Protocol 1277 Informed consent statement for Oral History Interviews

(This form can be sent in advance and signed or read into the tape at be beginning of the interview.)

The interview will be recorded, and I will use the audio file to make a transcript. The transcript will be shared with you, with an opportunity to correct it. The attached form indicates options for making the final edited transcript available.

My name is Fred Heller and I am a student at Duke University. I am in a course on the history of genomics that includes oral history. One goal is to produce a written transcript of interviews with important figures in genomics. Some of the interviews may be archived or made public through a website. The conditions for making the transcripts public (the audio tapes will not be public) are indicated in the accompanying form, and you can choose any of those options, or write in your own conditions.

I selected you as the person I would like to interview. The interview should last 30-45 minutes. Your participation in this interview is strictly voluntary, and you may withdraw at any time. You do not have to answer every question asked. The information that you choose to share publicly will be "on the record" and may be attributed to you, unless use is restricted the conditions you specify on the form.

This interview is being recorded and I may take notes during the interview. The interviews that are posted publicly will be archived as a history resource. If you prefer that the interview be used only for the course and not made public, please indicate this on the form.

One risk of this study is that you may disclose information that later could be requested for legal proceedings. Or you may say something that embarrasses you or offends someone else when they read it on a public website. The benefit of participating in this study is ensuring that your side of the story is properly portrayed in the history of genomics.

Signed: _____ Date: _8 Nov 2012____

Person interviewed: _Christian Burks__ Student Interviewer ____Fred Heller____
(Print clearly) (Print clearly)

Use of archived final transcript

Members of the Duke University community, students, faculty and staff at other institutions, or members of the general public may access the digital archives. Typical research uses of interview materials include scholarly or other publications, presentations, exhibits, class projects, or websites. However there may be other uses made as well, since the materials will be available to the general public. Investigative reporters and lawyers engaged in or contemplating litigation have, for example, used the Human Genome Archive.

Your permission to post the edited, written transcript of your interview, and any related documents, to a digital archive is completely voluntary. Unless you consent to their wider use, all materials from your interview will be available only to members of the research team affiliated with this project.

The form below provides you with different options for how, when, and with whom your interview materials will be shared.

(A) __X___ I place **no restrictions** on my interview materials.

OR

(B) ___My interview materials may be reviewed, used, and quoted by students and researchers affiliated with Duke University; *and in addition* (check all that apply):

___Researchers unaffiliated with the Center for Public Genomics may **read** the interview transcript and any related documents only after obtaining my permission.

___Researchers unaffiliated with the Center for Public Genomics may **quote** from the interview only after obtaining my permission.

___Researchers unaffiliated with the Center for Public Genomics **DO NOT HAVE** my permission to **read or quote** from the interview.

Posting interview materials to public digital archives: In spite of any restrictions listed above, I give permission for my interview materials to be made publicly available on the Internet by deposit in an institutionally affiliated archive:

___ 1 year from the date of this form

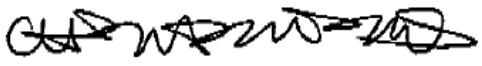
___ 5 years from the date of this form

___10 years from the date of this form

___25 years from the date of this form

___After my death

___Other: _____ (please specify a date or condition)



Signature: _____ Date: ____8 Nov 2012___

GenBank Interview

Questions provided before the interview:

Why do you think GenBank happened when and where it did?

Why did Los Alamos get the contract? Why did the specific contracts with BBN and, later, IntelliGenetics, win out over the competition?

How important was it that Margaret Dayhoff charged for and put restrictions on use and sharing of her Atlas? How did Walter Goad's "open science" ideas play out in practice at GenBank?

Biggest challenges in founding GenBank?

What did you bring to the Los Alamos Team beyond being a "card carrying biologist?" Or even, what did being that "card-carrying biologist" mean to the Los Alamos team?

Does you see GenBank evolving further? What do you think it's role will be into the future?

What are you doing now? How do you see GenBank as a part of your trajectory as a scientist?

Do you think there are lessons from GenBank that are relevant to today's debates about access to massive genomic databases?

The interview
Normal Font= Christian Burks, Bolded Font= Fred Heller (interviewer)

So I guess your first question is why do you think GenBank happened when and where it did?

