WEBVTT

NOTE duration:"00:31:41.1860000"

NOTE language:en-us

NOTE Confidence: 0.875510573387146

00:00:00.310 --> 00:00:11.570 Uh Indiana in our cans from our Cancer Center. Uh discuss I think topics that are are so critical in terms of being priorities for research initiative and.

NOTE Confidence: 0.929715156555176

00:00:12.260 --> 00:00:44.030 Start first with doctor Mark Mark Gerstein and as you many of you know doctor. Gerstein is the Albert L Williams professor of Biomedical Informatics and professor of molecular biophysics and biochemistry computer science statistics and data science. He's also a member of our genomics genetics and epigenetics research program at the Cancer Center and his research. I think his innovative research has advanced really the field of computational biology.

NOTE Confidence: 0.924071669578552

00:00:44.030 --> 00:01:11.720 Particularly in understanding data science data mining working through large datasets and you know as we continue to understand the reams of data. We get through both cancer genomics and beyond. I think this is leadership innovation expertise in this space. I think is certainly a critical in Cancer Research in our Cancer Center. So we're really very fortunate to have mark present us today.

NOTE Confidence: 0.874556720256805

00:01:15.790 --> 00:01:45.910 So I appreciate the opportunity to do so. I have to stand in front of this thing here, yeah. I appreciate the opportunity to talk to all of you and I want to make more connections with Cancer Center. I thought just to start off. I'd start with a Viking make that picture here. It's just I thought I'd show you picture. This is science. Hill on if you have it. This is where I am. It's been transformed. This is what it looks like in a year ago. so I thought I would start with some pictures of it bit of a bit of construction site.

NOTE Confidence: 0.884549736976624

00:01:46.670 --> 00:02:11.620 So then I just to motivate this all kind of jump right in. I'm going to be talking about passenger mutations. But this is a little bit of a quote from Barack Obama's State of the Union address. While we talked about Precision Medison. I thought was kind of interesting in the address he he didn't talk directly about mutations or genome, sequences, we did mention the genetic code and looking at that in particular.

NOTE Confidence: 0.897539794445038

00:02:12.270 --> 00:02:34.450 Cancer patients and kind of trying to interpret that an of course, looking at the mutations in cancer genomes is really a very important, and also very interesting thing and that stronger focus on today. I just want to give you a little context of what these variants are how to think about them. I mean, obviously there's tons of variants. We have in our genome.

NOTE Confidence: 0.862862944602966

00:02:35.090 --> 00:03:06.780 There's a most of the variance you poop, so sorry there. Most of the variants are obviously there germline variance. These break up into common and rare variants you probably all know this an account so we have semantic variants usually about 1000 but there could be 10,000 in a tumor and often. We think of these group of these variants as being possibly selected and kind of driving along the growth of a tumor and this is the kind of Canonical model.

NOTE Confidence: 0.866094172000885

00:03:06.780 --> 00:03:38.090 That we have where we might think we have, maybe say 1000 variants or semantic variants and it counts June but only maybe 2 to 8 under strong positive selection for the growth of those cells and we often determine these by looking across a cohort and seeing many of the individuals share these driving mutations, so this is the kind of Canonical view and you don't have a reference here to one of the famous papers by Vogelstein on this type of subject.

NOTE Confidence: 0.874995231628418

00:03:38.090 --> 00:04:08.600 And so I'm going to be talking a little bit about this Canonical picture, but not going to really focus on where all those passenger mutations doing do they have any effect and I'm going to be looking at the data set really developed from this big international collaboration. It has a very strange acromas called Peacock. But it's actually the union of the TCG. I seek cancer genomics efforts. An it should have actually it's kind of a big roll out in a couple of months where they'll sort of describe.

NOTE Confidence: 0.866775751113892

00:04:08.600 --> 00:04:14.830 About almost 3000 fully sequenced cancer genomes and these are whole genome sequences, not X homes.

NOTE Confidence: 0.901368856430054

00:04:15.540 --> 00:04:46.090 And they represent a very large amount of data and this of course, many different types of cancers that this group has worked on so I'm going to be mostly focusing on this data corpus. But just for current clarity. I'm also going to look at one particular one in detail. This is a particular type of kidney cancer where there's about 35 whole genome sequence is done just for this particular types give it a little bit more concreteness and it's also well characterized in 2 different types of through the efforts of the TCG.

