Winter February, 2018

De-identified interviews for the study: Data Challenges of Biomedical Researchers in the Age of Omics

Rolando Garcia-Milian
Denise Hersey, *Princeton University*
Milica Vukmirovic, *Yale University*

De-identified interviews for the study: Data Challenges of Biomedical Researchers in the Age of Omics

Authors: Rolando Garcia-Milian MLS, AHIP[1], Denise Hersey, MA, MLS[2], Milica Vukmirovic, PhD[3], Fanny Duprilot[4]

[1] Bioinformatics Support Program, Cushing/Whitney Medical Library, Yale University, 333 Cedar St. New Haven, CT 06510, USA Rolando.milian@yale.edu
[2]Head, Science Libraries, Lewis Science Library, Princeton University, Princeton, New Jersey, USA
[3]Associate Research Scientist, Pulmonary Critical Care & Sleep Medicine, Yale School of Medicine, Yale University
[4]Service commun de la documentation, Université Paris Diderot, Paris, FRANCE

Keywords (MeSH): Information Seeking Behavior; Data Interpretation; Statistical; Genomics; Computational Biology; Software; R; Survey

ABSTRACT

Background: High-throughput technologies are rapidly generating large amounts of diverse omics data. Although this offers a great opportunity, it also poses great challenges as data analysis becomes more complex. The purpose of this study was to identify the main challenges researchers face in analyzing data, and how academic libraries can support them in this endeavor.

Methods: A multimodal needs assessment analysis, combined an online survey of 860 Yale-affiliated researchers and 15 in-depth one-on-one semi-structured interviews. Interviews were recorded, transcribed, and analyzed using NVivo 10® software according to the thematic analysis approach.

Results: The survey response rate was 20.4%. Most respondents (78%) identified lack of adequate data analysis training (e.g. R, Python) as a main challenge, in addition to not having the proper database or software (54%) to expedite analysis. Two main themes emerged from the interviews: personnel and training needs. Researchers feel they could improve data analyses practices by having better access to the appropriate bioinformatics expertise, and/or training in data analyses tools. They also reported lack of time to acquire expertise in using bioinformatics tools and poor understanding of the resources available to facilitate analysis.

Conclusions: The main challenges identified by our study are: lack of adequate training for data analysis (including need to learn scripting language), need for more personnel at the University that provide data analysis and training, and inadequate communication between bioinformaticians and researchers. The positive impact of medical and/or science libraries by establishing bioinformatics support to researchers was identified.

Table of Contents

Interview 1

Which department or departments are you affiliated with at Yale?

What is your role with in that department?

Post-Doctoral Associate

How many people work in your lab or team?

I work in a big group, I would say between 10 and 15**.**

Which types of data analysis are most important to you?

Okay, I like this question. I can say for myself it is transcript omics data and then everything downstream from that which means differential gene expression analysis, signaling pathway network analysis and then many research questions can be associated with that transcription factors or some other or some other kind of more specific questions, but yeah it's mostly transcript omics analysis.

What are your current practices for analyzing your data? How many people on your team are involved in analyzing the data?

We have many tools in our current lab. We have some software that will allow us to make a flow and a kind of pipeline and to run it, but that software cannot answer all questions, so we have a bioinformatician with whom we are working closely and then some questions that cannot be answered by generic software that person will help with, so it's R-based packages. Basically, we use that.

Are you current paying for those packages, or are those free?

No, those are free. Some software we already paid and, they are quite expensive. Everything that is based on R or that can be retrieved from Bioconductor it's free.

So, you have to wait for people to do the analysis. That is important to know because if we can accelerate those workflows then it would be…

It would be good if we can figure out some kind of pipeline that biologists can use, some people who are not programmed trained because R still requires a certain amount of self-learning. So, I think for the platform that you proposed like Galaxy or that other. What is the other one? Partek Flow would be helpful for initial analysis, but I think that any how this data needs to be checked by a bioinformatician.

How much time does it take, on average, to analyze your data? Would you change anything if you could?

I think every person can tell you that. In my case, I have kind of periods because I need to do experiments to create data and then to do data analysis. Let's say for me, learning Metacore, it took me one-two months, but it's not a whole day. It's the time I had apart from my experiments. To be comfortable with the program, but again maybe some people learn faster.

