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 _____ 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: David Haussler Date: Oct 24 2012

Person interviewed: David Haussler Student Interviewer _____
(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) ___✓___ 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: David Hench Date: Oct 24, 2012

Briana Mittleman

A Social and Political History of Genetics

Bob Cook-Deegan

Interview with Dr. David Haussler

Briana Mittleman: I want to thank you so much for reading the consent form and sending it back, just to start with some background. I just want to ask how did UC Santa Cruz become a central player in the genome assembly?

David Haussler: So this is an interesting story, we had extensive research in bioinformatics throughout the 90s and we developed methods to recognize the gene sequences and map them into exons and introns in genomic sequences. These were applied to the fly genome in 1998 and then we got a call in 1999 to join the Human Genome Project and apply our gene finding method, but it turned out that after getting into the project several months they weren't successfully assembling the pieces of DNA into segments of chromosomes large enough to actually detect genes. A human gene can span hundreds of thousands of bases. If all the DNA is in tiny little pieces of a few thousand bases, you can't really find genes in those fragments. So we had to work on the assembly because others were not able at the time to produce an acceptable assembly--it became the central theme of the project. It was clear that the entire International Human Genome Project would fail in comparison to the competing project from Celera if we were not able to assemble the DNA sequence into a coherent genome. So Jim Kent from my group tackled that project with an incredible, incredible performance.

Briana Mittleman: Along with that, what was your main incentive to get this all done and put the sequence together before Celera did, as part of the public project?

David Haussler: We felt very strongly that there should be a reference copy of the human genome in the public domain that was available for research without paying a subscription price. This was very close to our hearts. The human genome is such a fundamental part of human biology, human history, what we are. It is such an incredible achievement of humanity to actually sequence our own genome and thereby understand our own origins and eventually our own biology The reference human genome had to be in the public domain to have its greatest impact. We were also somewhat concerned that there would be patents filed by Celera. We understood that they weren't aggressively going after a huge fraction of the genome in that regard; but nevertheless, it is better to have these things in the open intellectual arena from the patent perspective as well.

Briana Mittleman: So you believe that if Celera would have completed the genome first and they would have had it online, but with a fee, research wouldn't be as far now as it is because of the public sector?

David Haussler: I believe that if Celera had been the only source for a coherent assembled public genome then so long as that situation persisted, research would have been hampered because researchers wouldn't all be able or willing to pay a subscription fee to actually use the data.

Briana Mittleman: So I have read that you worked as a graduate student at Colorado, Boulder (The University of Colorado, Boulder) with Gene Myers. So how did it feel to compete with him when he was working with informatics at Celera and you were working in the public sector – was that more a friendly competition or how did that play out?

David Haussler: A very friendly competition, the same kind of competition when we were grad students. Gene is absolutely brilliant and I knew it was serious when he took charge of the Celera informatics effort. I knew they were going to do a good job and of course they did a very, very good job. Gene worked out the idea of whole genome shotgun and along with Granger Sutton led the team that developed the software to really implement that idea. The public project was saying that his approach wouldn't work and that Celera would fail, but you notice that within a short time after that competition, the public project also adopted whole genome shotgun and we use it to this day. So it turns out that the whole genome shotgun was a revolutionary idea.. I have the utmost respect for Gene's ability, so I knew this was serious and that we had to work really hard. After the dust settled and the papers were published in 2001, Gene and I organized a meeting that compared the informatics behind both projects and what we learned from that. We specifically did not invite the leaders of the project –Francis Collins or Eric Lander or John Sulston or Craig Venter, Bob Waterston, etc. - so that we could focus on science and not the acrimony that existed at that time.

Briana Mittleman: So how similar were your – the ways you were doing the informatics – how similar were your processes in the end when you looked at them?

David Haussler: There were a lot of differences because the Celera data was in a huge number of very tiny pieces and the public data was in a smaller number of somewhat larger pieces something, like 400,000. In the public project we used much more information. Jim Kent's assembly program used 13 different types of information to order and orient the pieces along the chromosomes of the human genome.They included the genetic map, the various physical maps that were constructed, including radiation hybrid maps and maps of clone overlaps, and a number of other technical pieces of information. Even RNA sequences that linked together several DNA sequences. All of these information sources were utilized in Jim Kent's magnificent program, GigAssembler. The program had to adjudicate all of the conflicts in these data. Not all these sources of information agreed about the order and orientation so the program had to weigh the evidence and make its choice.

Briana Mittleman: So along with that, what was the biggest challenge you faced trying to order the sequences from the main public sequencing centers?

David Haussler: When you say order, you mean order and orient of the pieces?

Briana Mittleman: Yes.

