**Dr. Gabriel Alves** 00:10
Welcome to Speaking of Mol Bio, a new podcast series about molecular biology and its trending application in life sciences. I'm Dr. Gabriel Alves,

**Steve Lewis** 00:20
And I'm Steve Lewis.

**Dr. Gabriel Alves** 00:22
Throughout our first seven episodes of Speaking of Mol Bio, we have explored CRISPR cell engineering, multiomics, exosomes, and single cell analysis. And for today's episode, we are returning to the world of multiomics. Today, it's Dr. Ben Sun's turn to share about his exciting work in this field.

**Steve Lewis** 00:44
Ben is the Head of Biomarker Genetics at Biogen, where he uses multiomics for biomarker discovery and disease prediction. In addition to his industry research, Ben also holds an MD and leverages his clinical experience to advance his work. We hope you enjoy our conversation.

**Dr. Ben Sun** 01:02
I guess I've got a quite unusual background in terms of how I ended up in the Boston area. So I initially trained in clinical medicine at the University of Cambridge in the United Kingdom, across the pond, where I did the integrated MD PhD program, where during my PhD, I specialized in statistical genetics and genetic epidemiology. And after that I practiced as a clinician, or I think you call it physicians, and my sort of research interests during my PhD time really brought me back into the research area. And the field we work in is really, where industry is doing a lot of really interesting work. And then that's what brought me into sort of Biogen, where there's a lot of excitement around using large scale genomic data. And so I've relocated from the UK to the Boston area, and also Cambridge, but Massachusetts rather than Cambridge, UK. Currently, I am within the human genetics group at Biogen, where we use utilize large scale population data, and genomic data to help us answer a lot of understanding in terms of disease pathology, and also contributing to development.

**Steve Lewis** 02:39
It's great to meet you, Ben. Thank you for being here and taking the time with us. I'd love to hear how you went from clinical medicine, to working even more within programming and what aspects of it are really interesting to you.

**Dr. Ben Sun** 02:55
There are both interesting aspects that drives what I do and what I want, when I might, where my passion lies. And also, there's also quite a, an interesting change in directions. But what I sort of ended there, sort of through starting at my PhD has always been interested in, you know, applying big data, which was popular at the time now is sort of diverged into machine learning and AI. Where I picked up a lot of the statistical background and mathematical background for applied and applying it in terms of large-scale data, which includes genomics and also large electronic health records. And I think, coming from the clinical background gives me quite a nice way to interpret the data. I think that's where that

hybrid cross-domain translation is really what drives me passionately, in terms of being able to do this. And not many people are able to do that, because they tend to stay in one the clinical domain or stay in the methods or statistical domain alone. But I think it really opens the, my eyes to like a bigger picture, being able to bridge the gap between the two.

**Dr. Gabriel Alves** 04:13
You mentioned about AI? I'm curious to hear your opinion about that. As we move forward, AI is becoming increasingly more popular in various fields, including molecular biology. And what's your opinion in the use of AI multiomics. And what do you think there'll be the impact of AI in the future of this field?

**Dr. Ben Sun** 04:33
It is becoming a hot topic over the last few years, especially with the emergence of large scale, computation, and tech. The popularity is gathered a lot of interest and also a lot of academic research, as well as industry research, really bolstered the field. The advancement has always been on the sort of methods side but also the computation as well. The computation ability has really enabled the training where, you know, if you take it back 20 years ago, there isn't that computing power to even do what we do now, despite the methods being there. But now with the emergence of cloud computing, also, we're getting more and more cores, our laptops and now more processing power. We're getting more and more computing power really to accelerate that process to really see that to the in theory, we get this but in practice, we apply this to a large set of data, we can also get valuable, meaningful results. It's transformative in a lot of fields so far, for example, in self driving, image processing, and now language recently as well. But there's also a lot of limitations, a lot of limitation really comes down to the amount and quality with data that we come in. I mean with healthcare has been notoriously difficult, where you know, diagnosis and patient management. The data itself is noisy. So obviously you know, the famous saying of garbage, in garbage out.  If you have bad data, or if you have noisy data, if you have data not well curated, then the model is not going to perform well. And that's why we see this struggle to for machine learning or AI to really impact the medical sciences in as much as it has done for things like imaging. But obviously the advances that's brought in definitely has moved us closer to in terms of where we want to be.