What type of hardware do you use? Do you have any problems with this hardware?

No, No. We're lucky in my lab. But I think that is a good questions. Many of these analyses cannot be done on simple computers. It just takes too long... It's blocking and annoying when these software are running like let's say for mapping or sleep analysis. You cannot do anything else. It takes four hours.

Storage capacity. Storage capacity people have access to Grace. So, Grace is a backup server and storage space and our lab has like terabytes and terabytes of storage there. That needs to be paid. I don't know how much it costs. I know we have it. We have lots of data, and I know it's there. I think everyone has access to Grace.

I think everyone on the medical can have that? I don't know the paths to get it and how much it costs.

What are your biggest challenges in analyzing the data? Is there a bottle neck in your data analysis? If so, where and what is it? How do you try to get around that bottleneck?

I would say it's not the research questions, so basically simply using all these new programs and software and packages. It takes time then it's really like, if you need to work in a lab bench …I wish we had a pipeline where we could plug in data and run it through and see.

Do you have data that you have not been able to analyze? Why?

No, No because I think every data can be analyzed. It's just a question of how much time it will take for someone to analyze. We have someone now is in the group who is analyzing regulation data. It just takes time to find good software to learn it, to plug in your data sets. The answer would be no, but it takes much more time than I would like for it to take to analyze it.

What would make it easier for you to analyze your data? (Specific resources, training, personnel)

I think what is doing is great like in kind of allowing people to use these kinds of platforms that are easy for biologist to use. Another way, definitely by having kind of human resource in terms of people whom we can go and ask questions. People who are available to answer the question in time that is less than two weeks. A lot of people do access to the platform for free and having people who can answer the question. Maybe, we can have some kind of group of available who can answer the questions in less than two weeks. That would be nice. Yes, but allowing people to have access to the platform for free or having more people who can answer the questions. Maybe you guys or we can have a group of available bioinformatician who can answer, but maybe that's too idealistic because I know all these people are very busy.

In the best case scenario yes, more people who can answer our questions and troubleshoot with us.

I think the one that he is proposing, Galaxy.

Partek Flow. I think that will definitely help us to have an initial grasp of data. I think with this Partek Flow we will have lots of questions. So if this platform comes with customer support. That's great that we could call someone. It would be very helpful. That person doesn't have to come from Yale. But someone who can answer questions. For example, the person who came from Metacore answered my questions. I could call her and my questions would be answered in a day or two. It really made my whole learning experience nicer.

I'm not sure, especially with cleaning data. I'm not sure honestly how many answers you can get from Galaxy. I know that many people use the program.

Is training needed at Yale? Any suggestions regarding the training? Any suggestions how we can help? How do you prefer training to be delivered? Are there others in your research group that could benefit from training? Why?

My preference is in person, but other options, such as webinars where there is a data of communication with support. That is okay.

I think you guys should invest in the R-learning things for people.

More training, like regular training-I think that would be a good investment.

Yes, R for transcriptomics or R for genomics, basically training a Bioconductor. That would be nice if someone could spend time.

Is there anything else you'd like to add about data analysis or is there anything else we should know?

I've discussed a lot of this with so I am kind of knowledgeable about you guys can provide. I know we have a group where sends emails. I am not sure how easy it is to find information that is shared within that group online. Simply, if I just join Yale or I'm here and not aware. If I type, Yale Bio…What would be the keywords to find the group or to find that information in one place? So… I'm sure you can create some kind of link on the medical library's website or some kind of data analysis, and maybe it exists already, but I didn't look specifically for it. But, I think that can be helpful.

There are a few groups on campus who are doing hard core data analysis. But they are doing just for their own group not for the whole campus.

Yes, something like that or maybe even a group… I'm not sure now. Is it better to show their names, or is it better keep them as a group and just send a question. I need help with this or this. I feel there is resistance from bioinformaticians part because they learn a lot of mathematics, biostatistics or statistics, and they know how to program. It is easier for them to write the code than to answer questions to create simple graph. But then, we are the ones who needs to come up with all these questions. Sometimes it is easy to come up with all these questions if you never studied data analysis in depth. Doing it in R, I think it's really, really a learning curve. I think it is good, all these details. It just takes time. While maybe some of the software can immediately help you with some questions and see how the data analysis look like, which I think on the other end bioinformaticians, I think they stay away from software…which I think that is one of the layers of discrepancy. We should come to the same result using Galaxy and R. But like if we have some questions with Galaxy, they won’t be able to jump in. We need to have some support who really knows.