David Haussler: Yeah, I think that it was precisely the conflicting information that we had. It turns out the human genome contains lots of patches of DNA that look very similar but are actually in different locations on the chromosomes, so it can be confusing where you are in the genome.. It makes the assembly problems very difficult, very, very difficult and so you need all

these other sources of information like the other maps that I mentioned to help sort out that the order and orientation of the pieces. The public project actually did a two stage effort that turns out to have extra information that whole genome shotgun does not. The public project divided the genome sequence into clones that were usually bacterial artificial chromosomes, which are about 150 to 300,000 bases long, and then separately read from each of those. So, for each read of DNA we had information about which bacterial artificial chromosome it came from, but then we had no absolute information about that the location of that exact clone. Some information (was) obtained by a physical map that was constructed at Washington University that gave a pretty good idea of where the clones overlapped, but was nowhere near 100% reliable.

Briana Mittleman: So, what are, in your view, the main positives and negatives to labs around the world working together on large projects such as the Human Genome Projects? Seeing that there are different motivators, such as Celera more dealing with the business side or the more scientific public goods side. What are the main positives and negatives for labs working together?

David Haussler: Well the positive is that we can bring different strengths to the project and you can do some divide and conquer. The public Human Genome Project was able to divide the chromosomes up, for example. Each [lab] collected DNA from a different part of the genome and then we put it all together--so many hands make light work. This is a very, very important principle, the fact that we can muster worldwide groups together who want to participate in an important public project like the Human Genome Project, and divide it up among them, is terrific. Nowadays you can actually put projects on the Web and crowd-source them so that

people all over the world can each do a little bit of work on the project. This idea of enormous international collaborations to do scientific work by crown-sourcing has amplified available computers and workers tremendously in some of these cases like the FoldIt program to fold protein sequences. The negative side of big collaborations is that they make it complicated to adjudicate who gets credit for what part, make sure that everybody is fairly treated, make sure that the project is organized in such a way that the work is not too redundant and makes efficient use of distributed resources. These things become much more difficult as compared to a very tight knit project in one institute with top down control. When it's a large public project among academics it has to be a little more democratic than a large project at a company, and it can be a complicated process to hash things out with that many people involved.

Briana Mittleman: Do you see labs around the world not wanting to work on some of these projects because they have projects that they believe they can get higher recognition for, even though it may not be as large of a project in the worldwide sense?

David Haussler: Absolutely, this is one of the key issues in science today. At a certain level, like for promotion for tenure in a faculty position, the academic community really only considers work that you led. They look at your papers and they ask whether you were the first author on the paper, or sometimes in biology for example they look if you were the last author. The last author is significant because the senior author is often last. If you are not first or last and you are somewhere in the middle of a long list of authors, they tend to discount the achievement even if it is a massive achievement. Having participated in the Human Genome Project under the current tenure evaluation system, being a middle author on the paper gives you practically nothing in

terms of credit for your work. Whereas if you solved a much more specific focused scientific problem and the work was clearly your own and you were the lead or senior author on the paper, that's treated as a serious scientific accomplishment. I think this is definitely a problem that scientists do not get enough credit for working in teams.

Briana Mittleman: Ok, so do you think this is the reason that many of the people that we have seen working with the Human Genome Project were already accredited or already had a lot of projects under their belt and now were ready to work with everyone else together more than at the beginning of their careers when they want to make a name for themselves?

David Haussler: Yeah, you do see that a little bit, but I always argue with my post doctoral students that it's great to be involved in these big projects because you meet people, you learn a lot of things, you get to complete with other labs to try to come up with results first--as long as you also have some work that is just yours to publish separately. Consortium work is a great part of career training, but clearly from a strategic point of view you want to spend maybe half your time on the big joint projects and half your time on a project you are leading.

Briana Mittleman: How has the work that you did on the Human Genome Project influenced the work you are doing now, in the recent times, with the cancer genomics browsers and the genomics zoo.

David Haussler: Oh, it is absolutely foundational to that. All of the systems we built to explore the first reference human genome are now being extended to explore other genomes from different species. The Genome10K project that I organized with Steve O'Brien and Ollie Ryder plans to get one reference genome for each of 10,000 different vertebrate species. In the Cancer Genome Atlas Project, where I serve on the coordinating committee and lead a Genome Data Analysis Center, we are trying to sequence at least 500 genomes from as many different cancer types as possible. The tools that my lab builds and the ideas, algorithms, databases, methodology, hardware, insights, and theoretical ways of analyzing genomes all come into play in these projects. They are all getting more and more mature as we apply our genomics knowledge to greater and greater problems. The field of genomics is explosive and blossoming at an enormous rate with quite beautiful results. And it is just coming into its own, I would say, a decade after the Human Genome Project.

Briana Mittleman: Finally, I have read about a lot of the big things that are about to happen and that are going on right now at UC Santa Cruz including becoming one of the major hubs for storing and processing individual genome data. How do you see that unfolding and what is kind of in the future for what you are doing?

