So, I am Ewan Birney. My current position is director of EMBL-EBI. And it's worth -- if anybody's going to look at this, my name -- my first name, Ewan, is actually a sort of family nickname that gets given so I'm actually formerly John Frederick William Birney and Ewan is the nickname everybody uses me by. So Ewan Birney is very much my scientific name. My legal name is John Frederick William Birney.

So, I was born in London in 1972, December the 6th, 1972.

So, there probably was. A guy called Bob Stevenson. I went to -- I did an English private education -- school -- public school in England just obviously, you know, the interpretation is quite the opposite so privately funded. And to Eton College and at Eton there was a guy called Bob Stevenson who was a really good teacher and scientist. And he fired me up for sure in biology and he was responsible -- at the time, we're coming to, you know, at the time Jim Watson who's basically a tremendous anglophile. I mean, he absolutely loves all things traditionally English. Jim had, I think gone to Eton and talked to some of the teachers there and seen and been entertained by some of the boys and decided to take a boy every year from Eton to work in what's called a gap year which is a very -- which you'll know is sort of a traditional English year that you spend between 18. After school and before university one takes a year out. So he made an offer to do that and Bob Stevenson was responsible for selecting the boys, so rather than there being some process it was just basically Bob's choice. And Bob suggested me. So when I was 19 I ended up at Cold Spring Harbor living with Jim Watson and working with Adrian Krainer.

So, it was -- I mean, again, when you're 19 it's -- you know, there's -- you're not aware of quite how odd and unique the situation is and slightly strange. So, I lived in Jim Watson's house with Liz, Jim's wife and Jim there sort of in loco parentis. His older son Rufus was also there in the house at the same time. It was very nice. You would -- I wouldn't spend too much time necessarily talking but you know, I was very -- we still exchange Christmas cards or whatever with Jim and Liz even now. And I'd walk in from -- they have a lovely house, top end of Cold Spring Harbor, the director's house -- and I would walk in every day and go to the lab. And Cold Spring Harbor was really, really fun. A really enjoyable experience. And that also showed you just sort of how open -- well, so I hadn't appreciated just how many unknown things there were. And so working in a lab where suddenly you're testing things and you have to work things out and everything else was really, really cool. I was only 19. It was a great privilege.

When I was there in Adrian Krainer's lab who studies RNA splicing really the databases had started to get big and people needed ways of searching them and looking at things and so I taught myself to program in that year. And a very lovely guy called Sanjay Kumar -- when I said, "Please, you know, can I -- you know, how do I do this?" And he was a programmer; I knew he was a programmer. And he said -- he bought me this -- the famous C book, this very thin book about how to program in C and I taught myself. And so my first paper was with Adrian before I went to university which was a very small paper about the presence of RNA binding domains in a particular protein that people didn't think had them. And we sort of just -- it was basically an alignment by hand. My second paper, which was far, far more substantial came about in my first year at university when I was still communicating with Adrian about this. And that paper which was on the RNA recognition motif -- I mean, it still gets cited now a little bit which is kind of cool. And it was, you know - it was a -- it was -- it was a good paper. So, that was a good thing to do. So I must have started that paper when I was 19 and it was published when I was 20. Published in my first year at university.

So, it's definitely -- and, you know, so obviously it gave me in some sense just as a timing sense a huge head start. But I think the other key thing it gave me was this -- a far better understanding of how science is done so that what you read in the textbook was not necessarily what was right. That, you know, when knowledge is founded by experiments and discussion and everything else. And I had a really, really great time. I mean, being British in America your accent makes you sound like you're, you know, you've got 10-plus IQ points. I worked really, really hard at Cold Spring Harbor and then I'd either go out in New York City or a friend of mine from Eton had gone to Harvard at the same time. And he hadn't had a gap year so he'd gone straight to Harvard so I'd take the train all the way up to Boston and hang out with him up in Boston for a weekend then I would come back down again. So, it was a really, really fun time. The other thing that I remember very distinctly is, you know, talks. When people came in to give a talk at Cold Spring Harbor everybody would gather. And one scientist, Winship Herr, would, you know, as soon as he didn't understand a slide he'd put his hand up and say, you know, "I don't understand what you're showing on this slide." And, he did it in a way where I think he -- he didn't do it -- he did it because he wanted to -- he wanted to understand what the other person was doing. He didn't mind looking a bit stupid sometimes. He really didn't care about that. He just wanted to make sure that he was following all the slides. And, again, I -- it's something that seeing really, really high end scientists feeling happy to ask stupid questions is -- was a -- is a good attitude to have to those talks, you know, kind of do I understand everything that's being presented at this time?

So, it's a good -- it's a kind of good rigor to have in your own head, you know, have they done all the controls? Has it been set up with the right fundamental thing behind it basically?

So Adrian Krainer who ran the lab in Cold Spring Harbor -- I mean, again, I think it's remarkable to think about someone who let a 19 year old come into their lab and then by the end of that year write a paper with them and then in the next year write a second paper with them. And that shows a lot of, you know, I think I -- I'm not sure -- I'm not sure everybody would sort of encourage that to happen and let me follow my own approach in this computational work. That Adrian wasn't necessarily so strong about but could understand what was going on underneath. So, Adrian was really, really good. And then in Oxford -- the Oxford system is four years and the fourth year you have to do your own project. And an electron scientist microscopist called Ian Campbell -- at that time I had published -- so I published a paper in my first year and in my second year I actually went to EMBL Heidelberg which is in fact now my parent organization that I work at. I worked with a guy called Toby Gibson and I wrote a second paper there called "Pairwise and SearchWise." And that is the precursor to a lot of my algorithms. So, I spent a whole summer basically in Germany in hot Heidelberg working with Toby. And I published two more papers then with Toby over my time. And then in my fourth year I went to Ian Campbell and effectively said, "Please, you know, I think I know what I'm doing with my own research. I have my own -- you know, I've got four publications already or something like that. Could you just let me fool around, you know, can I just have a computer? All I need is a computer and space and then that's fine. And you can, you know, be my supervisor. I'll obviously talk to you." And he said, "Fine." And at that point I knew that I wanted to do more things with databasing and more things around sequencing analysis. And the only open source database at the time was written by Richard Durbin and Jean Thierry-Mieg called AceDB. So, I wrote off -- I wrote email to -- at that point I had done all these profile things. I wrote an email to Richard saying, "Please can I come and learn how to use this?" And so he invited me. And I also gave a talk about my profile work. And I'd actually done -- I had incorrectly implemented an algorithm. Because I'd sort of taught myself all the way through. So I'd done something quite interesting in "Pairwise and SearchWise." Which wasn't quite correct. But, Richard Durbin and Sean Eddy was a post doctor and again, they both, you know, they sort of I think found my self-taughtness of computational -- you know, of these methods really quite interesting. So that was my fourth year at Oxford and at the end of my fourth year of Oxford I had to decide whether I was -- what I was going to do. And a couple of extra things here. So my father is an investment banker and my uncle is -- was an electron scientist microscopist and that was my uncle on my mother's side. And my father basically was rather skeptical that I would make a good career. That -- he felt that I was over focused on science. He felt that, you know, I was -- I was doing well but I hadn't -- you know, that I was going to choose this because it was interesting and then regret it in 10 years time because it didn't have enough money basically. I mean, not that he's very money orientated but he worried that I was closing down my options too quickly.