**Dr. Gabriel Alves** 06:26
In this first season of our podcast, we were, you know fortunate to have guests from various areas and trending applications, molecular biology, especially in exosome, or extracellular vesicle, research and all its potential in diagnostics and treatment. How familiar are you with exosome research? And what do you think about the use of omics in it?

**Dr. Ben Sun** 06:50
The exosomes are something that's is come up to the stage in recent years, it comes at a good time where omics research is becoming more and more mainstream. Especially with, you know, single cell sequencing, exosomes also previously, the microbiome. So you're getting all sorts of omics, you know, from the metabolomics to the proteomics. And now, you know, exosomes, and obviously, that spans both the technology that's been measuring the omics, you're getting more and more proteins being covered at faster, growing at a faster rate than ever. Same with metabolites, where previously, you're

lucky if you can do a dozen or so on the same samples. But now you can do hundreds or thousands of measurements on the same sample at the same time, through multiplexing. And through advances in mass spec and NMR approaches in metabolomics space. Despite the hype, I feel there's a lot of groundwork still to be done in terms of being able to gather and collate all that data. Initially, before we get to, you know, see the utility of different compartments and different tissue compartments. Because it depends on the disease area that you work with, and depends on the biology you're interested in. Different department, tissue departments could contribute differently to the effect that you see in, for example, you know, intuitively hematologic and metabolic diseases probably work really well in plasma and also in, you know, the exosome space. Whilst you know, other diseases, if you're going into the that might be sort of brain-specific or eye-specific, or other tissue specific areas where you need to get to that tissue in order to see any interesting biologically and clinically meaningful associations and effects.

**Steve Lewis** 08:47
Really interesting. Your comment earlier about just, in general, computation power. I think that it's enabling amazing work in biology. And I think in particular, like you said, the blood as a target. I think there's a lot of really interesting biomarkers that can be derived. So it's incredible what you do, I think it's mind blowing to kind of think about, like, all of these things actually happening in real time in the body. And we're just getting snippets and time as we gather more and more data. It's fascinating.

**Dr. Ben Sun** 09:29
Yeah, I mean, the most fascinating thing I find is how thing quickly things evolve. In the 90s, people couldn't do it at scale, though. It just wasn't the technology's not there. But now you can do things that thousands, tens-of-thousands, and also in the genetic space tens-of-millions. That's where it's extraordinary and we don't know in next five years how that's going to develop because it during my years of my PhD, there's always this dilemma. Do you wait a while to have something better that, you know, as you know, do you wait for the next iPhone 10 that's going to come up in next in two years' time? Or do you commit now commit the money now and buy, purchase the product, or purchase the iPhone or get the sequencing done now? So it's also an interesting question to dwell on from resource planning as well.

**Dr. Gabriel Alves** 10:22
Where do you see the field going the next five years?

**Dr. Ben Sun** 10:25
It's already happening is that, you know, this sample size is going to get a lot larger. Now we're working with biobanks, that the sizes, you know, half a million, or hundreds-of-thousands. And now, since UK Biobank, there's so many biobanks are emerging all over the place across the world. So the scale of data is going to increase by magnitude, at least, the amount of ancestry is going to get a lot wider. And the amount of information we're collecting on each sample, each person is also a lot wider with, you know, digital technology. Now we're able to do, you know, things remotely and digitally, to collect a lot of information, but also the other omics I was talking about, you know, that proteomics, metabolomics, and also transcriptomics, those data or you can accumulate many more of those data as well. I think that the bottleneck at the moment is not necessarily the raw computational power, is that is how we

really make it the most efficient way to analyze something, despite you know, the increasing computation, that cost comes from, you know, there's a cost to that there's a cost to getting things onto the cloud, and the financial cost is not trivial to run things. But also the carbon footprint is also nontrivial as well. You know, the amount of trees you're burning to, you know, to get to train a model. It's, you know, it's something that's underestimated and also under appreciated. So I think there's a lot of advances to be made there in the coming years.

**Dr. Gabriel Alves** 12:01
Interesting, and it's still sticking to the large scale multiomics data. Could you talk to us a little bit about the challenges and limitations that you currently face while working with these large scale data?