Yes

When it happened, I think really just two things pretty simply. One is that DNA sequencing in the laboratory really took off. I think the first sequence dates back to 1965 or so of a nucleic acid, but by the early to mid-seventies, methods had been developed that allowed people to sequence DNA much more rapidly- nothing like today's rates but, at the time, very, very rapidly. The community came to realize two things- the scientific community that was interested in analyzing and using the data realized that they were very rapidly and had already gone past their ability to keep track of all of the sequences by reprints and file folders and stacks of papers on the desks. And secondly, that the nature of sequence data made the use of computers to store them, retrieve them, and in particular to analyze them- computing was just going to be the way to go, plain and simple. And so the combination of wanting to be able to use computers on sequence data and two, needing to come up with a much more centralized and broad scale approach to accessing all the new sequences that

were being created really drove the community to figure out what they wanted to do about this. There were a series of workshops funded by various government agencies like the NIH, for example, that pulled together scientists in the community who were interested in this question and asked what should be done about it. And the answer to that was that the government agencies in the world- I didn't say in the US- NSF, NIH, DOE and others came together on the notion of providing support for a centralized computational database resource for the community to access.

Feel free to ask if you want to tease out any other details or specific aspects, but that's really why things got going when they did, which I guess would have been by… around 1981 or so, the NIH put out a call for a contract to run a sequence database- it wasn't called GenBank then- but they wanted to seek proposals from groups putting this out.

Was there at all an international competitive thing or cooperative thing going on?

There was international interest and there were a handful of groups like the one in Los Alamos that I ended up joining that were collecting data and sort of diving in and building up a database. As it ended up, several projects got funded. The US funded the Gen Bank initiative with Los Alamos and initially Bo Berwick and BBN as a partner. The European community funded the EMBL data library at the EMBL labs in Heidelberg. And then several years later Japan funded the DNA database of Japan or DDBJ. I think the idea was that, in principle, a single database would have been sufficient, but there was a notion that from early on that those communities wanted to have their own take on, their own control of, and their own visibility around managing what was going to be an incredibly important resource not only for academic research but for the biotech community as well. And so those 3 separate projects were funded, but with a strong mandate from the funders internationally and, at least equally importantly, a strong enthusiasm from the various teams and staffs to work together closely and to wherever possible divide and conquer. We tried to avoid, for example, all 3 groups doing initial entry of the same data sets. We divided up the world and one group collected one set of data and then [shared] it with the other groups.

OK- You want to move on? Whichever question- Why did Los Alamos get the contract?

Ok. So why did Los Alamos get the contract and I guess I'll get to the second part- why the specific contracts with BBN and later Intelligenetics. I'd say at Los Alamos there was a group that was called T-10 for short, but was theoretical biology, biophysics, that under Walter Goad's leadership. With help from Minoru Kanehisa, a post doc from Japan, and Jim Fickett, the US post doc, they'd already established what they called the Los Alamos sequence library. So, that was something dating back several years from then and this call for proposals kind of put Los Alamos on the map and already gave the group a significant amount of momentum. I think that obviously was a key component in getting the contract. They were already doing

very much what NIH wanted to see happen. They weren't the only group doing that but they were one of the few globally that were doing it. Not only were they collecting the data, but they were distributing it. And that ended up being, certainly not the only reason, but considered a very important reason why they got the project. They had almost an ironic, say disposition, towards the procession of Los Alamos as a weapons lab and [being] very concerned with [being] top secret in relation to things. Walter and our team's approach was incredibly what today would be called open access around their data, their tools. Anybody requesting the data sets they would provide it [to] in whatever form they could. They published the database as they built it up so that people could know what was there and use it for their research. I think that ended up being a significant part of our proposal to the NIH and I think something that the community that was reviewing the proposals when helping NIH to decide what to do really put a lot of emphasis on. I think that probably was the most important factor. I think another factor that was strong was that they already had a strong interdisciplinary team in place and there was the general perception of being in a very strong computing and communications context/framework at Los Alamos. Los Alamos was fond of saying in those days…what was the phrase? Something along the lines of "the most intense computing center in the free world." They had tons of mainframe computers, or what we'd call supercomputers, that were there because primarily for being able to support the weapons program- the design of nuclear weapons and analysis of all the related science. But, that we were in an environment that was used to large scale computing was a plus.

So I think that's enough on that. The fact that in the first go around BBN was our partner and in the second Intelligentics came out a little bit indirectly in sense that [Inosha] had been aware of Los Alamos's interest in this and, in fact, had provided some preliminary support while they got their workshops going and got sorted out what their proposals would look like. There was certainly an anticipation that Los Alamos would be bidding for the contract and would be a frontrunner applicant for it.