So at the same time I spent -- I did a summer working for an investment bank where I did all sorts of fun things -- including option pricing. And there was actually, I got quite close to -- it never happened, but I got quite close to spotting something that they would have created an option around which was kind of cool. You know, one of these exotic options. So that was one summer. I was mainly in equity research or pharmaceuticals there. And then, in fact the year after my -- after Oxford I spent a summer working for the mayor of Baltimore -- here just up the road in Baltimore. Which was a big education in kind of American life and politics and American politics. I mean, obviously not very scientific at all. So, both of those were slightly to prove to myself and prove to my father that I was making a good decision to stay in science. And at the time the Wellcome Trust had something called Prize studentships which was really designed I think for people like me where rather than -- so I did -- I applied to being an investment banker after going and working in the city. I, you know, went to Goldman's, I went to SBI. I had offers. Not from Goldman's but from a whole bunch of investment banks. So, and I think if there was just a straight British salary MRC style studentship where you really had us, you know, it wasn't enough to live on at all. It was pretty -- I don't think I'd have gone after it. But the Wellcome Trust Prize Studentship was healthy enough that you could live and, you know, you could have a car and, you know. Anyway, so I said yes to the Wellcome Trust Studentship with Richard. And so that's how -- that's how I got to Sanger and working with Richard basically.

So, I mean he's one of my -- he's still one of my collaborators. I mean, it's really interesting now being whatever, mid-40s thinking about that. Because I sort of started knowing of him in my 20s and then working as a student whenever I was 24 or something like that. But he's a remarkable guy Richard. You should get an oral history from him. So, he -- a mathematician from Cambridge. He's got these [unintelligible] things. I think he had a basketball scholarship that he came to Harvard for or something like that. He was a basketball player, a volleyball player -- something like that. And my, you know, maths in Cambridge is the highest level. And

I think he was -- I mean, he's an excellent mathematician but he's not one of those pure mathematicians. And so he had to study precisely what he was going to do. And one of the things he was doing and I can't remember quite how this worked with his undergraduate or his graduate work was that he started doing Confocal Microscopy. He programmed the confocal microscopy software and stuff like that. I think that was before he did his Ph.D. And then he did his Ph.D. with John Sulston and that was very much to bring computation in neural networks alongside the C. elegans developing brain. And I think he mapped out all the -- or many of the neuronal connections of C. elegans. Which is for somebody who's a maths person that's a -- that's a pretty big journey already. And John Sulston had a -- obviously a passion for the worm but a passion for genomes and passion for getting these things done. And so, as the worm genome project was coming up it was kind of clear that they needed information systems to do that. And I'm not sure how this happened but somehow Richard met and started working with this slightly mad Frenchman called Jean Thierry-Mieg who always came with his wife sort of this husband and wife -- Danielle -- pair. And Jean Thierry-Mieg wrote an open source database. Now, these days we would call it a document orientated semi-structured database. At that time they just wrote something that worked for this, yeah? And it's going to look, you know, it looks very clunky now and you've got to remember that this was just as Oracle, the database company Oracle was coming up and selling Oracle for massive amounts of money. And no academic would ever afford that. And also, before the web. And actually their database was an open source database before the MySQLs and the other famous open source databases came through. And it also had the concept of hyperlinks inside of the system of AceDB it had an integrated graphical interface with the database. And so, pre-web you had this rather amazing thing. And in fact, back in Adrian Krainer's lab I remember, so there was this year I did as my gap year. The next summer I came back to Cold Springs Harbor as [unintelligible] and that summer we installed AceDB on their Sun machine. And I remember, you know getting it up for something that's absolutely amazing. So Richard wrote that piece of software for the worm project and basically was -- has been this common thread through the worm genome project and then the human genome project and then in fact into the HapMap and then the 1,000 genomes and everything else of where his real skills are is algorithms and understanding what one should do with sequenced data. He also -- practically he did write this database and he practically did a number of things. There was a period when I was a student with Richard where the Sanger really ran on his software and occasionally, you know, things would break. Richard was the only person who could go and fix it and he would have to go and fix it. It was all not how you would run a system now. But he's a total legend. His algorithmic -- his maths and algorithms is just really top notch. And his understanding of biology and so it was a real pleasure. I mean it was a real experience actually working with him. It was the first time I had, you know, I had been self-taught and I had been the best programmer, sort of, or the best person up to then and then I met Richard and then it was very clear that there was somebody better. Yeah, I was going to -- somebody better than me in all these things. So it was -- it was -- that was, you know, it was -- it was -- it was great. And in my Ph.D. I both did a practical thing -- practical things and then I did very kind of algorithmic things. So my Ph.D., the practical things was running the Pfam database which was also a slightly kind of -- it was all -- again, you just wouldn't do it this way now. I mean, it's sort of -- it's generally -- not chaotic but kind of Heath Robinson-like approach to these things. And at the same time, I was focusing on developing algorithms, which I had thought about from "Pairwise and SearchWise." And sort of generalizing those algorithms and to sort of match how the maths worked with how the algorithm worked I wrote my own little miniature programming language. Now, again, because of my sort of slightly self-taughtness this was very poorly -- it's not a good piece of computer science. However, it did elevate dynamic programming, which is the key -- you probably -- key sequence alignment methods. It sort of elevated that as a first class language primitive. And, so I was able to write far more complicated algorithms. I think I probably still am able to write far more complicated algorithms and just, you know, throw them around, change things, test things, you know, do things in different ways and then know that my programming language would accurately generate all the right code for it so I didn't have to worry about debugging it and things like that.

As well as it being kind of practical for the -- for the final set of algorithms that I wrote, it actually gave me huge amounts of freedom to experiment. And that algorithm, you know, the best algorithm that came out from that was an algorithm called GeneWise which takes a protein sequence and matches it to the genome but it sort of handles splicing and handles errors. And in my thesis -- my major of biology thesis was about the language, which was called Dynamite. There's only one user -- that's me, really. There was a second user but she didn't -- unsurprisingly it's one of those things where there's really only one user, which is the person who created it in the first place.

Actually, I think there's a lot of computer science like that. I think a lot of computer scientists end up -- it's actually a very interesting debate about how you provide abstraction because, you know, ultimately, you know, when you start realizing, "Oh, I can write my own computer language that does this." And very often that starts off by writing -- itself writing a different lower level computer language like C and that's the major thing that Dynamite did. But actually I did go all the way to producing things that, where I could manipulate everything. So although I never wrote assembly compilers I got quite close because I tried to target it to different architectures where I had -- I had a different set of primitives to use, yeah. So, that was, you know, that was kind of fun obviously. But the practical thing that came out was GeneWise, which took a protein sequence say from a mouse or from rat and could map it to a genome but of a different species. So of human. I could deal with the fact that the splicing -- you didn't know where the splice pattern was and you didn't know where errors were. And that algorithm ended up being, you know, the most robust algorithm for predicting genes in the human genome, which was a big topic of conversation at the time. It was extortionately, computationally expensive, but very robust and did a better job than nearly anything and it's still running today.