NOTE Confidence: 0.897950649261475

00:04:46.770 --> 00:05:18.390 So this is an outline for the talk an basically what I'm going to do first is I want to give you a little background to star on it's going to be kind of quick on the annotation of the genome and stuff like that, and also a little background on sort of background mutation processes and so forth and how we have to take those into account and then I really going to focus on the bulk of the talk looking at the past three generations. Their overall functional impact how they're related to the signature or the mutational signature and then how we can kind of use a re purpose. Some of the machine we have from.

NOTE Confidence: 0.867404103279114

00:05:18.390 --> 00:05:25.430 I germline genomics to kind of look at their overall additive effect OK, so that's what we're going to do today?

NOTE Confidence: 0.890763342380524

00:05:26.250 --> 00:05:57.000 Answer first, a little bit about the annotating the genome and you probably all know that genome contains jeans, but jeans are only a very small part of our genome about a percent of the genome and most of the genome actually annotating. It is really different thing than looking at jeans and there's a number of projects that work on this one of the most prominent is the encode project, which is another international effort that aims to annotate the human genome and are very, very high level. I'm just going to say what this type of work does let's see if I can get the little pointer to.

NOTE Confidence: 0.885919451713562

00:05:57.000 --> 00:06:29.680 Is it pointing out that it was pointing one in case? What this actually you know when I can probably do I think I can move the mouse here? I think that's a little bit easier, yeah, So what? This project does it. You know I kind of do a lot of experiments on on the genome and they get some kind of signal of activity that you have transcription binding and so forth and then they process. The signal in some way and in the end, it turns into little regions that you have the microbes in finding sites or places where things are transcribed and then these things are kind of connected in networks and the other thing people do with the genome as we obviously compare.

NOTE Confidence: 0.907492339611053

00:06:29.680 --> 00:06:39.220 The human genome to that mouse or other organisms and also compare it amongst different people to get a sense of conserved regions of the genome and that's kind of an important thing.

NOTE Confidence: 0.885338127613068

00:06:40.200 --> 00:07:10.930 Now, this, the encode project actually spends a lot of time, collecting lots of data set so one of the key issues here is obviously the datasets that are pertinent to many different kinds of dreams have to come from different tissues to epigenetics and the activity of the genome is very different

issues. And so that so you can kind of think of it as a kind of a matrix. We have many different issues and this particular matrix actually made up for cancer genomics where there are kind of as a rough matching between many of the known cancer types and some of the cell lines that they encode crew.

NOTE Confidence: 0.890846192836761

00:07:10.930 --> 00:07:38.390 Tends to work on and then there's many different assays and of course. I'm not going to go into any detail about these assays and so forth thing people do RNA sequencing. Chip sequencing and a lot of advanced as he is like you know sort of the high CSA and so forth and in the end, they get this type of annotation that looks like this, where you have your jeans, but you have other regions may be linked to the jeans and they're kind of linked together and then they kind of maybe put these into these networks type of constructions.

NOTE Confidence: 0.895144522190094

00:07:39.390 --> 00:07:59.820 Now, when you come to assessing the impact of a mutation you might think yourself. Just not very simple level. If the mutation. Obviously disrupts a functional part of the genome. It's maybe has a bigger effect. And so you can kind of in a very simple way you can kind of think about taking the mutations you might have in uh.

NOTE Confidence: 0.853971123695374

00:08:00.640 --> 00:08:34.270 In the cancer genome and saying Oh do they impact into the non coding annotations? Are they in conserve parts of the non coding annotations. We use the word sensitive an ultrasensitive user group for this. These are all jargon words and so forth, then if the Mutation. Maybe has an obvious way it kind of breaks functionally breaks a transcription factor binding motif of a transcription factor. We might score that more highly and if it occupies a central position in some regulatory network is very simple.

NOTE Confidence: 0.809463202953339

00:08:34.270 --> 00:08:38.620 Minded way we might score the functional impact of Mutation.

NOTE Confidence: 0.872088134288788

00:08:39.140 --> 00:08:42.820 OK, now a little bit more background mutational processes.

