Scientific Curation: Untangling research data

Jackie MacArthur, Project Leader, Genome-Wide Association Studies, European Bioinformatics Institute

As Chris said it's been a while since I graduated, but it doesn't seem that long ago and I'm sure you'll find time will go really quickly for you too. I'm here today from the European Bioinformatics Institute and to talk about the work I do there, where I'm currently Project Leader on the GWAS Catalog.

This is a brief outline of my talk. I'm going to start off by describing my role at EBI. As I said, I'm currently project leader but until recently I was scientific curator, so I'm also going to tell you a bit about the role of a scientific curator. I am then going to talk about how I got to this role, which is quite a long and winding career path, through a PhD and then Post-Doctoral research. Finally I'm going to offer some advice on working in this field, then I'm happy to answer questions on any part of my career.

I joined the European Bioinformatics Institute in 2011 as a scientific curator on the Genome-wide Association Study Catalog. I worked as a curator for three years, before becoming project leader in September last year. For those who don't know, the EBI is a scientific institute. It's not really "academia", but it's very close to academia and one of the main things that the EBI does is to provide freely available data and bioinformatic services to the scientific community. For any of you who have heard of Ensembl that's produced by the EBI. The EBI is located about ten miles south of Cambridge, on the Wellcome Trust Genome campus, which it shares with the Sanger Institute.

So what is a scientific curator? Until I applied for the job I actually didn't really know. When we think of a curator we probably think of someone who works in a museum, organising or cataloguing the collection. Scientific curators have been described as the museum cataloguers of the internet age. They curate, collect, annotate and disseminate information that comes out of research studies and any other scientific data, which is then stored in databases and then disseminated to users.

Before I go on to explain what I do, I need to explain a bit about the data that I curate and organise. The GWAS Catalog is a manually curated, literature-derived catalog of all genome wide association study results that are published. Genome-Wide Association Studies analyse a huge number genetic variants across the genome in large numbers of individuals, to determine which variants are associated with a particular disease or trait. For example, the fact that BRCA mutations are associated with breast cancer.

In our database we currently have over 2000 scientific journal publications. We extract every association between every genetic loci and the disease or trait. There are over 15,000 associations included in the catalog. On the slide there's a table that shows the data we would extract for one of those studies. We extract information that describes the studies, so: Who published it? Where was it published? What was the disease or characteristic? So in the example this was hair colour, how many individuals they looked at, and where they came from - these were Eastern Europeans - and then which loci in the genome they found. For example you can't see, but there's one at the top near the HERC2 gene which was found associated with hair colour. The catalog would tell you the significance and a bit about the scientific design they used. Obviously all of these associations are important, some of them are with traits such as hair colour, eye colour but many, many more of

them are with diseases. This is the type of information that doctors, commissions and drug developers use to decide what causes diseases. Once we have all of this information in the database, it's also our job to make it available to users in a way that's easy for them to use and this is an example of one way you can access our data.

This is a diagram of the whole of the human genome. Every dot on that diagram is an association between a disease and a genetic variant at that position. The diagram is interactive, so you can filter it. In the example on the slide, I've filtered this for breast cancer. So if you were a researcher and you were interested in breast cancer you could do this; type "breast cancer" at the top and all those dots are positions in the genome associated with an increased or decreased risk of getting breast cancer. You can then click on each of those dots to find additional information; you can find out what gene that's in or near and whether it's been reported in multiple studies. There's statistical significance, that type of information, and there's also links now to get to other related material. The big advantage of this is that researchers don't have to go through and read all of those publications themselves.

How does all this data get from the publications in the journals into the catalog? That's the job of the scientific curators. They identify the papers, which are out there and decide which we need to include in our catalog. We do a PubMed search every week and find out which new GWAS papers have been studied, then the scientific curators read each of those GWAS publications, extract the information and enter it into an interface. That includes the study design, the methods and the association results. Now to be able to read those studies and to extract that information you really need to understand how the studies are carried out, which leads us into consideration of the qualifications and experience you need to become a scientific curator. All of the information is double checked by a second curator, and they're also responsible for checking and the public release of that data. I haven't mentioned it yet but the GWAS Catalog is a collaboration between the EBI and the National Human Genome Research Institute in the States, their main human genome research institute. The Catalog is a complete collaboration, so some of our collaborations are based with me in Cambridge and some of them are over in the States in Washington DC. We have weekly conference calls and we go over to visit them as well.