I mean with R you can do all of this, but you need to be able to write a code. Those are the long codes. I know even when we create a graph in R and we want to put colors in certain dots, it’s like a few lines of code.

How do you currently share data with collaborators? Do you have any challenges in sharing data with them?

No, No, not really. We use global transfer, global transfer system. We connect Grace to global to outside lab, and then it goes. Yes, it takes time to connect it. It takes two weeks. Again, it’s possible it’s here on campus. We can use it.

Interview 4

Which department or departments are you affiliated with at Yale?

I work in the Department of

What is your role with in that department?

I am a Post-Doctoral Associate in the

How many people work in your lab or team?

It is a little bit different because in our Lab under supervision of Professor Job we have one Post-Doc me, and one Lab manager,. But our team, is much bigger. Basically, there are two PIs, There is an Assistant Professor, Professor who recently hired a Post-Doc... There is also a computational core, which consists of Professor, her student program analyst…who and there is also a programmer Husner.

That's your team?

Yup.

Oh it's a huge team.

Yeah, so it is an YCAD team and clinical coordinators and clinical trial recruiters and some nurses who do sputum induction because it's all about

It's a huge team, more than 20 people, so it's one of the big teams.

What type of research are you involved in?

I am involved in translational research. Basically, I have couple of projects. Basically, I am doing Translational Immunology and Genomics. So, we are doing drug discovery and for. So, it's all about biomarkers and potential new targets for therapy of pulmonary diseases and asthma.

Which types of data analysis are most important to you?

I think for me personally, the most important parts are proteomics. Also, I work with PubMed databases, and now I have started RNAseq project, so I will generate a lot of RNAseq data and incorporation, and analysis because I just collected already yesterday and still I have to collect other experiments, so it will be a lot of RNAseq data.

What is the sample size in your projects?

Basically, has... my boss, Professor, has a huge biorepository because he is the. What's happening, every week five to ten patients are recruited. So, already in this database, there are around let's say 800 to 1,000 patients, and it's longitudinal data, so same patients came, and they come back. And like various, more than…not thousands, but there is a lot of clinical data in this database, like clinical phenotypes-then various data on blood. So there is like cytokines, profiling. Then, there is RNAseq data and whole exome sequencing data. So, it's a lot.

For all of these thousand samples?

More or less, Yes. It's like trying to implementation of…it will more or less be in the same space. Right now not all the data are together, but it is generally like what we have. But in my work, because I am doing this for clinical development. I usually have tissue culture samples, and it's around 100 samples a month, maybe 200 samples, and now it will increase because we just started a collaboration with. So it is like growing.

Let's talk about how you use data in your research. How do you store your data? Do you use Grace or some other storage option at Yale?

person's name, who is our assistant professor does orinasic data analysis but general RNaseq she uses Grace. Yes.

So you have a bioinformatician that helps with the projects?

Well, she is the director of our section. She is the. She is using Grace. I personally haven't touched Grace. Maybe when I will get orinasic data I will start to use it, but I haven't yet because I am still collecting samples for orinasic experiment. I haven't submitted them yet. It will be before Christmas.

What are your current practices for analyzing your data? How many people on your team are involved in analyzing the data?

Basically, my practice is about analyzing data. For example, recently, I had this huge dataset. I submitted 54 samples from experiments done exactly same-54 samples total. They were analyzed for 660 biomarkers by Biotech, so what I did, I normalized this data, using R-and we got to lower limit of detection, and then I had to do simple t test over and over in R. When it's something like more sophisticated, or if I have issues, I send emails to , who is our Assistant Professor, and if she can answer me, she answers me, but I mean, I had to learn a lot on computational stuff, and it is still a process, and I also find GraphPrism very useful software because it helps. There is also JUMP, which is available for free for Yale users. It is also very nice to do analysis, so…

So, when you send this data to the Bioinformatician, is there a waiting list or something, or is it right away?