NOTE Confidence: 0.882980763912201

00:08:43.340 --> 00:09:16.280 So when we look at recurrence in cancer cards. We often might look at something that looks like this will have regions of the genome annotation blocks that could be jeans. That could be transcription. Factor binding says we have groups of people we have mutations and what we were looking for and we're looking for a driving driving mutations is something that kind of looks like this, where there's a lot of mutations that and you kind of

see BYOD. There's more than you might expect right. Maybe in this particular cohort or across a bunch of cars but you got this is a very complicated. Statistical stuff you have to.

NOTE Confidence: 0.862176477909088

00:09:16.280 --> 00:09:35.890 Look at it kind of carefully and one of the big confounding issues is that the overall. The overall mutational processes. They can in cancer is very confounded by many covariates and one of the one of the well known covariance versus replication. Timing early replicating regions of the GM tend to mutate less than late replicated regions of the genome.

NOTE Confidence: 0.905295550823212

00:09:36.580 --> 00:09:44.510 Once you have to take that into account so for instance, you might find this region is actually very overburdened by mutations after you take that into account.

NOTE Confidence: 0.903194606304169

00:09:45.120 --> 00:10:10.700 And this is a picture actually that shows the correlation between say replication timing and the overall mutation rate. It's actually fairly well conserved and actually this picture also let me be a little hard to see but it shows the overall mutation rate for different people within a cohort you see it varies vary. Greatly and so there's a lot of complicated covariates that you have to take into account when you assess the degree of mutations you have in a particular cancer.

NOTE Confidence: 0.896664440631866

00:10:11.750 --> 00:10:44.130 Now we have, I'm not going to go into detail just give you a little flavor. We have a lot of models for doing this type of stuff. It's mystical models and so forth. This particular approach. We have breaks into 2 groups. You have these parametric models. We try to explicitly model, the mutational process and nonparametric models where you kind of shuffled mutations around the GM to get a sense of maybe what kind of random non selected mutational processes and that's very important because if you want to look for a driver gene or something that's positive selected you have to have a sense of what?

NOTE Confidence: 0.874515950679779

00:10:44.130 --> 00:11:16.780 Is not selected which is neutral or random and just to give you a flavor of how this might work. You could imagine a model where you might say the amount of mutations. You accumulate in particular, Genomic been my follow a simple binomial process. That's a little too, simple because you obviously. We talked about the rate of mutations changes of the genome so you might have might allow that rate vary. According to another distribution. This is a beta distribution and you might in particular allow the parameters of that distribution to covary with various Genomic.

NOTE Confidence: 0.867590725421906

00:11:16.780 --> 00:11:48.720 Covariate such as the replication rate or the level of openness to chromatin and then there's different. Permutations schemes are just going to give you a flavor for this type of thing here, so permutation. You might take the genome. You might permute take the variance in the GMM permute them and this is for instance. The Sanger Center uses this simple type of shuffling. One of the problems, though, is that like I said the different regions of the genome or not, that equivalent so you might do something where you make different reasons. the GM kind of equivalent in a way.

NOTE Confidence: 0.871329069137573

00:11:48.720 --> 00:12:03.490 Maybe they have similar levels of similar replication timing similar levels of openness is over then just shuffle. Within these things in this creates another type of amount. This is a type of model that's more favored by the Broad Institute.

NOTE Confidence: 0.897765934467316

00:12:04.210 --> 00:12:30.060 Any case, the different types of shopping malls and this will come up in a little bit as we as we talk and this just gives you a sense if you don't. If you do just a simple binomial model. This would be the distribution number of mutation counts you get and this is the observed real distribution that you actually have and you can see it's very different from what you get with a very simple model. But if you have a more complicated model that has allowed for this varying you can fit that distribution a little bit better.

NOTE Confidence: 0.89906769990921

00:12:30.820 --> 00:12:54.090 OK, so that was a little bit of background on some stuff that we're doing this, you mentation and background mutation process is now we want to talk about this impact of the passenger mutations, so first of all a little conceptual idea here. So there's this dichotomy in the classical dichotomy of a few drivers that are possibly selected OK and the thought is that?

NOTE Confidence: 0.890315115451813

00:12:54.700 --> 00:13:24.950 These have very strong impact OK and then we also have the notion of many other mutations being neutral, but let's just talk conceptually for a second OK? What else could happen for those mutations well. Some of the mutations that you're having cancer GM that thousands and maybe they're actually negatively their negative have a negative impact on cancer growth or cellular growth so they would be under negative selection as opposed to positive selection and we can imagine them being under a strong negative selection or strong.