**Dr. Ben Sun** 12:16
So that I mean, there's various sorts of challenges. So, the challenges could come from the technical aspects, but also from, you know, just biological aspects as well. And then there's also what's the best way to analyze something. And from the methods point of view, there's so many different ways to do things. How do you go about choosing the optimal way that you want to analyze things? I think that a lot of the heart, sort of bottlenecks comes from, you know, variation. And the variation can come from both technical aspects where you measure the same thing, twice, you get slightly different answer, you know, is, you know. Even if you measure someone's height, you get a slight variation. But that variation is so small, it doesn't really impact, you're gonna get a very accurate estimate of someone's height and weight through measuring it once. But sometimes in biology there's, or if you measure someone's blood component for things, you know, whether it's, you know, blood cell counts, biochemistry, proteomics, metabolomics, whatever. You name it. There's a lot of variability, if you are measuring the same thing twice, that comes from the technical limitations of the assay you're using, for example. And then there's also the biological variability within a person. You know, and that biological variability could just be through sampling, you sample that, you know, through different sites on the body, like the two different parts of the blood or blood from two different places, you might get different things or different bottles when you aliquot it, you know, from the tube that you collect when you obtain someone's blood. And that introduces a biological variability, as well, you know, person to person variability, you know. And then you also have a time component that's currently very, very sort of underappreciated, which is a big challenge, because we know for sure, you know, their circadian rhythms, their biological clocks where someone's protein could vary over time. It could vary during the day and during the year and during the month, especially in female reproductive biology, there's a monthly cycle, as well. And at the moment, because of various feasibility and cost limitations, we only taking a one-time point for a person. So you can see how that is not representative of that person's state across that day, maybe not even across that weekend, not across a month or not across that year. And that variability obviously is not captured either. And that introduces additional sort of uncertainty in terms of the value you obtained because that value is subject to change and be able to account for that is obviously a very difficult thing because there is no ground truth for it right now. And then the third sort of aspects of the challenges really comes from when you do things at scale. When you do a small experiment, there's less chance for, you know, batch effects to occur and batch effects to implement, because you're running things on one plate, you only doing one plates worth of experiment. But if you want to do things at scale, such as in the biobank, you have, you know, hundreds of plates that are done, hundreds-of-thousands of pipetting, that are done, although ultimate automated, but you know, there's a lot of procedures and lot of steps,

than you do in a single experiment. And then you can see, even, you know, in that pipeline, there's a lot of points where things can go slightly wrong, or things that, you know, that could fail QC or some steps that might not go, according to what you planned. And some of it is unexpected as well. You know, you despite if you put something and put it to scale, it might not always work end to end. I mean, for large amount of things, you know, when you go to meetings and conferences, there's always bound to be some tech aspect that goes wrong as well. You know, despite in theory, like everything should work, there's no nothing broken that we put into production. And when, you know, tried to do it for real things doesn't, there's always something that can go wrong in the process. And we're doing it at scale, I think these are becoming, these are no longer sort of trivial and negligible risks that you're dealing with. And be able to, firstly, find these issues, you know, batch effects, plate effects, for example, and differentiate those issues from true biological variation, which also confounds what you're seeing is a very difficult task, because there is no gold standard. And there's no sort of one way to do that. And then lastly comes from different methods being developed that will do similar things and be able to optimize and choose the best method for the data at hand and for the research question you're trying to answer. It's becoming also more and more difficult, because you know, even with neural networks, and with, you know, machine learning models, the different parameters and hyper parameters and different layouts and different neural network layouts that you can do is all are sometimes a little bit subjective. So it's difficult to home in on one specific setup that works for everything.

**Dr. Gabriel Alves** 17:37
We hope you're enjoying this episode of Speaking of Mol  Bio. We wanted to take a quick moment to tell you about the Invitrogen School of Molecular Biology. It is a great educational hub for molecular biology with rich and reliable technical content designed for new and experienced molecular biologists alike. Check it out today at thermofisher.com/ISMB. And now, back to our conversation.

**Steve Lewis** 18:04
That kind of leads me to one more question. Clinical medicine has, especially from like a drug design and drug development standpoint, has always kind of leaned on, I guess, working with different models, and then ultimately moving into clinical trials. And everything we've talked about so far in the omics space has really been kind of around the in-silico aspect of medicine. So, I'm curious as time goes on, how much of do you see for planning for drug development is going to shift more into that digital space? And then maybe, perhaps, move away from some of the either in vitro or even in vivo experiments that are that that have historically been done to date?