If it is something simple, they expect you to do it yourself. For RNAseq, I think it will be a waiting list.

You can now do RNAseq. You now have the tools.

For Qlucore heat plots 660 biomarkers-it took about one month to generate the plots. Access to Qlucore when granted trial access for Qlucore -guided me. It was a new software for me, but I was able to generate the… plots in one week. I was learning, so now I can do it in like a couple of clicks. So it was really, really amazing, and it will save a lot of time. Qlucore is also capable of running it over and over-like tests, and it's integrated with R, so it's really, really nice not to be on a waiting list forever.

How much time does it take, on average, to analyze your data? Would you change anything if you could?

It depends because I am always trying to find quicker ways to analyze data. Because you cannot research- you cannot one month. Basically, when I get data, it depends on sample size- but maximum one week to have results but generally my boss would expect sometime in the morning what we see in the data.

Would you change anything, if you could?

Right now, I am very happy with Qlucore, so I don't have to change anything. Basically, it is like de-code message, to learn R-understand how to do simple statistics yourself to be fast.

It's what I'm doing. It's a process. I think I progressed for this year because I attended computational school and immunology this summer. It was a nice and fun experience.

SPSS is not very practical in my opinion because I have it. I was not able to apply it to my data. It's not easy to integrate in SPSS.

I just use regular R. I know how to do simple things like t test or simple correlation, but I have to learn more soon.

What tools or resources do you use? Are these open access (free) resources or do you pay a fee?

Thanks to Yale, I don't pay a fee for Metacore or Ingenuity, but its's true I am now only using Metacore because it is more comprehensive for me.

I tried Partek Genomics Suite this summer. It was a trial. I tried Qlucor as a trial. I used R, Crofpad, and JUMP. Yale either has access, or it is free.

If you compare the two things, Qlucore and the Suite. Qlucore is much easier to understand.

What type of hardware do you use? Do you have any problems with this hardware?

No I have MacBook Pro and MacWeb Top. I use RAM.

Do you have any issues with storage or computing capacity?

No, especially now there is like code 24 backup option.

I see what you mean. That would be cool. I was fortunate bought this Mac laptop. It is a very powerful computer. But I understand, sometimes with laptops, it takes like two hours to run an analysis.

Do you work with a bioinformatician? If so, what is your interaction with them like?

I work with her. I have a relationship with. It is a learning relationship. The only way it can work. You have to pick up stuff from the bioinformatician and learn it. The bioinformatician cannot do everything for you.

What are your biggest challenges in analyzing the data? Is there a bottle neck in your data analysis? If so, where and what is it? How do you try to get around that bottleneck?

My biggest challenge will be when I get the RNAseq data, so my first challenge will be to wait for data to be delivered because it will take exactly one month to mix with data. Then my challenge will be to wait for RNAseq data. Then, it will be a challenge to sit down and pick up what is relevant because bioinformatician not always understand the biology behind, so the challenge is really extract what is pertinent to your project because bioinformatician can tell you okay this is significant but it may be total nonsense for you.

Is there a bottle neck in your data analysis? If so, where and what is it? How do you try to get around that bottleneck?

We will see by Christmas.

Okay, not yet.

I am trying to plan things and very happy that you have a license with Qlucore.

Do you have data that you have not been able to analyze? Why?

For that reason, last summer I hired a statistics tutor. Finally, it was analyzed. I had to hire somebody.

Because you need to learn to analyze the data.

Yes, Yes, Yes.

Because you didn't have the knowledge to analyze the data.

So I hired somebody.

But, it wasn't the best choice.

No.

Okay.

I hired three different people, so it was nice and fun to manage.

What would make it easier for you to analyze your data? (Specific resources, training, personnel)

I think we discussed it with couple of times. There should be some open code available in R. Like for standard things-proteomics-some open code. Like some standards-with summer school it was helpful-give us some kind of code if you want to generate heat map… Like if you want to do this, use this. Then you can apply it to your thing. Then I understand that it is not very straightforward because it might bug and you need some help from somebody and bioinformatician usually don't want to be there to debug you. So, I think it might be helpful. I like when our statistical students are available-but unfortunately, my user experience when I come to that desk, no one is there.