So, sort of at the end of my Ph.D. obviously the -- so the worm genome was chugging along very, very well at Sanger. Very clear it was going to work. And next door to us were the people analyzing the worm genome which they did a lot effectively not by -- not sort of in a -- in a computer assisted but fundamentally by hand way. And, there was a project to do a sort of similar thing with the human genome kind of matched -- you know, matched for the flow -- rate of flow coming through the genome project. Remember that the human genome project was projected to finish in 2010 or something like that with a pretty nice slow, steady beat all the way across the genome. So, it sort of -- halfway through or close -- yeah, halfway through Celera happened and it was this, kind of, you know, explosion of fear and excitement for the Sanger Institute. You know, excitement that this -- that, you know, the Sanger Institute was -- or the Sanger Center at the time was one of the most important parts of, you know, it was sort of validation that this was an incredibly important project. So everybody got quite excited about that, that there was this, you know, slightly mad American raising money in the stock market and doing all of this stuff. But then there was this sort of awful realization that everything would have to speed up and, you know, so one had to work through how one sped everything up. And some things were quite easy and other things were very unobvious about how that was going to work out. And so, there's a whole kind of saga here about moving away from a mapping first approach to a sequence backs as they come through and stuff like that. And I wasn't in the middle of it but I could see the kind of, you know, you could -- you knew that that debate was happening and how that was changing. What was interesting as they accelerated the speed of doing this -- the public project -- and making announcements that they would accelerate and match sort of Celera's run rate to this. Then Celera was also saying, "Well, that's great. We'll use the public data. That's all wonderful. Of course the public people don't know how to analyze it so everybody's going to be coming to Celera anyway because we've got the brains. And they had this, you know, very, very clever computational scientist called Gene Myers who was the person who originally sort of said that it was feasible to do the assembly. And he was a good friend of Richard Durbin's actually. Used to go back, you know, another thing -- despite all the kind of nastiness between the Celera project and the public project especially in the computational land there was a lot more sort of mutual respect of each other. And I'll come back to that through this. So anyway -- so it became clear that we had to have a solution to the analysis. We or Sanger had to have a solution to the analysis. And then three kind of things came together around that. So one was someone who Richard had hired to take over the running of the human annotation. That was Tim Hubbard. And he had seen this and he had created an even more extremely Heath Robinson system than the Pfam thing to try and keep track of all the bits of DNA that was being sequenced publicly and run it through some very, very basic analyses. Some time he had hired Michele Clamp and Michele I had known. She had been a post doc with Jeff Barton who I'd known from Oxford so I'd known them because of my undergrad at Oxford and I knew that there was this sort of slightly physics woman who could code really well and she had somehow done great things with Jeff. And so, she -- they -- she and her then boyfriend, James Cuff got hired to the EBI and then she got hired from the EBI to Sanger to help run some of this annotation stuff as well. And, me and Michele and James really hit it off actually as in terms of we all wanted to just, like, you know, make new things happen. And, so there was a combination of me and Michele also thought that AceDB really wasn't going to work for the -- for the -- for this. It just wasn't going to scale. And so we got this new fangled thing, MySQL, which nobody knew whether it was really robust enough and stable enough and was up to the job. We brought that in house. At the same time, Sanger had brought in Oracle in house to run its major database thing also realizing that AceDB really wasn't the sort of long-term solution. That always kind of annoyed Richard because to be fair to Richard, lots of things of AceDB worked well but some things just worked awfully, absolutely awfully. And, you know, so there was this sort of massive right lock on the database it was just a -- just not going to work. So, we brought in this MySQL and Michele and I were effectively exploring how to make a database schema for genomics and I was exploring, I had written GeneWise and I had -- was, you know, getting more and more confident that that was, you know, that any prediction that came through GeneWise was going to be pretty good, almost certainly correct, asterisks, except for pseudogenes predictions. So there's -- there is an asterisk there for this. And then this need to respond to Celera kind of drove this next phase of us really creating something. And this was just as I started my Ph.D. was about to finish and so another slightly complicated thing to this. Oh god, you see, this is why oral history is a good idea. So, I'd written this programming language called Dynamite.

And I was thinking about targeting it to different machine architectures, not just my own -- not just standard C programming. And there was a company called Paracel whose major customer was the NSA for text mining and stuff like that. And they had identified DNA matching as their other big opportunity. And my code, Pairwise and SearchWise and then GeneWise -- was probably -- was incredibly computationally intensive but worked directly off DNA. Most things didn't do that and was clearly good. It was clearly a really sensible thing to do. And so they really felt that was a match made in heaven. So they flew me over to California a number of times -- oh, and at the same time I was also doing the same thing with an Israeli company called Compugen had also done alternative hardware. So I was going to Israel to help see how my programming language could work with Compugen and I was going to California to see how it was with Paracel. And because that and it was the dot com boom. Then I was invited to be on the Paracel's scientific advisory board. And I was, you know, I was kind of a kid then but the other person on that scientific advisory board was Gene Myers so we were the two computational people. And then there were a couple of other people. And this was -- it was an independent company at the time. So, I had a lot of good interactions with Gene on the scientific advisory board in discussing how one does dynamic programming, and you know, or less how one does the algorithm and more what it's useful and how to construct it and how to make -- how to do different things. So, as I ended up my Ph.D. it was very clear to me that it was a bit unclear what I would do next but I had lots of options and it was the dot com boom that era. And, you know, Celera had founded with lots of money and there's -- you know, some startups and all sorts of different things. So I took myself off and went around America and I visited all sorts of places. Celera -- I visited Celera for a job. I visited Paracel for a job. I visited academic labs. I went to all -- I went to all sorts of different places. And, somehow my last one was Paracel and I remember them, you know, giving me a job offer for more money than I ever thought was sensible to give to a scientist. And I kind of -- I sort of said to myself, "I'm, you know, I'm going to take this job. This is my opportunity." And then I said, "No, no, no -- I better go home, talk through with my girlfriend, who's now my wife a little bit about that." And also just go home and talk with my -- talk it over before I signed on the dotted line. So, after this sort of week and a half this was still the height of the human genome sort of stuff, you know, and so it's quite surprising, you know, there's this whole narrative about people, you know, throwing shit at each other publicly through press releases. And yet, a perfectly pleasant and collegial conversations between many members of the science -- scientists in different parts of the project. So there's quite a funny contrast of that.

Well, I think -- I think -- I mean, I think there is a very particular Sanger and Wellcome Trust view as well. I think that -- I personally think already that there are alternative histories being written about some of this because I think it is - it is true that John at some point got the backing that if necessary the Wellcome Trust would finance Sanger to do all of it. And for John, I think that was an incredibly important thing that he could then come to all of these conversations and say, it does not matter what you guys decide, we will do X, Y, and Z, yeah? And I think that, you know, how important those -- that is is quite interesting question, yeah? But as part of the Wellcome Trust/Sanger mythology, you know, if that commitment hadn't been made and if John hadn't made those statements there was potentially a different branching pattern for what happened next.