NOTE Confidence: 0.887614250183105

00:13:24.950 --> 00:13:54.960 Or weak negative selection OK and also we could also have a situation where we have a mutation that actually encourages the

growth of the cancer but this mutation is not doesn't have a very strong effect. It's not one of these Canonical strong drivers and so we might have the notion of a week driver here and now the one of the problems with this notion of weak and strong as they get very confounded with the notion that notion of ascertainment so a lot of times.

NOTE Confidence: 0.909837007522583

00:13:54.960 --> 00:14:09.700 We could have a very strong driver, but it might only occur in a few of the individuals and so we wouldn't see it very well so it might look very weak, so for instance. The appearance of a strong undiscovered driver or a week week driver? That was more common.

NOTE Confidence: 0.918301701545715

00:14:10.400 --> 00:14:36.020 Would would be maybe fairly simple so these are different. Conceptual categories that we could put things that now. We don't know if anything is actually in any of these categories. But what I'm going to do is just look at some of the evidence when we look across these thousands of cancer genomes when we look at the impact of the mutations and we might think where do they fit in some of these categories? Are they only in the neutral passenger and strong driver bottle or do they may be fit into some other groups.

NOTE Confidence: 0.885595619678497

00:14:37.230 --> 00:15:07.340 So, in particular, if I look at the impact the functional impact. This is from a functional impact score of all the mutations. One of the things that's actually very striking when we do this? Is it has a Tri modal as a posed to bimodal distribution, so that now of course, we expect it to be bimodal. You know uh functional strong functional impact for drivers. Lots of other things with weak impact, but actually that's not the case. There's This Middle Group. Here, which it could be passengers, but maybe they have a little bit more functional impact.

NOTE Confidence: 0.889574468135834

00:15:07.340 --> 00:15:30.530 But actually it's kind of interesting to is if we also take a particular type of cancer. So this is a CLL an we are particular group of them and we fractionate the cohort into those that have the stronger impact mutations versus the weaker impact mutations. We actually can see a difference in survivability with actually the things that have the?

NOTE Confidence: 0.88928633928299

00:15:31.160 --> 00:16:03.010 Stronger impact mutations that people tend to survive longer. The implication of this is that those mutations are actually may be negatively select a lot of those mutations are actually discouraging the growth of the cancer now. I'm looking at across all the thousands of passenger mutations, not just a few drivers another thing we can do is we can take particular cohorts of cancer and we can ask as we have more mutations.

NOTE Confidence: 0.904527604579926

00:16:03.010 --> 00:16:34.360 OK, in a particular group is the fraction of impactful mutations does it increase or decrease now if the impact fullness of the mutations if it was completely neutral if it didn't make any difference right we would expect we have more mutations. We have the same amount of impact fullness right OK. But in fact, what we actually observe for some groups resisted for lung cancer. We observe a week negative slope and what is the week negative slope means it means that as we get more?

NOTE Confidence: 0.402841776609421

00:16:34.360 --> 00:16:35.590 You.

NOTE Confidence: 0.907822072505951

00:16:36.300 --> 00:17:00.800 Mutations they tend to be less impactful and overall and the implication of that is that there's a there's a tiny amount of negative selection right because because as we're getting more mutations where disfavoring the occurrence of those some of those mutations that are more negatively selected OK. So now I can grow if I can graph the slope of this line.

NOTE Confidence: 0.891172230243683

00:17:01.370 --> 00:17:32.300 For many different cohorts, you can see many of the cohorts sit in this area here where they have a negative slope a few of them are on the positive stuff but most are negative slope, which is maybe a tiny bit of evidence for some weak negative selection. Another thing we can do is we can look at the sub clonal architecture of the mutation so each of the mutations associated with pull up bath or variant allele frequency and that talks about how early that mutation occur did occur early in the cancer or late in cancer now while we expect.

NOTE Confidence: 0.888074398040771

00:17:32.300 --> 00:17:46.430 We expect that if we looked at for instance, driver mutations, we'd expect if we looked at the ratio of early mutation slate mutations that drivers would be enriched in early mutations. That's what we observe OK.