I know there are students here-statistical students available. I tried a couple of times, but unfortunately my user experience when I come to that desk, no one is there.

So that was my challenge.

So, the days that they are supposed to be there, they are not there? That's good to know.

I really didn't not once. There were times when they were supposed to be there and I realized that I can't rely on it, I moved on.

To have somebody sitting there at least once a week would be very helpful

A bioinformatician you mean?

Yes, like somebody it doesn't have to be a professor-like somebody knowing R. Like I don't expect this person to do your job-but if you ask a very specific questions. I did this-Can you please advise me? Like what's the problem?

Okay

Not like doing your job-but simply like- debug you.

So he has to know R? That's the thing.

Yes.

That's the thing because a Bioinformatician might not know R. They may use SAS for an example.

Yes.

So that's the problem. I really, really tried to use all the resources available at Yale. I tried to go to and folks they use SAS. Basically, they require payment. The first consultation is free. Then you have to pay. The issue is you can't pay out of your pocket. The PI should pay. There is a tough situation. The PI will not understand, if we have bioinformatician in house, why would we pay? Still, if this solution is overcome-they will use SAS. You have to use R; you cannot learn from them. The will require authorship on our project-you cannot contact them when your bioinformatician wants to be on your publication and if you need to put that bioinformatician-

I see.

Then, it will be a clash. Two bioinformatician will not survive under the same roof. Then there will be a struggle which bioinformatician goes first in the authorship.

Yes, that is what I mean, not a person who will do stuff for you. Let's imagine you spent your night on your R-code and something bugs-you really need-like hey this really is not appropriate-

Not only R, but statistics in general. Like advising try this or…

And then, there is authorship.

Yes, and I agree with you. I explained to you what I encountered in -like even if I was okay with it. They were not able to do stuff for me so…

Is training needed at Yale? Any suggestions regarding the training? Any suggestions how we can help? How do you prefer training to be delivered? Are there others in your research group that could benefit from training? Why?

R- Intermediate R, but with code being available…

R for RNAseq. R for Flow Cytometry-like specific codes.

Yes concrete so you can build your knowledge on it. In journals, they will ask you to publish your codes in supplements. You have to know it.

Yes, that is a very good idea.

How do you prefer training to be delivered? Are there others in your resource group that could benefit from this training?

Everybody can benefit from this training because it is everybody's problem. Everybody who does Genomics has this issue. Everybody has to learn anything for best way to be delivered. It can be delivered online. It doesn't have to be in presence-just code need to be uploaded. Then, everybody can download and if there are questions, they come to ask specific questions to students or whatever person available-not like do my project for me but I tried your code-but can you please tell me why does it make error.

Okay.

Do you prefer training in person or both that you can watch?

Recording online will save because everybody's schedule is different. It is accessible for a broader audience. It's better to make these resources with code being available online, and then if you have questions, you can find a way to send an email to the person who organized this training or something. I understand that there will be a wait time-because if it will just like classroom…

So you will have to have this code.

Okay.

How do you currently share data with collaborators? Do you have any challenges in sharing data with them?

Yale Box. No challenges. You upload, and it’s secure. We can download it.

Is there anything else you'd like to add about data analysis or is there anything else we should know?

That is more or less everything. Do you use resources for data analysis? I shared my experience with students; that I was waiting.

I attend workshops on R on Science Hill, so I attend R for beginners, and I attended intermediate R. Their code is not applied to the projects. It is applied generally.

So, then I attended this computational immunology school, and I attended a pre-workshop for it, where they were teaching it again. I attended most of your seminars you organized: Ingenuity, Metacore. Luckily, I had with Qlucore, who really spent a lot of time with me doing webex and stuff.

Do you use resources at Yale library for your data analysis? Are you aware of all resources that we have at Yale library?

I am totally aware.

Right now, actually I am going to go to a meeting with - there exists technology where you basically submit one sample basically, and they screen your sample for thousands of proteins. So, I am very curious because it is a little bit driven to my project-how they will deliver data. So, I will let you know-if like I will learn something interesting-which all came up because I plan to use this technology in the future.

Interview_5

Which department or departments are you affiliated with at Yale?

What is your role with in that department?

Graduate student

How many people work in your lab or team?

