yeah, so essentially my sense is that gets
0:57:06.57 –> 0:57:11.4 like a factor model where you have all
0:57:11.4 –> 0:57:16.12 of these sort of confounders that are inducing correlation
0:57:16.12 –> 0:57:21.12 among all the axis, but essentially like once you account
0:57:21.12 –> 0:57:25.17 for that confounding, it’s independent.
0:57:25.17 –> 0:57:28.043 - I see (indistinct) correlation.
So it’s fairly small correlation and essentially the reason for, and this is very different from genome-wide association studies. So it’s like, Oh is there some analog of Lincoln’s this equilibrium. And the key difference here is that it’s essentially a design experiments. So even though you’re not controlling exactly which cells receive what perturbations you are basically assigning them at random. So if it worked for this sort of pesky measurement mechanism business it would be an unconfounded problem. But essentially, so the only correlations are coming from this measurement. Yes so that is a great question because you can ask, well how did I do the slight of hand run? Like slide three all of a sudden I was working with like one enhancer with like one enhancer and where did all the rest of them go. And I think we’re actually not losing all too much by doing this, especially since we are controlling for those technical factors. Yeah thanks that makes sense to me. And another thing is maybe more, less than less statistical is how many confounding factors they are controlling and what are the important ones that you have identified? Yeah, I mean, so in this case
we’re doing essentially we’re following the lead of the original paper for which confounding factors with control for.

So in addition to sequencing depth. Yeah, so they do have a batch of fact

there’s also something called Percent Might’ve Country.

So it’s like what fraction of all the reads that you got in this particular cell came from mitochondrial DNA as opposed to, regular DNA, maybe a few others like just total number genes expressed in the cell, things of this nature.

So I think here we’re correcting for about five, but you could think of other things like cell cycle, this is a pretty K five is a pretty homogeneous cell line, but especially once you get to other kinds of, tissue samples you might need to think about, cell type and things of this nature.

So I think there are lots to consider here, we used kind of five easy ones.

Okay, thanks.

Any more questions for Eugene? Yeah, I think we are approximating the end of the talk, the seminar.

So thanks again for your great talk.

And if you have any further questions you can just send emails to Eugene offline.

- Yes, yes, definitely don’t hesitate to reach out.