NOTE Confidence: 0.891191780567169

00:17:47.110 --> 00:18:11.780 But, which kind of interesting is if we look for instance, at just high impact. Mutations just in general. We also find them. Someone rich so maybe the lot and maybe that has to do with the drivers. But if we look at mutations that are not necessarily as impactful. We still see a small enrichment and then if we breakdown amongst the different.

NOTE Confidence: 0.879133760929108

00:18:12.450 --> 00:18:42.980 Groups of jeans that you can have passenger mutations and you can see that tumor suppressor genes have a great enrichment in

early mutations. That sort of makes sense. Whereas oncogenes much less the implication. If you think about. This makes a lot of sense because if you have a random mutation in a tumor suppressor. It's probably going to kill that tumor suppressor gene and that's going to draw the cancer forward right. But Random Mutation Oncogene. Probably not going to make that oncogene function as not gene it probably has to be that exact particular spot.

NOTE Confidence: 0.85677433013916

00:18:42.980 --> 00:18:51.190 To create it, hence you see it kind of sort of go down a little bit evidence for maybe some small amounts of negative selection.

NOTE Confidence: 0.878888666629791

00:18:52.160 --> 00:19:22.340 So then the other thing we can do is we can take a measure of functional impact. This is that gurp word. But group is a word people use in genomics for conservation. So this is just as the as the site of Mutation becomes more concerned. That means more functionally important right. We can ask ourselves does the mutation tend to occur earlier in the cancer that means it's a stronger functional impact earlier in cancer so we take our driver variants we get a very nice.

NOTE Confidence: 0.883976697921753

00:19:22.340 --> 00:19:52.970 Straight line so the driver variance or there's a very clear correlation right. If we take unknown. Varient butter in driver jeans jeans with their drivers. We also get a week upward slope again. This is the signatures of positive selection right. But what's interesting is if I take all of the other mutations and cancer. The thousands of other mutations and I graph them. I get a tiny negative slope. The implication is as the site of Mutation becomes more conserved.

NOTE Confidence: 0.897813498973846

00:19:52.970 --> 00:20:10.940 More functionally important, it becomes less likely to be in early on occuring mutation more likely to be a late on occur. Mutation implication slightly deleterious OK, so another type of sort of maybe hint of a small amount of negative selection.

NOTE Confidence: 0.903076410293579

00:20:11.750 --> 00:20:43.610 So let me tell you a little bit about mutational processes and I tried to a little bit about the stuff earlier, so this is a little technical but we can think of the mutations in the cancer genome's coming coming about from different mutational processes different mutational signatures and the signatures are usually shown like this and this is a little technical where you have. This spectrum and I'm zooming in here each spot in the spectrum represent the number of mutations originate Ng from particular trinucleotide, so for instance, this is the number of mutations.

NOTE Confidence: 0.864088714122772

00:20:43.610 --> 00:21:13.680 That come from a CG that go to T in the middle, whereas this one is has a different try nucleotide structure, OK and each of these try new each of these mutational. Spectra is associated with different mutational process. For instance, one so she was smoking. This one's social with aging want to share with sun exposure right now, what we can do is we can look at the an overall cancer and I'm going to take kidney cancer make this more concrete here, we can look at its mutational Spectra.

NOTE Confidence: 0.889801204204559

00:21:13.700 --> 00:21:44.330 We have lots of peaks in these CDT transitions. This well known object here and this is its overall spectrum. But then we can ask what are the Spectra of mutations that have a very strong impact or the mutational Spectra of mutations attend to disrupt particular transcription factor binding site so these are particularly this is the transcription factor binding sites for say SP one or EWSR and you can see the spectrum. Mutations is very different that has the strong functional impact than the overall.

NOTE Confidence: 0.824981987476349

00:21:44.330 --> 00:21:46.770 Distribution in the canister.

NOTE Confidence: 0.89542555809021

00:21:47.690 --> 00:22:18.560 And So what the implication of that is, is if I say to myself, I take the kidney cancer. and I asked Oh. Let's ask. Let's look at the low impact mutations. Those mutations are dominated by signature one mutational process. One second, Secondly, mutational process 5, whereas if I look at the higher impact mutations. Those are the mutations. The medium impact mutations. Those mutations are dominated by signature 5.

NOTE Confidence: 0.879218816757202

00:22:18.560 --> 00:22:24.520 Different mutational signature then signature one so the idea is just shifting the mutational process.