So, anyway. Going back to -- so the geeks, were, so -- we got along I think well. And, so I'd gone around all of these different places. And before I'd left Richard said -- Richard said to me, "We want to keep you at Sanger and why don't you run the mouse annotation group?" Which was going to be two people and wasn't kind of the heat of the problem which was the human annotation stuff. And I kind of said, you know, "I don't -- I'm not sure I really want to do that." And I went around. So, when I came back, Richard was incredibly keen to almost immediately talk to me and he sort of almost physically dragged me off to the EBI to talk to Graham Cameron. And Graham -- and Graham and Richard sort of in the time that I had cooked up the idea that EBI would offer me a position and back half of the annotation project and that was what triggered the Ensembl project being a joint project between EBI and Sanger. And the commitment for -- of EMBL was really some money but also making me a PI. So I became a PI -- I actually became a PI before I got my Ph.D., which was a little bit. I got appointed to EMBL before I got my Ph.D. There's this letter from Francis [unintelligible] that says basically if you don't submit your Ph.D. in the next month, then, you know, this is going to -- you know, you know. This I kind of, yeah, things are going to explode. Awful things will happen. And at the time we were building up to the big Tony Blair what ends up being the announcement thing. So this was sort of -- this was January before that announcement. And so I was working like an idiot with Michele and other people and trying to get everything to work. And then, you know, then in the evening I was trying to finish off my Ph.D. It was just excruciating. But anyway, that all worked out. So, we got -- we set up Ensembl and then we also realized that it had to be funded. So we had a meeting with -- yeah, exactly. We had a meeting with Wellcome Trust kind of in the back of the room. And, so the first thing is we did a number of proposals. And the first proposal had something like either people and John came back with this -- almost like a -- I remember the one, almost one line email that said, "Double it." And so we wrote it for 16 and John said, "Double it again." And so we wrote it for 25 or something like that which was a huge grant. I mean, you know, when most people -- what are you doing? So, you know, it's now a team of about 45 people. I now am totally used to the idea that this kind of engineering requires quite a lot of -- it just requires a lot of personnel and muscle and stuff like that. And we had a meeting with the great and the good there and again, it's one of those cases where the Wellcome Trust clearly had to take a risk. I mean, if it had gone through normal peer review it would have just been torn into lots of little pieces. Plus the fact that it has 25 people to it, yeah. And, you know, no panel would have -- would have swallowed that at the time especially with someone who is straight out as a graduate student. I mean, it was ridiculous. But nobody knew what to do and they knew there wasn't enough computational biologists and they knew that I had a lot of, kind of, you know, self -- you know, delivered by hook or by crook. By persuading, you know, all sorts of different things. And we -- there were four -- three or four PIs. Richard, Graham, myself, and Tim Hubbard. And it's too my regret that Michele Clamp wasn't a named PI. I think that's right. Because I think it's quite easy -- I see Ensembl being founded by Tim, Michele, and myself. And the fact that Michele was sort of under Tim was a -- was always a bit of a -- something slightly wrong about that set up. Anyway, so Ensembl started up and that gave us then a very big, you know, effort. We -- and we really did deliver that and make that happen over time. So, if you look back at it again it looks clunky but a lot of the, for example, the fundamental, you know the fundamental data model that we had was, especially of the genes, transcripts, proteins, it's all obvious stuff. But it's actually the, you know, it's stood the test of time. The same concepts hang around, basically. And GeneWise, for example, was right in the middle of the gene prediction, but Michele's code is well about how you call GeneWise, how you make -- where you do it, how you call it where you do it, and how you tidy up afterwards, which is a key part of the whole process. Still, I think it's almost the same now, as it was way back then.

I can't remember which Cold Spring Harbor this was. So this was 2000, I think. Yes, it was 2000. And it was very clear that there was one session, and it was all about gene prediction, and I was there explaining Ensembl and GeneWise and stuff like that. Jean Weissenbach and Hugues Roest Crollius from Genoscope, with sequencing takifugu or tetraodon, I can't remember which one, and was very smart. And they had used those reads to estimate the number of genes in the human genome. And they had come up with this shockingly low number. And it was already kind of rumored before we had the actual session in Cold Spring Harbor that this was going to happen. And if you remember, I did spend some time in this investment bank. If you spend time with traders, you discover that you, you know, they bet on anything and everything, you know, anything you can bet, they will bet on. And back then, I -- you know, it teaches you a lot about running markets, so I actually ran a book once for -- so be the bookie, to go-cart racing, with these people. And it's actually very, very interesting when you're the bookie, rather than the person betting, because you have to offer odds which are long enough to attract bets, and short enough -- you know, as soon as you run a bet, you have this phrase being over- round or under-round, and you always want to be over-round. Over-round meaning that you win no matter who wins the race, no matter the outcome, exactly. And that's basically -- traders have to be in a, you know, you want to be in an over-round situation, not in a situation where you're asking for results to go your way. So I had that experience. So I realized that it would be fun to do this, and I -- there was a time when I wondered whether I should kind of offer odds, but I realized that wouldn't work at all. It was pretty clear that, you know, you couldn't offer odds on the number of genes. So I set that up as a sweep skate. And I didn't realize this, but Francis Collins had the same idea, and was a bi pissed off [laughs] that I had suggested this. But I just sort of waved this book in the air, and of course it was at Cold Spring Harbor, so I had to go into Adrian Krainer's lab, and say, "Could I just have a lab notebook?" and I pasted GeneSweep, whatever. And that -- so I had great fun that night. So I had to write down the rules of the bet, and then take on bets. And we -- I decided, this was all decided sort of on the hoof that you could buy one -- so this was the only sort of fun thing, slightly different thing I did, because you could buy one number for one dollar on the first year, so it was going to be decided in 2003. And then it was going to be five dollars a bet on the second year, and then 20 dollars a bet on the third year, and the fourth year we would decide. And so that's all written in. So the idea being that the information was much better as you got later in this. But it was a great -- you know, I think lots of -- I met lots of people, lots of people remember me because I was this young Brit, I used to swear a lot as well, as a much more sort of dirty mouth at the time. Young Brit, precocious, doing this, and then I came round with this book and persuaded effectively everybody in the meeting to put a dollar and put a number in. And of course the really amusing thing about it is that everybody was wrong. I mean, everybody was absolutely horribly, everybody overestimated the number. And it shows you that when the crowds with no data -- you know, so, crowds with data probably will make a good estimate. Crowds with no data, absolutely don't make a good estimate. We just -- we were all doing it from, you know, sort of crappy backlog, and stuff like that. And don't forget that Insight was around as an EST company, and they were selling access to 100,000 human genes. So it was considered to be quite radical to put down 50,000. You were really saying that [laughs] Insight was lying, basically, to put down that kind of number. And for Hugues to get up and say there are 26,000 predecoding genes in the human genome, I mean, you know, people thought he was a mad Frenchman. I mean it was just -- I mean, people really, genuinely thought he had lost it at some level, for there to be that low number.

So by the time we got to -- so the interesting thing is Hugues put down his estimate, which was, I can't remember what it was, like 26,200. And a couple of people realizing the sweepstakes rules sensibly went below Hugues. But only two people went below Hugues, which is, when you think about it, very poor strategizing by, you know, those probably 500 votes. And it's only two that go below, who didn't go down, yeah? So when we came to 2003, and by the rules of the bet we should settle it. And actually the number still wasn't known at this time, and in fact, if you wouldn't actually have a number now, you'd have a range, a band, which would be above 19,500, below 20,500, yeah. With a lot of kind of definitional stuff going on.

So we decided in 2003, or the suggestion was is that the pot was split three ways, between the last two people and Hugues, and Lee Rowen from Seattle, she actually had the lowest, lowest bet, so she kind of won. But I really gave Hugues the biggest and the most amount of credit. I mean, nobody was down there, except for the fact that Hugues was down there. If Hugues hadn't gone down there, nobody would've written a bet below 30,000, I think, which was kind of interesting, for sure. Yes, so that's the story of the GeneSweep. Also, I met lots and lots of people at Cold Spring Harbor, though.

The human genome, the draft of the human genome was the first time we'd done anything of that scale in terms of genomics. And it was the first time we'd done this big consortia sort of method. And it showed that it was the first time that -- the analysis showed, the paper structure showed, whatever. By the time you got to the mouse genome, it was better, far better. By the time you got from the mouse genome to the chicken genome, even better. You know, we really honed it by the end. So it was done in an ad hoc way, for sure. An absolutely key individual here, and again I really hope you get his oral history as well as Jim Kent, because, you know, their, you know, without -- so, David Haussler and I, and I think Jim, one of the Cold Spring Harbor's, and I don't know if it was 2000 or 1999. There was a whole business of putting together the human genome, and even how one thought about an assembly. And so we had this thing called a golden path, which was a concept from Phil Green, in his assembler, which was a golden path of reeds. And we sort of took that up into an assembly concept. And a very early version of Ensembl allowed simultaneously different golden paths to exist, so different alternative assemblies. So our very first design, Michele and mine, very first design had this. And, you know, actually these days -- so somebody could, you know, people talk about this now, for graph genomes, the idea of having flexible, more than one assembly with the same backbends, and stuff like that. And we've never really done it, yeah. And it was quite funny that we wrote the system at the start to do that. And in fact we ended up throwing away that part, just dealing with a single linear reference. It's slightly sad that we did that, but still.