4 graduate students and 1 undergrad

What type of research are you involved in?

Anything to do with biochemistry or cell biology. We are doing a lot of transcriptomics.
We did a genome screen, RNA-seq

What type of data analysis is most important to you?

RNAseq. I have used pathway analysis. In the future I might like to do ChIPseq

What is the sample size in your projects?

We work with cell cultures so it is a little different. Everything has to have 3 biological replicates. But that is about as far as it goes.

So it is more like small samples.

Yes. We are not doing anything gigantic.

Let's talk about how you use data in your research. How do you store your data? Do you use Grace or some other storage option at Yale?

All of our RNAseq data is stored on the RNAseq core facility. Some of our Data was stored there. Other than that we just have external hard drives in our lab and that is about it.

Do you have a lot of external hard drives?

No We probably have 3 or 4 then we have some really old data that is saved on floppy disks.

Rolando: Floppy disks?

I don't know what is on them.

What are your current practices for analyzing your data? How many people on your team are involved in analyzing the data?

I would say two out of the 4 graduate students are involved in analyzing high-throughput data We are trying to bring in our undergrad who is working with us. The other 2 students do mostly yeast or high-throughput so they don't really analyze to the level that we analyze.

How many people are involved in your team in analyzing the data?

I don't know yet because we are still trying to figure out project flow.

How much time does it take, on average, to analyze your data? Would you change anything if you could?

If I had to put a number on it maybe 2 days. Doing the whole analysis and the follow up. For the three samples. Going very slowly.
Rolando: Would you do anything about it if you could? I don't know enough about it yet to change it. I would change my own computer skills. I could use some more general knowledge of coding and how to work programs and things like that. Yeah.

What tools or resources do you use? Are these open access (free) resources or do you pay a fee?

We use Partek Flow and we pay for that. (Yale pays for that.) We use a lot of free software from the Yale software library. GraphPad prism things like that. I have used Matlab in the past but that was mostly for classwork.

Why don't you use matlab for working?

Honestly, I am not that good at it. I have not had a real reason to sit down and work it out because we have other programs that get around that.

What type of hardware do you use? Do you have any problems with this hardware?

I use my personal laptop a lot and it is fairly new so I am lucky. I struggle a little bit because our lab has Mac computers so a lot of our data being compiled ends up on a Mac and I am not a Mac user so the interface is different and a little bit of a struggle. Our data is stored on hard drives and is pretty easy to access.

Do you work with a bioinformatician? If so, what is your interaction with them like?

Yes. It has been alright. The interaction with him we gave him the data and he analyzed it quickly. Which is great and give us sort of a printout but I guess our issue with that is understanding what the bioinformatician did. There is a little disconnect between how he did the analyses and how we understand what he did. Part of it that they don't do any biology and we only do biology so there is that difference there. Part of the reason we switched to Partek Flow and doing this on our own is that we can see all the steps you can't do that when he just gives you that final read out of the genes that are good.
I think this happens in many fields with people who work with computers and people who work with something else.
You almost need a translator.

What are your biggest challenges in analyzing the data? Is there a bottle neck in your data analysis? If so, where and what is it? How do you try to get around that bottleneck?

A learning curve I am graduate student and everything is new so every time you do something you totally have to learn how to do it. I have no background in data analysis. I was a very heavy chemistry undergrad - and did not do any high throughput at all. It has been a lot to try to learn the computer. Just how to do RNAseq.

Bottleneck: at the end now because we get a list of genes and we try to figure out what follow up studies to do with these genes. We may have 15 interesting genes out of that and you have to pick one or 2 something that would be reasonable to study.

Do you have data that you have not been able to analyze? Why?

No right now no. I spoke to you about the Partek Flow issue and I was able to get it uploaded and everything looks ok now.

We go some extra reads and my problem originally was that I had to combine 2 files. It is actually pretty easy if you know command line in a Mac. Again this is the Mac issue. Sam in the lab knows Macs and I know command line and what I wanted to put. Between the 2 of us we figured it out. It was a command line thing and it has been a long time. We even talked to the bioinformatician. Again that computer difference it seemed very easy when he explained it but when you try to do it yourself there is some sort of disconnect. Finally figured it out then I uploaded it Partek Flow and hopefully it will run today.