NOTE Confidence: 0.89551317691803

00:22:25.530 --> 00:22:31.580 In crime inexorably changes the functional impact of the mutations whether or not, their selection or not.

NOTE Confidence: 0.885712683200836

00:22:32.770 --> 00:23:03.320 And this is a kind of global picture of that where you see all the different cancer. Cohorts here OK and these many different transcription factors and you can see each of the different cohorts has different mutational processes going on it, and so you can see how the amount of that particular the transcription factor binding sites is differentially impacted in each of the cohorts probably to do with slightly different mutational processes.

NOTE Confidence: 0.902204394340515

00:23:03.320 --> 00:23:33.840 And this picture I'm going to skip this one is a little more complicated talks about how there's a change between high and low pack low impact. Mutations in the different colours, OK, so now. I just quickly want to go into the end here, so we might say I've been kind of hinting that the passenger mutations are not passengers. There even though they're in the back seat. There actually directing the car right and I've been hinting that there might be weak positive and negative selection involved in these pastors and this is a little counter to the receive dogma.

NOTE Confidence: 0.902432084083557

00:23:33.840 --> 00:24:03.950 OK, so how can we kind of mathematically show this well one way of mathematically looking this is see if we could predict if a particular sample is cancerous versus not by taking into account. These mutations again. So what we did here is we reposition a particular type of model that's very commonly used in germ line genetics and that's called this random effects are out of effects model and so you might know for instance, that people are very successful in predicting.

NOTE Confidence: 0.883648037910461

00:24:03.950 --> 00:24:36.040 In relating the genetic contributions to height, but there's many, many variants thousands of various related to height and only a few are related very strongly tight. But if you sum all the various together in one of these models. You can actually see it accounts for height very well and that's this model here when you think about that rate here is the particular the particular variance. And here's some coefficient on the variant now. I'm going to just go through this kind of quickly what we do here is we make a sort of similar type of model where we try to predict that rate being.

NOTE Confidence: 0.868904769420624

00:24:36.040 --> 00:24:44.630 Essentially growth or cancerous versus not we related to this semantic mutations OK and we refine these.

NOTE Confidence: 0.917911529541016

00:24:45.290 --> 00:25:00.110 Coefficients on each Arm Mutation and there's a little bit of mathematical discussion of how this random effects model doesn't actually refine a particular parameter. It just puts a random number that you don't know what that number is and I'll explain to you, a little bit more about that in a second.

NOTE Confidence: 0.899029314517975

00:25:00.720 --> 00:25:20.260 And then we can actually put we can divide our model into different groups for say mutations in promoters or mutations in jeans or mutations in different groups to compare the relative importance of these different groups and this is the kind of result, we get when we do this so here.

NOTE Confidence: 0.89752858877182

00:25:20.960 --> 00:25:52.420 This is the amount of variation explained OK, so if we can explain all the variation we can completely account for that right. So so that's the sort of language when we talk about how much of the variation height. We can explain in terms of next week and say it's very hot so we want to do here is we want to explain can. We predict if something is cancerous by looking at all. The variants so if we just take the driver. The driver, the known values. The drone driver variance we can explain about 50% of the variation.

NOTE Confidence: 0.880371570587158

00:25:52.420 --> 00:26:23.690 In these cohorts you can see it varies amongst the different cores, but let's just say with a number by 50%, OK now. If I take all the valiance now if I go from taking say the 8. Key variance to take the thousand or so in a cancer. I can explain 9% more than variation no that's not huge, but it's definitely appreciable so that means that those thousands of other variants in the cancers have predicted value to saying that person has that particular type of cancer.

NOTE Confidence: 0.523160696029663

00:26:23.690 --> 00:26:24.280 OK.

NOTE Confidence: 0.893539071083069

00:26:25.880 --> 00:26:55.950 Then I can do more. I can say, Well, let's split up this discussion of variance between coding coding and promoters. Encoding pros and other non coding regions and you can see if I do this that I get a slight increase again. When I add in the non coding regions. The passenger mutations in a non coding regions. Small amount of added variants now of course I figure out the amount of additive variance.

NOTE Confidence: 0.879898548126221

00:26:55.950 --> 00:27:18.580 Per nucleotide because they're non countries are so big it. Obviously falls very little and obviously the dominant component of explaining the variation is the mutations in the coding. Regions certainly per nucleotide. That's unquestionably true but there is appreciable amount of additive variance in the non coding regions.