So Jim was absolutely key, in doing this. And in some ways it became clear that Jim was making this wonderful browser. We were, as academics are, competing and collaborating in some sense at the same time. And he had thought through a way of making the assembly. And so Michele and I focused on making genes. And then, you know, putting the paper together, the person who really made the -- who sort of drummed, made a drumbeat for the paper, was Eric Lander, for sure. There were many people sort of round him, so John Sulston and Bob Waterston were sort of with him. But Eric was the person who wanted to put his fingers in everything during the analysis. And he kind of called -- and I think he gave the name, Hardcore Analysis Group, was from him. And, you know, I think that's both because of his enthusiasm and his desire to control, so that, you know, both of those things are wrapped up together. And so that's how that kind of process happened. And then there were always these phone calls to coordinate the production of the genome. And then came another run of phone calls around the analysis.

Michele and I and Tim, would go to the G5 calls occasionally. But that is -- we're always very focused on production, yeah. And it became more and more obvious, as things developed, that we needed a separate thing that was more analysis-focused, and that's how the Hardcore Analysis thing came. Now there was a meeting, I remember going to Boston when, you know, when it wasn't the Broad, it was still the Whitehead, and in that kind of, that funny semi-factory-like building, and it was all snowy. And I sort of remember that as one of the first physical get-togethers of all these analyst-style people. But it was a real raggle-taggle bunch of people. And when you read the analysis, it's not very good. And I think, I've got to be honest about it [laughs]. It was the first time we did it, but it's not really very good. In our defense, the Celera analysis wasn't very good either. And the Celera analysts were also using GeneWise, for example, in the middle of that pipeline. So -- and there was this whole hoo-ha, less -- more for this Tony Blair, Craig Venter, you know, Bill Clinton announcement, about the number of human genes. And that led into -- it was all sort of wrapped up in the same thing as this sort of betting book thing. And, you know, that was a great example where I was tempted to send a little message to Margy Andal at Celera, and say, "Okay, what number have you got? Why don't we just agree, you know, together so that we're not all totally out of whack?" Because there was this huge fear that we would, you know, Michele and I were making estimates in the 20,000, which we called 20,000 confident human genes. And we're basically being told that that number looked too low. And I remember Michele's first estimate was up, it was about 24,000, whatever, and we felt it was too low. And when we presented one of the G5 cells [unintelligible]. Like Celera is going, you know, it was "This is going to be awful. We're going to look bad, that we can only find 24,000 genes when Celera can find 35,000 genes," and whatever, yeah. And InSIGHT has 100,000 genes.

That did color me for some of the later things. For example, with ENCODE, where I've -- you've got to stick by the data analysis -- you got to say to yourself, you've got to have a good talking to yourself before you can open the box, and you look at it. You've got to say to yourself, "Okay, you know, which things of my analysis am I confident about, which things am I going to question if the result doesn't look like how it looks like?" I wish I'd stuck to my guns, Michele and I had stuck to our guns a little bit harder. That we were closer to the money on our first analysis, far closer than the initial draft paper. And it has an awful phrase, something like, "We can find confident evidence of up to 26,000 genes," and, you know, "exploratory evidence up to 35,000," and I mean, you know.

But the foldout's kind of interesting, because, you know, we were putting together a paper in the Hardcore Analysis Group, and there's just -- another awful bit in that paper, by the way, is "evidence for horizontal gene transmission." It got shoved in at the end. Again, you know, we got much better at doing this process, of checking things, you know, of not having these sort of last-minute "Oh, my God, I've found the most amazing thing ever," thing coming in. But when we came round, and then Francis said, Francis in one of these phone calls said, "We need a poster. And we need a poster, and I want to have a poster with every gene in the genome on the poster." Now it's quite hard to -- it's quite hard -- you know, you can't do this by hand. Yep, you can't do it by hand. So -- and it's quite hard to do, it's actually quite hard to do good layout. It's hard to alternate good layout of these things. And so it's huge, and so -- eventually I said, "Well, I can program -- I can write postscript." And so I wrote a system to write these posters, and -- that's why I took a photo, because I love when I see it, because I don't think anybody appreciates the level of detail. One of the more complicated, no it's not, the most complicated algorithm. But I had to use my dynamite programming language to write a very specific model, to get the layout, to look aesthetically correct, yeah? So afterwards, I'll go to the corridor and I'll point out all the little things that the algorithm has to do to get this lovely sort of shape in the, yeah. So yeah, there I am, writing some layout, postscript layout program. And I can remember doing this vividly, because this, you know, the production timelines of this had to be earlier than some of the other ones, so that the foldouts would happen, and stuff like that. And so I had to -- and I worked with an artist at NHGRI, Darryl Leja, and so in the middle of having all these other deadlines, I had another, you know, deadline, which was producing posters that people like the look of. So I'm very proud of those posters, at the end of the day.

I certainly had the feeling that we hadn't -- the draft wasn't anywhere near the end of the human genome. It was, you know, it was, you know, I knew how messy it was. And I knew how this -- we couldn't leave it like that, basically. So that's one thing. At the same time there was a whole thing about the mouse genome going at the same time. It was very interesting there, because of course there was this whole business the whole genome shotgun wouldn't work, and then suddenly Eric kind of learned to throw himself and took a complete 180, and said, "Oh, no, whole genome shotgun is totally fine for this." And, you know, it shows that Eric is smart. It also shows that he will change his mind on a sixpence when necessary.

All that set aside so there's a kind of theme of genomes. And the mouse paper, the mouse assembly was better, the mouse gene set was better, the analysis was better. We knew what we were doing better across the board there. And then by the time we got to chicken, for example, we were, it was, you know, we were like -- I mean, there were problems with chicken genome because of these micro- chromosomes. But kind of how one thought about doing the analysis was now becoming really a much more structured process in many ways. So there was that theme. But the -- so it was also very clear that there were going to be two threads afterwards. I mean, NHGRI said this, but also it was clear inside of Sanger as well. So one of them was human variation, HapMap, that ends up in 1000 Genomes. And the other one is basically beyond protein coating genes, right? So what is beyond protein coating genes? And I was interested in both, but because of Ensembl being so focused on annotation, I really felt that the, you know, the thing I should throw myself into was the, you know, stuff beyond protein coating genes. And so that, of course, ends up being the ENCODE project. There you have the bit before the ENCODE project, which is, "What should we do, and how big should it be," and all of that. So we had those meetings. And I think something that people sort of, again, forget at that time is that we really didn't have good technologies for studying the human genome at scale. So this was in the epic period of arrays, micro-arrays, and then tiling arrays. They were awful, you know. You never -- they were just, you know, they were not -- some people are nostalgic for them, but I am not nostalgic for them. They are -- they, I mean, they -- if that was the only way you could do things, as in the first ENCODE, that would be the only thing you could do. But they were absolutely awful. Batch effects all over the place, completely impossibly to compare between platforms.