In a very short period just before the proposals were all due, DOE, which provided oversight and under whose wing Los Alamos ran, provided the guidance that they wouldn't allow us to bid for a government contract. Though officially we were UC California employees, because the University of California ran the project for the Department of Energy, the Department of Energy would not sign off on our contract bid and they said it was inappropriate from their perspective for one government agency to be bidding for something from another government agency. That created a bit of a rapid readjustment of how we were going to be doing this, because we basically needed someone outside the US government to partner with us in our proposal, in which case DOE was fine with us submitting a contract or two. For instance, a private sector applicant, but that'd come from outside.

The second thing DOE put down as a mandate was that we couldn't show favoritism to any particular external applicant. We couldn't go out and say, good news, from our perspective this would be the best partner and we're going to put in one proposal with them. We had to agree to put in proposals to basically anyone that was qualified and wanted to bid for the project. So in both competitions we

ended up being listed as partner for more than one proposal and with more than one private sector account.

Were you gunning for a specific private sector applicant, or did it not really matter to you guys which company?

Oh I think it mattered but we were…I was there for the second round of proposal; we took very seriously the spirit of that constraint. We did our best to provide the same ideas, the same written input, the same "let's put our heads together and figure out how we're going to solve this particular problem or this issue" with all involved, which was a little frustrating for our partners, because they're all looking for an edge on their proposals, but that, eventually, I think, was provided through the particular functionality and value that they were providing in their proposals. I think BBN, certainly the first time around, had an established history in helping to get ARPANET, one of the internet precursors, going and in providing access to NIH and support for NIH access to, internet resources and for the community that wanted to access NIH resources interact with NIH resources. They had been doing some contract work for the NIH already so they were a known entity and had sort of established themselves as sort of at the life science and Internet interface. I think that would've been a strong component of their success on their first bid, with us as their partner.

And I think the second time around Intelligentics just put a lot of thought into, and provided some attractive perspectives of what they could provide in terms of developing interfaces for the community, both for accessing the data but in some cases for submitting data to the database. I think they had positive suggestions to make in the proposals about the underlying technology they'd be providing that made that proposal attractive. But at the end of the day, I wasn't on the review panel. I don't know in detail why one proposal won out over the other.

OK.

So I think I'm ready to go onto the next question.

Go for it.

So there's the question how important was it that Margaret Dayhoff charged for and put restrictions on the use of her Atlas and how did Walter Goad's open science ideas play out in practice at GenBank. And I think I answered that-

Yeah, you were talking along the same lines

I think- I'll put it this way. I think that that certainly was one of the most important elements. From a history of science perspective I think that's… how do I put it… I think that's been cultivated and has drawn a lot of attention and been very interesting to people, both as an early precursor of many of the debates over everything from patenting sequence data to open access science and so on. But also I

think that, because of the ironical component of this most open access approach coming out of a government weapons laboratory, that there's a nice resonance with the original juxtaposition of J Robert Oppenheimer, the original director of the lab, and Leslie Gross, the military guy whose overseeing it there, their wrestling back and forth on whether or not and how open information should be shared within the laboratory and between the lab and the outside world. Just as an interesting reflection on that, maybe even more historical context, it keys into this approach with Walter Goad. He came out of the weapons program- he was at Los Alamos for a couple of decades at least, coming out of grad school and staying there as a theoretical physicist. He was subject to, as were we all- those of us who came in to work in T-10, were largely working on sort of open publications and grants from the NIH research projects, but we all had to get top secret clearance in order to work there, because that was just a fundamental policy of the lab, which sounds very restrictive, but that actually was something that was instituted early on because they wanted everyone in the laboratory, regardless of what they were working on to be able to freely exchange information. There was a philosophy started with Oppenheimer and certainly reflected in Walter Goad's insight and leadership there that spoke to the notion that science went faster and got done better when information was freely shared.

OK, that makes sense.

Let's see. I guess there's the question how did this play out in practice at GenBank. In other words, it was great to talk about that, but did we really behave that way? I'd say plain and simple yes. Certainly I was there from an early stage- I joined the group as a post-doc before our first proposal went into NIH and so was involved in both in proposing and then in implementing the project once we were funded by the NIH. It was completely true through and through, from the start. There was never any sense of the data being owned by or for preferential use by us or Los Alamos or any group out there. We operated as if, and we took very seriously the notion that, we were caretaking these data for the community- from the community for the community- and I think that approach, the open science approach, did play out for us.

Do you think if you hadn't taken that approach that Los Alamos wouldn't have gotten the contract in the first place?