**Dr. Ben Sun** 19:00
Yeah, I mean, that's an interesting question, actually. Because I think with the emergence of digital data, and large-scale digital data, especially in humans, but also the current sort of interest in sort of digital data and machine learning, and large scale analysis, I think more and more places are becoming more aware and want to leverage the this resource. But however, I think, you know, in vitro and in vivo experiment has come a long way. It's been, you know, clearly there aspects of it that has worked well and has been, you know, tried, and tested. So there's bound to be, you know, is that it's not going to go away. It's not, I think is more how we integrate the two together, rather than trying to get one to replace the other. I think being able to merge the evidence in a non-bias way, is where the challenge lies. But now I think, you know, even in vitro in vivo space, there's a lot of advances being made as well, for

example, you know, CRISPR technology is something that's occurred in this decade, essentially. And that's, you know, really revolutionize the in vitro and as well as in vivo space. So I don't think is, one would go out fashion in favor of the other. But I do feel it really is also contextualized, on the disease at hand, because in vivo, and in vitro experiments, you want to do something that's representative of a human disease pathophysiology, or a disease mechanism or disease model, as some diseases are harder to model than others. Especially the diseases where there tend to be more bespoke to humans, there is no, you know, mouse equivalent, just purely because, you know, the way humans are distinct, both in terms of their environment that, you know, the social interactions, as well as, you know, things that are less likely to generalize across other animal species. So, I think a lot of those diseases would definitely be harder compared to something that's more fundamental that might be translatable across different species, where, you know, you have models for cardiovascular disease, and potentially some, you know, movement disorders, and also for, you know, respiratory diseases, all sorts of diseases. But there are models that work well, and then as models that work less well. So I think the ones the in vitro models, where, you know, there's been tried and tested models that reflect human biology very well, are there to complement what we do. But it's really how we weigh the evidence, that's going to be the challenge. Do you, for example, if the big data in epidemiology or human data, gives you one answer, and then in vitro, in vivo experiments gave you a different answer. How do you combine those two together? How do you weigh that up? Do you believe in one or the other? I think that's something we don't have enough data for and will not be, we won't be able to accumulate that data at enough pace and scale to answer that question empirically. So I think that's where the challenge will be. So I think the jury's still out on the how we weight, how we put different weights to those different elements, not necessarily in clinical trials, but also in all sorts of in scientific field in general, I think.

**Dr. Gabriel Alves** 22:52
Great. And Steve's question was in regards drug development, drug development, mine will be about diagnostics and precision medicine. Going back to precision medicine, how do you see the field of multiomics impact the future of personalized medicine and diagnostics?

**Dr. Ben Sun** 23:11
Yeah, um, so I personally feel, you know, if you look at, you know, the, the amount of evidence is accumulating over the last few years, there's definitely additional values, additional impact in integrating you know, omics, right, in addition to the standard demographic and standard biochemistry, or biomarkers, that are used clinically right now. Because if you supplement, you know, these experimental omics approaches, if you supplement that data into the existing sort of gold standard, if you like to call it, you know, predictors for various diseases, you see an improvement. There's no doubt that you see a marginal improvement and real improvement might not be as drastic as people hope it to be. But there's definitely an improvement there. So and that's, you know, generalize, there's been cases has been generalized across different cohorts. So, the value that additional predictive value is there to stay. And that's obviously intuitive, because you measure on a lot more things, of course, you're going to see again. I think the difficulty comes in in terms of being able to pick which of those proteins or which of those metabolites that you should consistently measure in the clinical use domain, right? Because in research, the threshold might be different to what you use clinically, because your margin of error that you're able to tolerate is different. And also depends on how severe the disease is and what's the impact of predicting someone with high risk of disease in terms of the downstream