I'm now GWAS Project leader. The slide lists the responsibilities of the project leader, but some of these are also carried out by the curators – it really is a team project. The other things we do are to work on improving the user interface, in other words the availability of the data for users. The diagram I showed earlier was something we developed over the last couple of years, and we're still developing, to make it easier to search. We also increase or improve the scientific content of the Catalog. At the moment were trying to increase and standardised the ancestry and the ethnicity information that we extract from all the samples. It's really important at the moment; different diseases affect different ancestries to different extents. I also prepare scientific reports, publications and presentations. I go to conferences to talk about the GWAS Catalog, as do the other curators. I answer queries from GWAS Catalog users. For example, we get queries from users asking how they can use our data and what the data means. I manage the project, training the curation staff and I'm also responsible for meeting all the deliverables in our current grant.

What are the person requirements for the scientific curator? Well this largely depends on what information you are curating. This is the GWAS Catalog, but there's also Proteomics databases that

need curators, all kinds of things. You generally need an MSc or a PhD in the field that you're interested in and an in depth understanding of, in this case of variation and how the studies are carried out. You really need to enjoy reading scientific papers. Attention to detail and good organisation skills and self-motivation are crucial, but I think most of those skills match up with a lot of other scientific careers as well.

The next slide shows my long, long winding path to get to my current role. There's a good reason why it isn't a straight line, because when I set out I really didn't know what I wanted to be. I definitely enjoyed the journey, and I think I'm not expecting the present job to be the end of my career path. I started off in Leicester with my degree in Biological Sciences with Genetics and then did a PhD in Nottingham. I did three Postdocs at Newcastle, at the University of California San Francisco, and Sanger Institute and then I moved to the EBI.

So why did I do a PhD? I did a PhD because I loved science, but even in my third year I wasn't actually expecting to do a PhD. However when I looked at the jobs I was interested in they all asked for a PhD. It was really important for me that the PhD was on the right subject. I was really interested in human variation, human evolution. I looked at human evolution and population genetics of the H-Ras minisatellite and the link between that minisatellite and cancer. As I talk through my career you'll see that there's been a big link between human variation and disease. Even though it wasn't planned in that way, I think if you're interested in something, even if your career kind of winds around, as long as you stay true to what you're interested in at the end it looks like it was planned… almost.

Doing a PhD helped me develop my research skills and scientific knowledge. I did a lot of PCRs and Southern blots in those three years, a lot, but I really loved the subject. I loved working in the lab, but it also gave me the experience of managing a project and working in an academic environment.

After finishing my PhD, I didn't know what to do. I knew I didn't want to be a lecturer or a principal investigator, I didn't want to write grants and always having to think of ideas to write grants, but I really liked working in science and I liked working on a project and problem solving. I also liked working for not-for-profit organisations, to help the greater good. I tried looking around in my final year of PhD for non-research, for non-academic jobs but none of these really matched what I wanted; I didn't really find any of them interesting and it seemed at that point that the thing that matched most to what I wanted was being a Postdoc.

I ended up doing three Postdocs which definitely wasn't in my life plan, but every one I've really enjoyed and the reason I've ended up doing three was because as I've did each one I really loved doing it and then another project would come along and I'd think "wow, that looks really interesting!", so I'll go and do that. I've had the opportunity to move around and seen the world. If you do stay in science there are big advantages in doing a Postdoc or working in the States. It's not just because you get to live in a really cool place, although that is really good, because you get personal flexibility and resilience. You also get experience of working in the US academic environment. Before I went everyone said to me "you should really go and do a Postdoc in the States it's really important", but its only now that I'm coming to realise how important US science is to the whole world of science. As I went forward past that point, the fact that I had time as a US Postdoc on my CV definitely helped me. Once again I'd like to point out that this was without me planning really, I didn't have a specific goal "I will do three Postdocs and they all will be on this topic". As it turned

out they have all been to do with human variation and disease. I should also mention that I found these jobs would fit in with my personal life. I actually had two children whilst I was doing my Postdocs, one whilst I was in San Francisco and one whilst I worked at the Sanger Institute. So you can work a scientific career around the rest of your life.