I think that had a big factor, maybe the prevailing factor, in that decision on the part of the NIH and its review panel in awarding the first contract, yes.

Ok, that's interesting. Biggest challenges in the founding of GenBank? Or any big challenges you had in your experience at GenBank- let's broaden that.

Sure, Sure. I'd say there were several. One, plain and simple, was the exponential growth of sequence data. If the community was concerned enough to start to talk about starting a database in the mid-seventies, by the time it got started, it had

established and has maintained- the scientific world has established and maintained to this day- a doubling rate of the data that is somewhere between every one and two years. There's twice as much sequence data as there was before. So, that means that you're always looking at what looks like an avalanche coming at you in terms of the quantity of data that we made available. So that was a significant challenge. The fact that data entry initially was paper based, and post publication, and- AND- relied on a sort of hunt and forage approach through the world literature on the part of the database staff made it incredibly challenging, and in fact impossible to keep up with the literature of the labs generating the data. When things first got going, the modus operandi was a scientific group would submit a paper to a journal with a new gene sequence and a journal would go through a peer review process, decide whether or not they're publishing the paper. Once the paper was published in hard copy form, it would go out to libraries worldwide and we, of course, tried to subscribe to as many journals as were relevant, but of course, as you'd guess, we couldn't subscribe to every journal in the world that might publish a DNA sequence. So we had this model where we had to go find the papers, wherever they were. We then had to enter the sequence data through data entry back into the computer so we could put it into the database, and all this was happening after publication. So you have this incredible mismatch: the day the journal is published and lands in the library, you have hundreds, if not thousands of scientists in the world who would liked to turnaround- and the functionality of the database was imagined around the notion that they could turn around- and access the data from the database. But of course, we might not be seeing the data at the same time as them, or even slightly after. And then we had a process to get it in, and we were backed up. It became clear relatively quickly that we were underfunded, under resourced for the quantity of work that was involved. We had a tremendous backlog. I think another thing that wouldn't have been guessed at but certainly was true was the computation, the computing technology and software and operating systems were all not a great match to large scale text computing. All that mainframe stuff that we bragged about and which people were excited about having- the database already at Los Alamos was very oriented towards number crunching and doing large-scale mathematical computation- the traditional reasons that people went off to use computers in the first place. And even when you looked at things like database technology that was out then, the large commercial systems and, for the most part, any systems that were available were very number oriented. You either, [used them] for traditional scientific computing purposes, or for support of number crunching in the financial world. It actually was quite difficult for us to figure out what, in some ways, could be seen as a humongous word processing project. Getting all these strings of characters into the database. Being able to manipulate them, store them, retrieve them, analyze them, but it was character strings. And, from the way things were defined back then, infinitely long fields. As a result, we were always probing for and working with vendors to get technology on board that would work and that was a better match to the task at hand. So we pretty quickly moved off of the mainframes and the patch-job environment that existed and using the mainframe that existed at Los Alamos, and into personal computing that we could oversee and use ourselves. We shifted from the supercomputer operating system to DOS operating system, and pretty soon

to Unix operating system to support the kind of analysis and to be able to build our own software tools for analyzing the data. Lastly, and most importantly perhaps, we shifted from paper to people submitting sequences on digital media to finally use of the Internet for getting data.

You got them to start submitting before publication too?

Well I would say the other big thing, and probably the biggest paradigm shift that we introduced that I think has had a major impact on the world, well beyond just DNA sequences and GenBank, was shifting to electronic submission prior to publication. And in particular, we were very lucky to have the engagement and leadership of a number of journals and their editorial leadership to shift to a model where they were requiring evidence that people had submitted data to GenBank before they would agree to publish the data. Very few of them were… not wanting to in a position of trying to police all of that, but just the fact that they stated that as their policy, that they had the community's attention at that most crucial point when you want to get a paper published. And if a journal says you need to dot your I's with blue ink, you tend to dot your I's with blue ink. Because right on the other side of that silly little request is the possibility of having a paper published and so I think that ended up really being a key turnaround.