management, downstream prevention, and downstream clinical management. And obviously, we in the clinical space you want to be sure, right? Whilst in the research space, you might have a slightly more tolerance for error, because there's less of a direct impact on human health on someone. So that difference in threshold really makes it difficult to convert what you find in the research space into what you have, into what you use clinically, because that this tolerance for risk is different. Because if I told you, “Alright listen, 85% or 90% chance of you get having increased risk of getting certain disease and I'm 90% or 80% sure of it.” In the research phase, or if something that doesn't really impact you clinically, then that's something you might go, “Cool, I think that's useful to know.” Right? To a person. But if if I told you, you know, “I'm 90% sure, 80% sure clinically, in a clinical domain, that that means, you know, there's 15% of error that I'm not sure about.” And if I told you, “You're high risk, I'm gonna have to intervene on these and obviously, these interventions come up with, that come with their own risk,” then you can see how I'm suddenly not so confident about implementing that anymore. One is for information that I would like to know, and statistically, I'll be more right than wrong, likely to be right than wrong. But in the clinical space, I want to be 99% sure, rather than 90% sure, that I've got this, you know, increased risk or raced biomarker, for example, for a disease. I think that's where the impact from, translation from research, epidemiological research or basic science research into clinical areas has been the difficulty. But because the threshold is different, because, you know, we have to limit the amount of harm we do to patients.  I mean, the first rule of you know, clinical medicine is do no harm. So whilst that in the research base, especially, come from, from a computational background, if you're more interested in expectations, you know, on average, as long as I'm doing better than chance, or I'm doing better than what's currently available, that's good. Because in, in medicine, you have to bring in the additional weighting in terms of the harm that could potentially be caused. And some of that is actually quite difficult to quantify as well, objectively.

**Dr. Gabriel Alves** 27:32
Right, and, but in any other side, you will have the chance of increasing prevention rates, like, oh, you are on the path to developing, you know, hypertension, you know, you should take these measurements now. And so, don't you think that can also, on the other hand, help on the preventive side?

**Dr. Ben Sun** 27:51
Absolutely, I think there's still a lot of trials or, you know, to be done in terms of prevention, to see the utility and efficacy of that, you know, in a non-biased way. And, obviously, you know, you have screening programs for various diseases. But some of the screening program, it, you know, is in the research space works, but when you employ clinically, the efficacy is, the quality of life, you know, that's improved over time, is not as different, as is not drastically different, it's not enough to make an impact, to justify the intervention that you're making. Because whatever you intervention you made, has its own risks as well, that you're introducing to a patient. So that I think that the paradox is that there is value. Of course, in doing this, if you implemented, you know, if you integrated multiomics approaches into disease prediction, or into various sort of prevention or prediction algorithms, you're gonna get, again, on average, most of the time. But there'll be people where that doesn't work, but that's incorrect. And that small population would suffer, and you can't afford that, in the clinical space. I think there's a thing there's a should be a lot of studies that should be done to evaluate these scores into, you know, multiple cohorts and multiple population to demonstrate there is a, you know, a clear improvement over

what you have right now. That I think it would convince a lot of the, you know, the practitioners to adopt, you know, the, these multiomic, additional multiomics data into this. But obviously, there's another aspect is the cost, the health burden, and the cost element that's on the health service in order to do these because, you know, these tests, if you're going to enroll it clinically from a health planning perspective, is not necessarily going to be equitable and it's not going to necessarily going to be affordable because they're not exactly cheap tests. And then you're run into the issue of, you know, equitable and equality and how you, you know, divide the resources up, so that everyone's getting fair and equal access to this, so that there's a scientific element. But there's also, you know, a resource element involved as well. I think definitely over, you know, in the next 10 years, you're gonna see more and more biomarkers been integrated into risk or models. I think I firmly believe there's value, genuinely clinical value in doing that. But, you know, there's a lot of moving pieces, there still a lot of, you know, different parties and point of views and needs to be considered before you're making it into part of a clinical guideline, for example.

**Steve Lewis** 30:53
Really interesting, and a great question, Gabriel. That that almost kind of, to me, begs the question for where cell and gene therapy is headed, right? Because we're at this really interesting inflection point where we can now go inside the cell, right, and with our drugs, that we design, right, or biologics, and that for a long time was something that was taboo. It was something that you didn't do because you, you know, focused on extra cellular membrane receptors, for example. I'm curious, how do you reconcile that? How do you reconcile the consideration that we do now have personalized medicine coming out that's available? And then the idea that there are those cost constraints and equitability issues to consider as well. There's a lot that goes into that. I'm just curious, your opinion of what it's going to look like over the next 10 years as it does become more personalized.