You might ask "Why do a Postdoc if you don't want to stay in academia?" Firstly, you get experience. So that's lots of experience in research and science. You get project management skills, which I'm not sure would be as easy to gain in other fields. Another really big thing is networking, you get to know lots more people in science. Develop your transferable skills and, because Postdocs are seen as a trainee position, you get lots of opportunities to go on courses, to attend conferences and develop your communication skills. If you're interested in science communication, whichever area you want to go in to, you really want to take advantage of all these things as you go through your career.

After Postdoc three I decided it was time to leave research, but what I really wanted was an interesting job in science and, because I had two young children, it was important to be working near Cambridge. I went through all the job adverts, looking for any jobs that caught my eye, that I thought looked interesting. I applied for the job of scientific curator at EBI because when I saw the job advert I thought it matched really well to the things I enjoyed, to the skills that I had and what I wanted out of a job. However, I wasn't sure, so I emailed and met with the manager of the advertised post to discuss the role first. I would really recommend doing that; if you see a job that you think sounds interesting, try and get in touch with the person there and see if you can meet to talk about the role. Not only does it give you an idea about whether you want to do the job, but I also think if you come across well that gets your foot in the door ahead of a big pile of applications.

So was this a good choice? Yes definitely. I think I have found my ideal job, although I probably would have said that about all my previous jobs too. I'm still part of the scientific community, but I don't work in a lab. A really good thing is that the work is looking at the bigger picture. While you might spend three years in a lab doing lots and lots of PCRs which gets published as a brilliant paper, how many people actually read that paper and use the results? On the other hand, I know the GWAS Catalog is used by lots of people you see people putting the map diagram up at conferences and you know that it is used to decide whether a locus leads to a disease. It's really good that I still get to publish and present my work at conferences. However, it's still not a permanent position; it's still based on grants. I think that's the main thing I compromised on, but I put other things higher on my priority list when I was looking for the job.

My advice for getting jobs, and really this is jobs in science that are outside research or outside the academic department, would be as follows. If people are really, really interested in science, then get a PhD because a lot of jobs you'll see out there do ask for a PhD. Even they don't specify one, it still gives you a useful step up. Build up your network is another thing I'd advise. When I was a student, when I was a PhD student, people used to say "you should network, go and network" and I used to hate going to speak to people. I'd never go to speak to a PI of some lab who was competing with us, that kind of thing. I think this was partly because "networking" sounded like some big deal. I'd actually like to rename it as just being friendly. Be friendly to your peers, like the people here with you today, to your lecturers, to people you work with, to people who are also at the same training event as you. In all probability they won't be useful contacts at the moment, but in ten years' time they may alert you about a job or help you in some other way.

Think about what your skills are, and what you enjoy, and try to match these to a job. If you really like communicating and talking to people, or writing and you like science, try and put the two together. Keep an open mind because the job you're best suited to may not even exist yet. I don't think this job of scientific curator existed when I graduated. Also think about the other priorities in your life because you can if you move your career around you can definitely fit them all in.

Final thing, my career has definitely been a long and winding road and I'm sure it will continue to be. But I think the main thing is to continue to enjoy the journey and not necessarily think about the final destination; just try and match your next step to what suits you and then you'll be happier along the way and get somewhere in the end.

Thank you.