At the end of the first 5-year contract the project and its oversight were really in turmoil because we really were quite far behind. I don't remember now, but pretty close to two years, on average; a paper would take maybe one and half to two years to get into the database after it was published. We and everyone agreed that that was not at all what the database was meant to be and what it needed to be and so, in the second contract we put in, we asked for roughly tenfold more resources, and got tenfold more resources, and so we actually had the wherewithal to tackle the problem at hand. And two, we went out and began working with the journals and the scientific community to really shift to this notion that it wasn't up to us to go and find papers and enter the data in; it was up to the community to take responsibility for sending us the data in electronic form prior to publication. The combination of the resources and that model, the shift to a model of the database role as one of really providing highly automated archiving, really ended up turning things around, and so by the end of the second 5-year contract we ended up for the most part- most of the data were getting into the database and being held until date of publication if the author requested it, but then almost instantaneously [being released] after publication. For the most part we had managed not only to catch up on our backlog but were keeping up with the ever-increasing literature. Which is really essential, because if you've got responsibility for processing something that's growing exponentially, and you don't have a model that allows you to grow your staff exponentially, which of course you don't, then you know you're doomed to failure somewhere down the road. I mean, you might start out with enough staff to keep up with exponential growth but at some point the lines are going to cross. We divorced ourselves from that- the burden of that exponential work from us and shifted it back to the community. If there is going to be twice as many scientists generating twice as much data or the same number of scientists generating ever

larger data sets, it was up to them individually, as a part of what they do day to day in analyzing and publishing their data to get the data to us. That was a key element of that success. And it was a very important success, both for us- the taking pride in our responsibility for the database, but very important for the community, to see that the database was a timely and up to date resource that it could draw on.

Let's see…What did I bring to the Los Alamos team? What did I bring to the Los Alamos team beyond being my, quote-unquote, "card-carrying biologist." I think I say that humorously. I was, I think at the time, the first PhD biologist to join the team. The other people there obviously, or maybe not obviously… but it's certainly true that they were very intelligent and very aware of what was going on in molecular biology. It wasn't that there was no on there who didn't understand what DNA, genes, and cells were, by any means. They, I would say, knew as much as I did about that. But I'd come out of the framework of molecular biology, lab work, use of sequence data and just that whole perspective of a biologist in an academic environment trying to crank out research results, get them published, relate to the world, relate to journalists, relate to readers, and so I think that perspective was valuable, really. And then also just probably also valuable in just the marquee of being able to just say we've got one of those guys on board as being part of our team.

I think that was certainly a key part of what I brought to the table. Further that- what was I able to contribute or why was it helpful to have me there- I think I had certainly been fascinated at the interface between molecular biology and computing as I went through my graduate work- which started out as very much a venture in experimental project- I found myself gravitating towards the computational part of that: we need analyze some data, we need to build up a data set, we need to write some software. Whereas some people when confronted with that would scurry back to the lab- what for them was more interesting elements of designing and carrying out experiments, and I found myself really drawn to the computational aspect. And so I was definitely looking for post-docs that would carry me in that direction and so being enthusiastic about that and being interested in the interface between the lab and computational approaches was good to have on the board for Los Alamos. I think something that was- and again, I wouldn't necessarily have known it myself in going there, and I think it's turned out to be evident in looking back, is that I really enjoyed the idea of designing and supporting and building broad scale resources. That's pretty unusual. Or at least sort of counter to the typical academic approach which will build up resources, toolsets, reagents, enough to publish the paper that establishes you learned something about nature or you proved a particular hypothesis was tenable, but then probably the worst thing in the world is to have that around your neck like an albatross, with everyone in the world saying "Oh that's great. Could you please-" having a hundred correspondents saying after you publish, "could you please provide me that plasmid," or "could you get that software to work on my machine," or "I need this from your lab." That's not a high priority for groups typically, because that typically is not helping them get further down the road in terms of leading edge research. Being fascinated with the idea of what does it mean to provide a resource to the world and the real value proposition of being able to say, if, in doing a very highly centralized and well put

together resource we can make thousands or hundreds of thousand of scientists in the community able to do their research faster, then that's a good thing to work on. That certainly was important.