**Dr. Ben Sun** 32:03
Yeah, I think that's a really, really tough question. That depends who you ask, you might be getting, you know, different answers, depending on what hat you put on. If you I mean, from a pure science perspective, or, you know, there's benefits clearly in doing that. But then the, but the ethical dilemma really revolves around, you know, how, how you best distribute there. Because not everyone's going to be able to benefit from that. And also different countries, you know, different points as well, in terms of how equitable things are. And I mean, if you look at, you know, the COVID vaccine, for example. You know, the, the distribution of vaccines is probably from scientists with scientific perspective, if you treat everyone equally, probably hasn't been optimized. But, you know, there's so many forces in play here that what you see is, it's something that's very complicated. It's, it's difficult to put up, you know, the way to sort of best, to best distribute and answer that question. Quite frankly, I think, personally coming, you know, from a science background, I like to see the field evolving, I like to see, you know, to see whether things work or not, and how you distribute that technology is probably outside of scientists' area of expertise is probably up to the resource planners. But I think from a scientist perspective, you're interested in whether something brings value to the field to life sciences, to medicine, to people's lives. Does it improve? I think it's similar, if you go back in history, you know, with advances dependent penicillin scientists, you know, care about this penicillin work, or does you know, a drug or does something work? Whether how much you sell the drug for, you know, what, or how you distribute the drug and how you produce it and how you distribute, who should get it. I think it's something as pretty

out of the limits of what we do in science. Yeah, but it's a very important thing to consider. But I think the I think there's no easy answer to that.

**Steve Lewis** 34:31
Certainly. I figured I'd take the opportunity since you were both a clinician and a scientist to get some of your perspective there. So yeah, thank you. I know that's, that's a really long question maybe better had at a bar over many beverages and ways to solve the world crises.

**Dr. Gabriel Alves** 34:56
Yes, that sounds good, Steve. Ben, I ask this question to every single guest that comes in this podcast, which is their opinion in what is the most important ingredient for their success in their careers? So would you mind telling us what is the most important one for you?

**Dr. Ben Sun** 35:17
I mean, I can think of several but the one. I think the one most important trait in a, it's, I think is, is more about being passionate and believing what you do without being affected by external influences. So, I don't think you should pursue success as something in itself. But in, just do what you believe, and you're passionate about, and what you want to do. If you're interested in, for example, a mechanism or you know, a research study or something that really sort of, you know, get you up in the morning, then you probably along the right tracks in doing something that you feel is meaningful to yourself. That might not necessarily lead to, you know, success in the different ways that you want to measure it, however, you know, by too, by today's standards. But at least you can, you know, look at yourself, you know, look back on yourself and go what I did, at the time I felt was worthwhile, or what I did know, what was the thing I pursued was worthwhile. I think that's something that defines what you feel is successful, then I think that's really what's important. It's not about what other people define what success is. I think it's all about when you looking back, if you if you're able to leap forward in time, you know, at your funeral, and go look back on yourself in your life and go what I did there, I would have done again, I think that's something you've achieved, that you've proved to yourself, you've done something worthwhile.

**Dr. Gabriel Alves** 37:06
Awesome. Thanks for your perspective on that. It's great.

**Dr. Ben Sun** 37:10
sounds good, but didn't,

**Steve Lewis** 37:13
Yeah, I was gonna say didn't expect the existential question, science. But it's only like, we like to keep the guest’s guessing.

**Dr. Ben Sun** 37:27
Well, I mean, there's so many cases in science instances where people will do the work and not necessarily rewarded in the same in proportion. If you look at the case in their DNA, you know, GFP, you know, there's so many things where, you know, success is not doesn't this hard work and doing

meaningful things doesn't necessarily lead to, you know, success in double quotation marks. But I feel that doesn't mean what you did, wasn't worthwhile to yourself.

**Steve Lewis** 37:58
That was Dr. Ben Sun, Head of Biomarker Genetics at Biogen in Cambridge, Massachusetts. If you'd like to hear even more of today's conversation, you can view the extended video version of this interview by visiting the URL in the Episode Notes. And if you enjoyed listening to our interview with Ben, try to share something interesting that you learned with a friend or colleague this week. We hope you brighten their day and pique their curiosity. This episode was produced by Matt Ferris, Sarah Briganti, and Matthew Stock