Questions

Q: You mentioned the scientific curation team, how many people would be involved in that team?
A: So, we don't actually have that many. I've got one scientific curator in the EBI with me and there's one in Washington but we also have a couple of other people in Washington that help us with curation and doing other bits. But there's a whole other side of my team, the ones I'm directly involved with. They're doing things like development of the user interfaces. Currently we're actually migrating the Catalog from the States to the EBI and we're having a whole new search engine. We've created the diagrams, so there's probably about four bioinformaticians or software developers, however you want to describe them, at the EBI with me. I think I was hired as a geneticist who could speak to the bioinformaticians. Because I'd used bioinformatics in my analysis all through my Postdocs, I was used to the language, even though it wasn't a massive skill that I had initially.

Q: What would you say is the rough income range for a position like that?
A: It is not low; I would say they start at about £30,000, £35-£40,000. It depends if you are looking at scientific curators or if you're looking wider. The only downside is there aren't that many places that hire curators. The EBI hire quite a few curators and their starting salaries are around £35,000. Quite often it forms part of another job. So, as I said, I'm team leader but I also do a bit of curation. If you end up leading a project or something else as well, then the salary range can go up.

Q. What Bioinformatics skills have helped you with your career, and what skills do you think are most helpful I the industry right now?
A: The thing that helped me was, during my Postdocs I had to do all my own analysis' so would used Unix, a bit of R, that kind of thing. I haven't used R since I've worked at the EBI, but I've used Unix and just the fact that I know what people are talking about is really helpful. I've user Perl as well, so Perl and Unix. The other thing is, I've talked about kind of my career route from PhD, to Postdoc, to scientific curator and project leader, now because I've gone through that route I have lots of friends who have done PhD and several postdocs and they're now working as editors or bioinformaticians at places like Illumina. The career path is similar, I would say, and I'd give the same advice about getting skills in what you are interested in, whatever it might be. If you don't want to go directly into bioinformatics then anything that shows you are interested in it is good.

Q (CJRW): Jackie, you mentioned that you've done three Postdocs. Not everyone may know what a Postdoc does. Do you want to just give a brief outline of what that kind of role does?

A: So a Postdoc works in research, normally at a university or an academic institute, normally under a Principal Investigators as they're often called, but essentially they're the same as a lecturer. Pretty much all the lecturers that teach you will be PIs. They apply for a grant because their interested in investigating a certain problem and then they'll pretty much always recruit a Postdoc to do the work for them. So as a Postdoc you're given a title and you're given some ideas by this PI, who's your boss, and then you're generally given it to run with. Normally you'll have three years for your funding on the grant. One of the ones I did, in Newcastle, I was looking at the genetics of retinoblastoma. They had a grant from a children's brain tumour charity to look at that, so I was recruited to look at that so managed to get in lots of retinoblastoma samples and then I looked at the losses and gains in the DNA that lead to that. Then generally what you do is write it up and try and publish papers, go to conferences. Being a Postdoc is often seen as a route towards academia; generally you'd do a couple of Postdocs, then aim to become a lecturer or a PI. I didn't really want to do that, but if you do go down that route then you need to be aware that there are lots more Postdocs going through the process than there are academic positions, lots more. That's why I mentioned I have lots of friends who have done that route and have got really good alternative careers after it, even though they may or may not of wanted to be PIs.

Chris Willmott: Jackie has highlighted the opportunity that science gives to travel over to America and places like that. One of the recognised downsides, however, is that there is not a particularly clear career structure through science but people go on to do PhDs and then a Postdoc. As Jackie was saying the normal expectation of someone going into that step is that they probably want to try and become a PI, a chief researcher and run their own group in a university setting. But there's a very definite pyramid in terms of progression along that 'normal' route – there are far fewer opportunities to be the PI than to be a Postdoc. It is difficult to know with confidence that you will make it all the way along that path. Some people make a deliberate decision to step away from that route, other people perhaps step away reluctantly realising they're not going to get into the PI role themselves. However, as Jackie has already indicated, and what I think our next two speakers on PhDs are going to endorse, is that there's actually merit in having a PhD as a key for opening a diverse range of careers, not just to say "I'm going to be a Nobel Prize winning lead scientist on a particular project".