NOTE Confidence: 0.892066776752472

00:27:19.170 --> 00:27:44.890 OK and then finally this is a little bit of an abstracted slide. But I can just say that what we can also do here is as I was saying. This is a random effects model. We're not actually saying which particular mutations are actually contributing to the mall. We can actually recast this model into a more productive contact context, finding the best linear unbiased predictor.

NOTE Confidence: 0.906381189823151

00:27:45.600 --> 00:28:02.230 And in that context, we can find the our estimate of the mutations that are contributing most to that additional additive variance and therefore we can find small group of week additional week driver mutations and when we do this we basically find that there's about.

NOTE Confidence: 0.884905278682709

00:28:03.110 --> 00:28:11.650 7 or 8 more week drivers in each of the pan cancer cohorts. In addition to the known drivers OK.

NOTE Confidence: 0.841547966003418

00:28:12.230 --> 00:28:14.910 OK so let me summarize where I talked about.

NOTE Confidence: 0.895953953266144

00:28:16.040 --> 00:28:46.390 So today, I told you about a lot of some topics in Kent in cancer genomics in particular, looking at the effect of these passenger mutations first little background on the annotation background mutational processes and so forth and then I looked at the overall impact of the passengers key idea. There's a kind of trichotomy of impact as opposed to dichotomy just a lot of things don't scale the way we might if they were completely neutral. If you have more mutations. There actually their impact decreases on average.

NOTE Confidence: 0.889369428157806

00:28:46.390 --> 00:29:19.060 They they if you look at the early versus late. Mutations is not quite what you expect for mutations monitor selection, then we looked at the impact of signatures and we can just clearly see that signatures naturally give rise to different types of impactful mutations, so signature change and the different signatures in the different cancers could be giving rise to different types of impactful mutations and finally. We talked about trying to bundle this into kind of a predictive framework using this Adam affect small and I think here what we can show quite precisely.

NOTE Confidence: 0.884392321109772

00:29:19.060 --> 00:29:28.960 There's a small additional amount of outer variance explained if we include them any passenger variants into the model, particularly the noncoding passenger of variance.

NOTE Confidence: 0.907137393951416

00:29:29.730 --> 00:29:39.490 And we can actually use this to develop an estimate for the number of weak drivers in a number of the cancers and.

NOTE Confidence: 0.849660575389862

00:29:40.400 --> 00:30:11.250 That I'm going to end as we have more construction pictures to see this is chewing on the building there in science. Hill and I just acknowledge some of the people that worked on it, so there's a lot of people that work on the annotation effort related to cancer genomes. This has been work that a lot of people worked on it, alot, dot, dot dots here, but the

key. People just in my in my individual apps. I've been haijing's on this also in collaboration with Shirley Lou at Harvard.

NOTE Confidence: 0.835774481296539

00:30:11.290 --> 00:30:42.480 And Kevin White at the University of Chicago and then the sort of Lancaster thing. There's hundreds of people involved in this. But the particular group that worked on the stuff in particular here is Sushant Kumar is Associate Research Science in my lab and Jonathan Worrall. There's also been collaborating with Gabby. Get and yoga. Peterson and kidney cancer stuff that we highlighted was done by a graduate student my lab. Chantele also with Brian Shock, who is was here awhile ago, an without I think.

NOTE Confidence: 0.904190838336945

00:30:42.480 --> 00:30:44.840 And and thank you for your attention.

NOTE Confidence: 0.388367474079132

00:30:50.320 --> 00:30:51.200 Start.

NOTE Confidence: 0.820196509361267

00:30:51.700 --> 00:30:55.160 Thank you more questions.

NOTE Confidence: 0.863053143024445

00:30:57.680 --> 00:31:28.790 So let me ask you know, given what is work deciphering the passenger events beyond what you described is sort of the context right that the interactions between these events? How do you model that in well that's it like I said so that word people tend to use in German genomics is epistatic affection of sort of the correlation between different things. So we're not taking that into account in this at all, and I think that's another level of analysis. Of course, it makes the analysis more complicated to look at pairs.

NOTE Confidence: 0.873571813106537

00:31:28.790 --> 00:31:40.010 Of variance or I guess triples and so forth, but definitely that something that people want to do in the future. Of course requires more and more statistics bigger and bigger data set so forth.