The final thing for me was, which again has just sort of been reflected in my career, was an enthusiasm for and an interest and fascination with start-up and the culture and climate of startups. Even though GenBank was embedded in Los Alamos National Laboratory, which had been around for decades and had six or seven thousand employees, and so by most measures would be considered a big institution, theoretical biology, informatics, and the GenBank database itself embedded in there were very much in start-up mode. That is, we were scrambling to get funding, we didn't really know how to do what we were doing, we didn't even know what the end product would be- you know, we had a general notion, the database would be a resource for DNA sequences, but what did that look like, what's the right way to do that, what's the right computing technology to use, what people are going to be good at this, what's the model look like? All of that was figured out on the fly. And figured out on the fly in a context where you already had a very demanding user expectation for delivering something useful. That sort of "bring order out of chaos," and, I mean this sounds incredibly boring from a scientific perspective, providing a really solid operational framework and a sort of strategic stepwise plan on how you're going to get where you want to go, I think ended up being incredibly important for the project, something that I was good at and happy to dive in and help with when I was there. Just to give one example, at the end of the first five years, when we were one or two years behind the literature and we successfully asked for additional resources and we said we had a plan on how we were going to catch up to the literature, that- everyone involved was pretty, frankly, skeptical about our ability to do that because we were so far behind and everyone was just incredibly upset and just really worried about how could they possibly not only catch up, but keep up with the ever increasing literature. It would've been easy to not succeed at that, not because we didn't have the right ideas but because operationally it's pretty complex and you've got a moving target and you're also delivering a product every single month to the community. Your notion of "we've got a great idea how we can solve this better or increase our throughput twofold;" you don't have the luxury of saying "we're going to take a time out from delivering data to the world while we explore this and validate it and QC our tools and then open our doors again in 12 months." We had to do it all at the same time as we were meeting the demand for the product that we were providing- the database. It took a really strong operational imprint on how we were going to approach that and be organized for it, and that certainly was something I enjoyed helping out with.

When you were two years behind, was the scientific community at all upset with GenBank or disgruntled?

Oh, some were upset and outraged (chuckling).

Or were most understanding? What was the general…?

I'd say certainly people who were close and the advisors of the database- the NIH had an advisory panel which were stellar people from the community who understood the need for and, to some extent, the underlying technology providing resources and all that. Where good representatives of the community thirst for a solution to the, having a resource to get its sequence data, not ev… they were certainly understanding, they were aware of the limitations we were working under. They were very supportive when we switched into this mode of, in particular, wanting journal editors- there were a couple of journal editors on the NIH advisory board and they jumped right in and decided to use their journals as bully pulpits for the rest of the community, around which there was a lot of reluctance. We spent a lot of time pushing the idea of journal involvement in all of this. It wasn't a task the journals wanted to take on. They were worried about getting on the wrong side of the authors that they were eager to see submit papers to their journals. Of course, many of the journals were run by and staffed by members of the scientific community, or who had been members of the scientific community, so they weren't very sympathetic to anyone stepping in and telling them they had to do something.

Was it that much an imposition to ask? It doesn't seem like it would be that difficult to do, just from a modern perspective, to hand in the sequence data before as opposed to [after]. Was it that big a paradigm shift? Was it that difficult?

Part of it was simply being told that they had to do something. I think that was a point of concern- are we opening the door to further incursions on our liberty to publish the data as we see fit or to decide when and where. There was also the competitive interest that many labs have, which is to delay other people's access to their data. I mean, on the one hadn't they're all open to and they understand that the way that the community moves ahead is by sharing data, but there's been a long tradition in academia of publishing your results and deciding when and where you enable the rest of the community to race ahead of you by giving them your reagents and tools and detailed knowhow on experimental methodology. In a paper you publish enough information so that peer reviewers can indicate whether or not they believe that result and think that you've done sufficient work to justify the conclusions you've reached. That's very different than being able to say, "I can do that experiment in my lab tomorrow." I think the community was concerned about the underlying data being plucked out of their hands. And this only increased with time, as it got easier to do much more sequence data, in that the labs doing ever longer regions of DNA thought that they had a gold mine on their hands and that it would just take a long time to go through there and do all the experimental work that would give evidence for whatever they thought they saw in the sequence data that was interesting. They're very nervous about putting the data out there initially. The community overcame that. I think it just got to the point where the community- it's never let go of it's competitive interest in not being beaten by other labs with their own resources, but it's certainly come around to being culturally the default to put your data out there as soon as possible so that the community can run with it.