David Haussler: Well we are extremely excited to be the first national repository for cancer genome data. We are proud to have built the UCSC Cancer Genomics Hub (CGHub) for the National Cancer Institute. It will hold an enormous amount of data—the data that will be sequenced as part of The Cancer Genome Atlas project and the TARGET project and several other large cancer projects. The Cancer Genome Atlas encompasses 20 adult cancers and the

TARGET project studies 5 childhood cancers. The genomic data from these cancers are of incredible value to research. There have been multiple reports in the highest end journals--*Science*,*Nature*, and *Cell*--in the last few years revealing the molecular insights that are gained from these cancer genomes. They are of incredible value to the field and we will learn an enormous amount about cancer and even just basically about how genomes work and control the cells by sequencing cancer genomes and finding out how the mutations of cancer genomes alter the cells in ways that make them behave as a cancer. That is the fundamental research of cancer and also has a huge bearing on the basic question of biology: how do genomes control cells? So those data will be of extreme value to research. The fact that we are able to house them all in one database for the National Cancer Institute makes it possible to look at a large number of genomes at the same time. This is incredibly important for cancer and for all types of basic research because cancer exists in many, many subtypes. You can't just talk about breast cancer, that is not a single disease, you have to distinguish between basal and luminal and then you have to further distinguish between luminal A and luminal B and so forth. So it turns out that after you fully classified the type of cancer according to its molecular type then it is just one of many thousands of different types of cancer. To really understand any one sub-type you need thousands and thousands of patients.So with thousands of sub-types and needing thousands of patients for each sub-type, millions of cancer genomes will be required to really understand the cancer in all of its various manifestations. We will only get to there if we can build databases that enable the aggregation of cancer data collected from multiple sources. That is the key challenge today. We hope to set an example with the CGHub to show how aggregating large cancer genome collections can promote research. My goal going forward is to try to use that as an example and

help others to build databases that ultimately will be aggregated together for the purposes of research.

Briana Mittleman: Have you found it difficult to find patients who want to give their samples and help be a part of this project?

David Haussler: Kenna Shaw at the National Cancer Institute collects the samples. I am not responsible for directly asking patients, because I don't personally work directly with patients. But projects that are collecting large sets of cancer genomes at this point are facing issues regarding sample collection. As a national project, the Cancer Genome Atlas had to work with different hospitals and medical centers to get tissues. They had to make sure that the samples had the required phenotype information and that everything was properly consented for research use. It was really an incredible challenge to collect those tissues. Its one of the most difficult parts of the project.. I think the vast majority of patients would want to have their cancer genomic data used for research. The problem is that they aren't being routinely asked and there isn't a program to have a common consent that they can sign so that the data will be aggregated. Each medical center has its own rules and for various reasons wants to keep the data separate in what we call data silos. That will prevent cancer research from progressing as fast as it could. We need to break the silos. One of the good things along this lines is a consent form written by John Wilbanks that is called a portable consent--a kind of universal consent to have your genomics data used for research purposes. If something like that becomes widespread we will have many consented samples that are all legally and ethically appropriate for aggregation into one database. It probably will take the activity of cancer patient advocates to make this happen though. From

some points of view, it really is not in the interest of medical centers to share data with other centers. Patients must demand it. They basically have to say, "Look, it's my data and I can do what I want with it – I want to aggregate it into this national database, I understand that there is a portable consent." For all of that though to happen, we must have the funding and infrastructure for the large database and that is quite up in the air at this point. There are a lot of discussion about it, but there is as of now no clear path forward.

Briana Mittleman: Is there a competitor from a private sector in this new cancer genome bank like there was in the Human Genome Project with Celera?

David Haussler: Not at this point and I think that this time it shouldn't end up like the competition between the public Human Genome Project and Celera. I think a better model would be the so-called, "SNP consortium" which was actually a public/private partnership. I don't know whether you are familiar with that, but shortly after the reference human genome it became clear that, of course, we aren't all genetically identical and there are common genomic variations in the population that may affect medical treatment outcomes or a person's chance of getting certain diseases. Your response to a drug might be determined by a genetic variant that you carry. There was worry that particular pharmaceutical or biotech companies would patent these variant and then nobody else could use them in diagnostics or other areas of medicine. To combat that, the SNP consortium was formed. I think it has a more official name which I could get to you, but its goal was to create a public database of human variation so that this information would be prior art and not patentable. It was funded as a public/private consortium. The pharmaceutical companies got together with governments and scientists and they decided it was

in everybody's best interest to make these data available on a pre-competitive basis to all for research and use. I view the cancer genome database as a similar kind of proposition and would welcome both private and public contributions to it.

Briana Mittleman: Ok, that's all the questions I have for you, but thank you very much for taking time out of your day to help the project my class is doing and learn more about the Human Genome Project and what is going on at UC Santa Cruz.

David Haussler: That's great, thanks so much, are you going to make a full transcript of this?

Briana Mittleman: There will be a full transcript that we will make and put together with all the different class interviews.

David Haussler: Great, that's so great. Can you please share the full transcript with me?

Briana Mittleman: Of course, thank you.

David Haussler: Great talking with you.

Briana Mittleman: Great talking with you too.