So the first ENCODE, it was clear that this technology was not going to scale across the entire genome with tiling array technology. It was clear that we didn't know what we should even be measuring, lots of people had different ideas. So, I think, you know, the right decision was not to try and do genome-wide things, but to instead concentrate on one percent of the human genome, which was this one percent project for everybody to do. But we also were bringing it into a completely different community of people, so all of these sort of transcriptional cell biologists, chromatin people, that sort of stuff. And so there was also a sort of big social thing going on about this, and I think it's right -- I mean, it's interesting, so, are these very big international consortia the right ways of exploring and doing these things? One, you know, there's post-hoc justification, because these things, this is the way things happen. It's far less clear-cut, what is the right structures for these things. However, that one should do it was very clear-cut, I mean, you know, that scientifically one should go and work out what was going in the rest of the genome was very important.

So ENCODE 1 -- so I can't quite remember quite how so -- Ian Dunham from the Sanger was one of the projects that were funded by NHGRI in ENCODE 1, and we were involved partly because of Ensembl, and maybe for the genes as well, through that -- through Sanger or something that. Anyway, I turned up at all the meetings. That's probably the most important thing to mention here. And in ENCODE 1, it was very unclear about how we should talk about the results. And so there was a decision that there were going to be five different papers. This was an awful decision, but, you know, it was cat-herding, and not with cats, with lions. They really didn't want to get on. Everybody had their own view, everybody was jostling for position. So you had five different sort of things, and therefore five different analysis groups. And I think I did one, that's right, around comparative genomics with Elliot Margulies, who's here at NHGRI Intramural. And anyway, it was just a big mess, basically. Now how -- I'm buzzing, let me just do a little check that -- okay, cool. So it was a big old mess for these five things. And probably the most important thing, which -- we sent them into review at the same time. So two things about this: So somehow, I think I became chair because of my experience with both the human and mouse genome papers. Francis appointed me, basically, as chair of the analysis group, and I ran these phone calls, I was used to this now. And I obviously had the computational smarts. But there were a lot of people who really didn't, you know, I didn't know a lot about chromatin, I didn't know a lot about quite a few things there, I mean -- on the flip side, they didn't necessarily know about computational techniques and things like that. So it was very interesting, kind of just trying to manage all of these people. And as we tried to put together these five papers, I mean, there was an awful moment where there was a false marriage of two groups to lead one of the papers, John Stamatoyannopoulos, and Anindya Dutta. And for example, I had to effectively chair a four-hour phone call where we went through the ordering of the authors for that paper. We went down to sort of the, you know, I can't remember, like the tenth, you know, number 10 in the list before we went into alphabetical, or whatever, and then the last -- and I had never really been -- I hadn't myself been exposed to that level of distrust between scientists on writing papers and things like that, and all of that. It was quite novel for me to see all of this. So we had those five papers. The -- it was, you know, it was all over the shop, it was very clear to everybody, you know, every time you did statistical analysis, you had this massive lab effect, so experiments clustered by lab more than they clustered by, you know, any other feature. This is, you know, it's actually very, very common on lots of micros. I mean, you know, so lots of micros have the same aspect, so it wasn't like we were awful, it was -- well, and we were pretty bad, but it was a feature -- it was an aspect of the platform that, you know, people didn't really like talking about, including in gene expression. So Affy[max] arrays, and Lumina arrays, if you took them, you clustered them, you would cluster by array before you clustered by anything else. And also the experimental design was also awful in the sense that people were being given freedom to do experiments on the cell lines they wanted to do them on, that they would make work, rather than having any sense of a common cell line. So we could also not really do comparison, we didn't have anything that looked like a good matrix of comparisons. So, anyway, the reviewers, unsurprisingly, read this, and said it was crap. [laughs] So it came back with, you know, "This is awful," was the, you know, "What is going on?" And in particular, I think the thing that particularly annoyed the reviewers at that time was that, you know, paper two says something, and paper four says something, and when you read these paragraphs they are in conflict with each other, you know. And it's okay to have two papers from two different groups saying that, but not two papers that are coming off the same data sets, or supposedly coordinated so you can go work it out. So then we had an awful kind of meeting where I chaired the analysis group, and I basically -- I think I had talked to Francis, or Francis and Eric together, and I said, "Well, look, obviously, we've got to have only one analysis like the genome papers, and obviously that's the only way through it. And we've got all these problems, we've got to chuck out a whole bunch of these things. We're not going to be able to talk about everything that everybody can talk about, because some of it is not -- it's not solid enough for us to be able to talk about the things that we had hoped -- you want to talk about, but you can't." So I'd kind of agreed that with people here, maybe that was more with Eric than with Francis. And then I went into a huge meeting with everybody there, and basically I talked them to death, so that the only answer left on the table was one paper. And so then we spent another year or something putting together this one paper. It was like the march from Moscow, it was really not the, you know, not a nice experience at all. But we did, and the paper's pretty, you know, it's a bit -- I see it very similar to the draft human genome paper, so it's not a good paper. It is the first paper that tries to a number of those things at that scale. Perhaps a bit unlike the draft human genome, actually the take-home message is, "We don't have the right technologies." But interesting enough, you know, after all of this hiatus around this, what actually had happened was that Solexa had come online, and was -- at this period had been bought up by Illumina, I guess. And it was very, very clear that -- it was very clear that there was a completely new option on the table that could scale incredibly well, which is rather than using tiling arrays, use sequencing as your readout. And I remember then writing people -- there's a question about what ENCODE 2 should be, and whether it should be whole genome or not. And again, I think if you just took the results of ENCODE 1 at face value, you'd say, "You must be kidding, this is not going to work out," you know. We've got bad enough batch effects -- at one percent it's just -- it's going to be impossible. But there was, you know, enough of this new technology had come out to persuade people that it was doable. And so lots of people wrote, I mean, the grants for ENCODE 2 very much with a, "Well, Plan A is tiling arrays, but as soon as we can go to Plan B, we'll go to Plan B." And in fact by the time the grants were awarded, everybody had gone to already to Plan B, which was doing it with an Illumina/Solexa readout. And that was much better. And the other good thing between ENCODE 1 and ENCODE 2 is, I think I had a -- because of the experience of pulling together the paper -- I mean, I had saved the projects. I -- well, not I, sort like an individual -- but pulling this together had prevented a project, the project looking like a complete disaster, to it looking like, you know, a success of some sorts, you know.

So I had got the trust a lot more. I wonder when all of these histories come out, but -- you know, I got the trust a lot more of NHGRI project staff, so Elise Feingold, and Peter Good. And I think they, you know, they had a lot of people pushing lots of different opinions on them about how to structure things, and how to do things. And so I became a stronger voice in that. And one of the things which I was kind of happy and proud about was not only justifying ENCODE 2, but dealing with -- saying "We must have common cell lines. We must have a core set of cell lines that you must do these experiments on." Those were called Tier Zero, Tier One, and stuff like that. There were six of them we chose. And that made -- the ENCODE 2 was much, much better designed, so the sort of fundamental experimental design of ENCODE 2 was better. We then also learned about the Q.C. process better, and so we were at least trying to address Q.C. up front, though in fact we -- I don't think that really came to fruition until ENCODE 3, realistically. But at least we were -- I mean, we were much, much better. We were in a completely different space from the first ENCODE. So ENCODE 2 was good, actually. And although there was, you know, still people rubbing -- so also all the social stuff that had gone wrong in ENCODE 1, effectively some people had just got so pissed off they didn't bid for ENCODE 2, so that's one way of solving this problem. But also people had sort of come down a little bit, and known where people are coming from just understood each other a little bit better, more. So although there was still jostling about the PIs, and about who was doing what, who was in charge, and stuff like that, and whose asset was the most important, or whose viewpoint was the most important. It was a much more collegiate process.