I think one big concern that I think was actually a valid concern to ask was people were concerned about submitting data prior to all of the peer review, worried about the notion that we were going to the model where we were saying once we got the data, it just goes into the database plain and simple. There was a very valid concern raised and a lot of debate around what does it mean. Originally people thought, well, one, sequence data will only be submitted to the database after publication and that means that the sequence data had been subjected to peer review and all of the corrections and curatorial improvements that come out of the peer review process. So if we're asking for data before all of that happens, when the manuscript is submitted to the journal or even later on in the process before it's published, people were concerned that there might be data appearing that was lower quality. I think, second, people were hoping and wanting it to be the case that we did a lot of curatorial work. The notion was that we go out and find those papers, we'd read the papers, understand the underlying biology and the experimental framework- you know, what hypothesis was being tested, and in entering the data we'd regularize nomenclature, we'd regularize the annotation, we'd notice that in the figure where they reported that sequence they'd mislabeled the start codon for the starting sequence- all that kind of stuff- and that we'd either correct it or we'd correspond with the original authors and correct it based on their input. And it was very clear to us when we were proposing all this that anything like that that we can do computationally, we'll do. If an author sends us a sequence and says that position 12 to position 999 is a protein coding sequence, we can put that sequence through a bit of analysis and tell you whether there are any stop codons in the middle of that region, and if there are that would raise a flag and we'd automatically be able to bounce back to the author that observation doesn't agree with what you said about the sequence. But if it couldn't be done computationally, we were trying to get humans out of the loop, at least humans on our staff. Again, the community was concerned about junk- not what biologists call junk DNA, but illegitimate DNA getting into the database, or poorly characterized and ill reviewed data getting into the database.

I think that particular concern was, I think, kind of overcome and nibbled away at by a couple of things. One is, we had so many instances where data that had gone through the peer review process were still problematic. It was pretty clear that peer review at the best journals was really focused on the scientific context and the conceptual framework for the data and the underlying technology, but peer reviewers weren't going through and typing in the sequences and analyzing them and figuring out what was really there and correcting authors on that basis. You'd have instances- there were not many of these, but there were instances- where a paper would come out in a leading journal, had gone through peer review, would get into the database, and we or someone else in the community would say that really exciting new hormone receptor gene from a human you published is actually nine tenths bacterial vector. It may well have been the case that, while the experiment was correct, a secretary entering a file into a figure prior to publication had put in the wrong file folder. There were cases where actually the whole experimental framework had to be reevaluated and withdrawn because it turned out that they had, in the early days, missed the opportunity to compare their sequence to a library

of bacterial sequences and realized they somehow or another ended up with a lot of bacterial DNA in the clone that they were analyzing. So, on the one had we were able to say that peer review was not really delivering pristine, checked-in-every-way-you-could-imagine data to the database, so in getting data prior to peer review we don't think we're giving up a whole lot for just the sequence itself. And if someone told us, "well we submitted this paper and please hold the sequence data until publication, don't make it public in the database" and six months later they sent us a note saying well we decided not to publish that data or the journal never accepted it, we don't put the sequence data in unless the author said go ahead and publish it even though we didn't get a paper out of it. I think on the one hand that concern was not so significant. The concern about whether or not people were going to get scooped- basically, as funding agencies came in and said "get over it guys, we're paying for this, we need you to make the data available," people overcame those concerns. And then there was a highly cooperative effect. Once a few journals dived in and said, "this is important, we're going to do it," those concerns went away also. There was just a big, I would say highly cooperative transition from a lot of people and stakeholders in the community, saying "that's not right to submit data ahead of time, it's going to create a really crappy kind of resource, etc." to people saying "It makes all sense, it's the best thing in the world." And I think the biggest pressure or leverage we had all along was, "do you want the data or not?" If you want to be able to go to the library or go to the mail and see that a competitor of yours published a new gene sequence a hundred thousand bases long of something or another, if you want to be able to get it out of GenBank at that moment, you have to agree to all these other things. And that was the carrot, I think, that really drove people. At the end of the day, they wanted the data as soon as possible as a resource.

Any of these last questions? GenBank evolving further? I kind of had asked that with a view towards synthetic genomics and if you think it will encompass any of that information, but however you want to answer that is fine.

I think that the two ways that GenBank has evolved and, I think, will continue to evolve is, one, as new approaches to sequencing and sequencing strategy and technology have evolved, it's pressured and burdened the data, the quote unquote "GenBank database" in different ways. And so EST sequences were something a few years back where that notion that we're going to work 5 years on characterizing this one gene and what it stands for and everything we know about it... When ESTs came along there were these people submitting all of these fragment that didn't have near the amount of experimental verification of what they were and what they might mean. And also just tons more data. When people went to metagenomic sequencing, whole genome sequencing, DNA barcoding, all of these created pressures on the database staff and the community to understand "are these the same kind of sequence data?" We used to say it had to have these three data items associated with it- like a species name. It has to have a species name attached to it before you put it in the database. But then when you go to metagenomic sequencing, where you don't necessarily know what organisms the DNA came from- "I sequenced everything in this tablespoon of soil" means you're creating data that the world

could well be interested in and would want to take advantage of and would want to have access to, but you can't tell necessarily with any certainty what species it came from. And that's just one example.