The consortium acted like a consortium far better. We had a very functional, I think, analysis group. But that was saved for sure by two individuals. So Ian Dunham and Anshul Kundaje. So Ian Dunham was previously at Sanger, and Sanger made an awful decision, an absolutely crazy decision, where they sort of, they cut about 20 percent of the Institute out. And they sort of pretended that it was not -- that it was strategic rather than quality. And because of that, there were a lot of people who weren't, you know, weren't tremendously good who were cut out of Sanger when the 20 percent got lost. Two really, really excellent people were cut. One was Ian Dunham. The other one was Stephan Beck. I do not understand why somebody didn't say, "For God's sake, guys, let's, you know, just, these guys are the guys." So they, you know, it's crazy. And it meant that Sanger didn't still, you know, didn't have the depth of functional studies on the genome for a long period, and they had to rebuild it. Anyway, so Ian had kind of got effectively a redundancy package from Sanger, a very nice redundancy package, he could go wherever he goes. And he was being interviewed to go for -- institute directors, people were trying to court him to do all sorts of different things. But Ian's -- he's a great guy, and -- two things about him, he was very, he didn't really want to move his family. So he wanted to be in the Cambridge area, so that was quite a driver. The second thing is he didn't actually -- he doesn't actually want to be the person having to be the salesman and the justifying of the money, and all of that. He just wants to do the science far more. So he came with me, and he said, "I want to --" he was also an experimentist, "So I want to learn more bioinformatics. Could I use my redundancy money to work in your lab?" And I was like, [laughs]." So Ian had been the supervisor of my journal clubs, you know, when I was a Ph.D. student at Sanger. I mean, you know, he's 15 years my senior, he is extremely clever, wise. It was very odd to think that he -- So I said, "Sure, fine great, come along." [laughs]. "Feel free. Come play with ENCODE, you know, learn Perl, Do what, you know, fine for you to be there, yeah." And then when ENCODE 2 came around, actually we argued for deliberate funding of this analysis group. And I said to Ian, "Why don't you," you know, "why don't you do this? Why don't you be the person who makes all of this happen with me? We'll go and do it together. We've worked enough together." And he said yes, thank the Lord. So he and I, and it was far more him, organized the analysis meetings. Though on the phone, you know, it's, I'm the talker, chitchat, you know, all of that kind of things. So I'm definitely the front man for the Ian and Ewan show, as it were. Because our names are very close, lots of people can't even keep our names straight as well, so they would, they call me Ian and -- I forget. I mean, it's quite, you know, it's all sort of funny at some level. So we were doing quite well at doing this organization. Not perhaps as well as the 1000 Genomes project, but he did have more complex data. But then we had to run, and we really wanted to run everything through a standard pipeline with good Q.C. and everything else. And we just couldn't persuade people to do it in a way between the different groups. And there was this great graduate student that was Serafim's graduate student, that had somehow got attached to -- he was at Stanford, and I don't think Mike Snyder had moved to Stanford at that point, called Anshul. Yeah, that's right. It was -- he runs DNANexus, now. I can really visualize him. He is German. I'll have to find his name. Anyway, he was somehow between Anna and Serafim, Anshul was. And Angel basically put his hand up, and said, "Well, I'll run this, I can do this, I'll run all these things." And so Anshul created a standardized pipeline for the major chip-seq data sets. And coupled with that, is we had this legend of non-parametric statistics, called Peter Bickle.

And I'd met Peter in Oxford -- no, not Oxford, in Cambridge, when he was coming over from -- so he is a Cobb's Prize-winning statistician. He's, you know, statisticians don't quite bow to him, but you know, he is a major force in non-parametric statistics. And he's a charming guy, lovely, absolutely lovely. And when I first met him in Cambridge, he wanted -- it was around ENCODE 1, and he wanted to talk to me about this and that. And I wanted to -- and he had done something around genome heterogeneity. And you know, it was like getting -- he was very tolerant of me asking stupid questions. A stupid question was like, "I don't understand what these things mean. How does this thing --?" And so we brought in the Bickle Group as part of this analysis group, and that was really great. And then so the Bickle Group developed the statistics, Anshul made a pipeline, Ian corralled all the experimentalists with their metadata to stuff, I was the ringmaster on this analysis [inaudible], to make this work out. So I was -- yeah, so ENCODE 2 worked far better, from my perspective. And by this time I'd done a lot of genome papers after that, so I'd also kind of [laughs] really wanted this ENCODE 2 paper not to be the mess that the ENCODE 1 was. And so we structured it better. We built up to it better. We then went into this slightly surreal business, because it was very clear, we were presenting at meetings, you could see the data, that we would be making a publication. And so we sort of had both Nature and Science come to us to say that. And so we had a whole kind of, "Well, what can you do for us?" sort of thing. And basically Nature pulled out all the stops. And so we had, you know, epic amounts of print space, we had the main paper, we had the other papers. We got them to do all sorts of things, which I think was really good. So it's with some regret [laughs] that this whole 80 percent function thing kind of blew up.

So it was always a question inside of the consortium. So the definition was to find the function elements of the human genome. And so you end up sort of circling around the definition of function, yeah? And you can have a long, old debate about this. And the stuff that we measure, that you can measure things on the genome sort of doing, or being done to it, depends on your perspective. And pragmatically that was most of what ENCODE was. It was measurements of things on the genome, biochemical things. And then the question is whether you call that biochemical activity, and then reserve the word "function" for something else. Or you use the word "function" for this. And this debate went on and on and on inside of the consortium. And there were a group of people who were much happier with the word "biochemical activity," some of which have function. And there were a group of people who were very keen that had a too-controlling view of the word "function," so you need to talk to Tom Gingeras and John Stamatoyannopoulos, they are the -- when we observe biological, biochemical things happening on the genome, we can't just dismiss them because we don't understand them, is that perspective. And there's this other perspective, which is lots of things can happen, and they can happen thermodynamically and stuff like that. So you can't just say because things happen, they're important.

So this debate went on and on and on. And as we closed in to the paper, we actually had to use a phrase. And my regret -- so we did a phone call. I got everybody to discuss it. My regret was not to have a formal vote. Because a lot of people think -- one of the slightly alternative histories where I solely came up with the phrase "biochemical function," rather than it being a consortium thing, and it would be much better -- I mean, maybe we did record those phone calls. I wonder if we did. It'd be much, much better to have

About reproducibility of biochemical events, gets mapped into Peter Bickle's statistics, run through Anshul's, pipeline, and you then discover that an awful lot of the genome fulfill this criteria, and people shouldn't be -- and I -- and if -- when I started presenting this, I said, "Don't be surprised." You know, transcription from this perspective is, you know, is absolutely there as one of these reproducible events. And not only transcription as in making RNA, but the passage of the polymerase across the DNA and the histone modifications that get deposited as that happens. And there is something weird to say, "Oh, you know H3K36 isn't functional." You know, there's something -- there's something very odd about, you know, taking the very hard-core extreme, which is it's got to be on the evolutionary selection because there's a final group of people, yeah, which is going to be on the evolutionary selection. Yeah, there's a -- there's a -- sort of this complete sub-transition from these people. If it's not in the selection, yeah, it's not useful to -- if I can measure it, I don't want it off the table. And then there's a group of people that basically --there was a middle -- this middle group is sort of the Barbeau. This is the Chris Ponting -- obviously Dan Graur is like to the left of this group over here. Then there's this group, which is Barbara Wold and a bunch of people who accept that there's things that one -- that there's plenty more than what we can detect by pure selection measures but wants to have a very call it “high bar” for the use of the word "function.” And then you get to Tom Gingeras and John Stamatoyannopoulos who's like, you know, the mistake of molecular biology over the last thirty years has been to only talk about the things that we understand now. And they always would say microRNAs and these things and how if we hadn't been so closed-minded, you know, lots of things would have been discovered earlier. So yeah, the regret here somewhat is, in an early draft of the paper that I wrote, there's more ambiguity of the levels in the abstract and in the main thing. By the time we get to the end, we get less ambiguity, but with a kind of chain of definitions. And then we do have a section which goes through all of this and says, "Well, if you define it like this, it's this percentage. If you define it like this, it's this percentage. If you define it like this, it's this percentage." It's a march of the percentages, but you end up with 80 percent. So my, you know, out of all the things that I arranged for, or tried to make happen for ENCODE 2, where we got the QC a lot better, the experimental designs a lot better, we did this virtual machine of all our results and stuff like that. It's quite -- I find it quite frustrating that this fixation about this phrase and this number is the, you know, the biggest thing. And actually I don't think it's the biggest thing scientifically if you look at how it's being used. If you say it gets used routinely, fine. But if you look at the kind of next layer of why is it remembered, unfortunately, for better or for worse, this story is a big part of it. And that kind of pisses me off. And if I had my time again, I would have done my chess moves in a slightly different way. And, you know, maybe I would have insisted on a different phrase, which might have been "biochemical activity". At the very least, I would have got the whole consortium to vote very explicitly. So at the very least it would be very clear that we're all in it together. Because there was a period when Dan Graur kind of exploded, where a lot of people said, "Well, that's got nothing to do with me, Ewan wrote the paper." I'm like, oh, for God's sake, we all knew about this phrase for a long time. We can't walk away from it like that. But I'm in the -- I'm most comfortable with the use of the word "biochemical activity" for the broader thing, "a substantial amount of which we believe has cellular function" or something like that, would be -- would probably be my optimal phraseology, and I should find the early draft of the paper where there's -- where one tries to transmit more ambiguity of this percentage number. But I don't have sympathy with this hardcore end of people who think, you know, if it's -- these people, I think, are nuts. So the group of people who thinks that, I don't know, that everything that comes from a transposon is not relevant is -- I mean, they're a bit loopy, basically. And I'm sure they wouldn't -- I'm sure I'm characterizing their position a little bit. They wouldn't say that. But I think they have this very, very strong view that it's all about evolution. And as soon as you think about other things, like cancer, for example. Cancer would be a great example, so there's substantial deregulation of all sorts of different things. And of course you want to know whether MiCK binds here or not, even if it's not under selection. If the MiCK binding gives rise to cancerous cells, yeah. You know, you really want to know it. Now, do I call it function or not? I don't know. Do I want to know where it is, understand it, measure it, characterize it? For sure. Yeah? If we just leave aside the function of it. So yeah, have I learned? Yeah. And then it was -- I mean it was -- I find it personally very, very difficult when Dan Graur and some people really laid into me on the Internet. And this -- and some people still do. And I've, you know, I thought I had thick skin, but I didn't. Now I have thick skin because now I look at some things that are written about me and I'm like, "Well, clearly you don't know me. You've kind of got this voodoo doll person, which is Ewan Birney, and you enjoy sticking pins into him, you know, fine. You know, I'm not even going to try and debate because what on earth is the point? To persuade you that that was, you know, you're so locked into a mindset where I'm the anti-Christ. I'm not -- I'm just not going to attempt to untangle it for you." And it has given me insight when people attack other scientists passionately, or politicians. It's given to me an insight that, you know, most of the time, it's very unlikely that people really hold the extreme positions that they're said to have done by other people. It's -- especially if people are clever -- I mean on television. They won't -- you know, I think people are -- when people sort of -- this process of creating these sort of caricatures of people -- it's much more about how people want to frame the debate than really about, you know, understanding what is going on. And yeah, but as you can see, I have a lot of kind of regret because I feel I did lots of things right, in ENCODE 2. And it's my best -- my best thing, my best-run consortium, and it's sort of marred, has this really bad end-of-life taste right at the end because of this.

The interesting thing about Dan Graur is that -- so Des Higgins, who used to be at the EBI was really well, not a friend, but early Dan was an early phylogeneticist tree-builder, all of this sort of thing, which is why he's into this. And he -- and Des knows him well enough to have beers when they're over there on a conference together. So it was quite interesting to talk to Des about Dan and, you know, Dan clearly holds hard positions and, I mean, I kind of -- I understand the passion that scientists bring to intellectual purity of all sorts of different things. And so there's a part of me that thinks, "You know what, if I just have a beer with Dan Graur, you know, we'll leave friends. You know, if he can have beers with Des Higgins, you know, there's a kind of transitive beer-having process that says that this should work out." And I, you know, there's part of me that's tempted and then I just think, oh come on, your life is too short. And he has made such a thing with me on the other side that I just don't think he could come to like me. So I have kind of mentally said to myself, "You've just got to forget about it. You just -- you know I can't make everybody like me," which is -- if I have -- one of my failings is I kind of want everybody to get along and everybody to be happy and stuff like that. And it's regrettable. But yes, I've said this you know, Twitter and blogs and stuff like that. We don't on-line criticism -- we don't have constructive online criticism - it's hard. And you know, it degenerates very, very quickly into extreme positions and I've actually tried to make sure that I am not someone who feeds that. But it's quite interesting because sometimes you can get -- I mean -- this whole problem with email, which is, you know, never write an angry email until you've at least slept on it. But Twitter, you know, moves that ability to go from anger to tweet in a very quick thing.

And also the other thing I think -- which perhaps I didn't -- my antennae were tuned in a slightly different way was that of course over in the U.S. the junk DNA thing is wrapped up in the creationism debate by intelligent design, all of that. And I actually think that has made some of the people who are the kind of adamant -- does lots of junk DNA people. They close their eyes to data. They can't, I mean, it's slightly weird. I've had some of these interchanges. And of course, from my perspective, you know, creationism is Class A bonkers. I mean, you know, it's in the -- it's in the non-science zone. Yes, we've got to work out how to educate these people, but it's a completely different class of discussion from this discussion, which is you know, "I've seen this. I've done this. How has it happened that these things have emerged over evolution?" But for some people, they see every -- they see the support, or the fact that creationists use these statements as indication that they've got to attack science. And I was actually reassured -- you know, there are two key members because people were kind of laying in Twitter, blogs, and Facebook as well. So there's a whole Facebook kind of thing. And there you're -- it's a web of friends kind of process, yes, what have you. So somebody was really going at it with Mike Eisen and stuff like that. And then I -- Mike was pretty -- Mike was annoyed that the consortium had so much Nature publicity. That was Mike's main beef. But other people were annoyed about this other thing. And so I joined in that Facebook thing and said, you know, just to make sure, you know, this is the part in the paper, whatever whatever, and I said something which was, I don't know, something about how -- you know, this is being quite personally difficult for me to hear so many people who don't know me really vilify me. And I give Mike quite a lot of credit that both in that Facebook thing and in Twitter he said, "Look, you know, I don't like ENCODE for this, I don't like IT for that, I don't like him for the other, but Ewan Birney is neither an idiot nor should he get, kind of, you know, whacked around by this." And that was nice of him, actually. Well, it's not nice of him. That shows -- you know, that's the right attitude to have. And I have got a lot of time for Mike. Not everybody has time for Mike. I'm sure Francis Collins, Francis doesn't have time for Mike. But I think that fundamentally he's got -- he's got a strong, you know, he matches his passion with the human side, so it's a good thing, you know.