I think the other big thing, and this came with the initial realization that our primary focus for the main database was to be an archive and a sort of timely and rapidly available archive for the original data, was that the database has been a great platform to build all kinds of very scholarly, well-wrought, well-curated, well-researched specialized collections of sequence data on top of. It's a great discovery resource. If it's very important for you to know that you had every HIV sequence in the world, an initial pass through GenBank, you'd probably bump into- like if you search with one HIV sequence, you'd probably find 99% of all the known HIV sequences out there. But, they wouldn't be organized, or named, or curated in a way that would make them very useful for you as a study platform. I use that as an example; when one of these very specialized databases that came out- actually from Los Alamos, run by a guy named Gerry Myers, originally, and later by Bette Korber, was focused on HIV sequences. There was tremendous interest in the community, of course, in trying to figure out why is it so hard to target this virus and build up vaccines against it. What is so special, in biological senses, about its ability to mutate away from any net we try to put around it? A very focused database, very well curated, very scholarly- everything that was said in it was incredibly useful to that community. But that was a separate database, drawn from GenBank but not part of the GenBank project itself. I think that kind of thing will continue to happen as the database goes forward.

I don't know if you want me to say anything about what am I am doing now or lessons from GenBank?

I mean, if you have the time and you have anything you want to say, please go.

I've got a few minutes, I guess. In terms of what am I doing now, or what I've been doing since working on the GenBank project, I say that the two things I mention being especially rewarding about GenBank, one was being involved in a startup and the dynamic of a startup and trying to get a startup out of the more chaotic mode and into the water mode has been sort of repeated as a fun and good process and [I've been] involved in several startups, more on the biotech side, in the commercial sector since I left Los Alamos. That experience and that realization that I really enjoyed that [start-ups] has certainly been reflected in the kinds of projects I've taken on and companies I've worked with over the last twenty years. I think the other thing, which was working on something that really made a difference, that can't always be true and sometimes you think it's going to be true and it's not (chuckles). I think [of it] as sort of as a guideline, for any time in my life where I've been saying "What do I do next?" I've really been interested and focused on whether it's a project or a company or something along the lines, what is going to make the- or at least what has the potential to make a big difference in the world? I had that first experience in GenBank and it's one that's carried over nicely.

In terms of the last question there- "Do you think that there are lessons from GenBank that are relevant to today's debates about access to massive genomics

databases?," that seemed a little open ended, so I didn't have a pat response to that. I think it's certainly true that GenBank, I think, was really, certainly in the life science community, was the first that I'm aware of to really overtly and explicitly breach that "wait till it's published" barrier. And in conjunction with that, and sort of hand in hand with it, to tie that to the idea of electronic submission of data to a central database prior to publication. I think when the Human Genome Project was taking off there were a lot of debates about, "Well gee- we're going to fund these labs to sequence the human genome. Do they get to pick and choose when to release the data? We'll wait ten years for them to put the big paper together where they report the genome?" I think in large part because of the example and experience with GenBank, the community and the funding agencies and the scientific community that were going to be involved were very comfortable with a plan that really called for immediate release of data as it came off the quality control from the experimental pipeline. It certainly encouraged and wanted the community to analyze the data and understand it and do good things with it, but it was the responsibility of the groups generating the data to put it out there under the public eye. And that's carried over- that's more DNA sequencing, but I think there are also great examples from other types of molecular biological data that have drawn on that paradigm shift that GenBank introduced at that time.

So I'd say that's probably a good enough example there.

Yeah, that answers the question. Any parting thoughts, any last words you want to say about your experiences with GenBank, or about anything really?

Anything in the world (laughing). I mean, just, it was a fantastic project to be involved with, fantastic group of people. That's one thing- it's a great example from early on of- I mentioned some of the people who were involved in this but there are folks like George Bell, who was in the theoretical division, and he and Walter I think early on really incubated this idea. There were some visiting scientists at the lab, I think Mike Waterman and Temple Smith were very keyed into helping to understand what the community wanted and bringing that message back into Los Alamos. And in addition to people like myself and Minoru Kanehisa and Jim Fickett there were a number of other key and central staff: Paul Marn [?], Mike Zukowski, Paul Gilman[?]. Literally staff of dozens of people that were central to making this happen. And then, of course, our collaborating databases in Europe and Japan and our partners in the private sectors. It's one of those big science and big project efforts that wouldn't have happened without a really large team, and it's a really key element of what made that work at that time.

Ok. Well thank you so much for agreeing to meet with me and for your time- well to skype with me, and your time and your insight.