What would make it easier for you to analyze your data? (Specific resources, training, personnel)

Basic training would be good. I think part of the problem is that CSSI offers coding classes but it is too far to go up there all the time. It is prohibitory far. Yes, then I have to interrupt my experiment. If something were here I would go to it. I have friends go to the one at science hill and they say it is useful.

Is training needed at Yale? Any suggestions regarding the training? Any suggestions how we can help? How do you prefer training to be delivered? Are there others in your research group that could benefit from training? Why?

More command line and coding I think anything in R really.

They like to do their own analyses and if you code your own thing then it is better. They view it as better and I don't know because I can't code anything, those interfaces I can't decide for myself. In general, the things that you have provided for me are fine so far. And what they do I think RNAseq people. I don't know if you could get a person who codes to switch to Partek flow even if it is offered. I think people learn their specific way of doing it and they like all their subtleties and maybe a user interface would not give them the power they want to control their analysis. But I think for people who have never done Partek or RNAseq before it is a way to get them to use RNAseq.

We had someone in our lab using Cryogen and he did everything on Linux so if I wanted to do it, there is no software that will easily do it you need some coding. If you are doing anything structurally related it is similar. I use a program sometimes called Pimol. It is half user interface and it is half coding (python based coding) and if you know the coding it is easier. That allows you to look at protein structures and you can pull up from the PDB and you can zoom in on the area you want and can color it to make figures for your graphs.

You don’t absolutely need the coding but since it is half coding and half not if you want to do anything beyond the basics you need the coding. It has been interesting.

How do you currently share data with collaborators? Do you have any challenges in sharing data with them?

So far I have not had to share data with collaborators. I have shared basic experiments but nothing like high-throughput so I don’t know.

Is there anything else you'd like to add about data analysis or is there anything else we should know?

I don’t think so.

Do you use resources at Yale library for your data analysis? Are you aware of all resources that we have at Yale library?

We use all the programs all the time. PubMed all the library search engines…. I think I am aware. I learn about new ones every other day.

Interview 6

Which department or departments are you affiliated with at Yale?

What is your role with in that department?

I am a graduate student.

How many people work in your lab or team?

About ten people

What type of research are you involved in?

I would say mostly genomics and transcriptomics.

Which types of data analysis are most important to you?

Chip-Seq and RNA-seq. Those are the two that I am doing directly. People in my group also do lots of single-cell sequencing, and processing those have also been challenging as a group as a whole. We also like to look at methylation data as well.

What is the sample size in your projects?

For example, for my project I'm working not with mice or patients but with a cell line. I have different treatment groups, so in total for my experiments, I have 16 samples of Chip-seq and 16 samples of RNA-seq.

Let's talk about how you use data in your research. How do you store your data? Do you use Grace or some other storage option at Yale?

So, I'm not familiar with Grace. Initially, I stored a lot of my data on to Luis, the HPC, the cluster, but then I was that is not reliable in case something happens. Storing the data became a big deal for us, and to that end, our lab got a three terabyte hard drive just for my project, so that I could put all my raw data in there at least.

Just for one project?

Right, yes.

I am thinking about the other people and projects and how this become…

So the storage of data is a big deal because it makes us nervous about having the data in the cluster. It shouldn't be used as a storage at all.

You have to back up the data.

Right, yeah.

What are your current practices for analyzing your data? How many people on your team are involved in analyzing the data?

I actually am the first person in the lab to have a bioinformatics projects. And following me a year ago there was a research scientist came in, and he is trying to do more bioinformatics. Right now, the Bioinformatics is not smooth in our team at all, and I'm working with this one bioinformatician who is in another lab, and I'm working with that person very closely to make sure I'm doing everything correctly, and everything is right. Yeah, but mostly we use the traditional ways of mapping and calling the peaks and everything for RNA-Seq top hat, cuff links, and cuff difs. Those are the ones we are using.

So you do methylation analysis?

Not directly, no, but I like to go on the USCG genome browser and download those raw data or download the peaks and look at them.

Okay.

On IGV, yeah.

How many people on your team are involved in analysis of data?

I would say two.

In a team of ten, that is not bad.

Yeah, Yeah.

How much time does it take, on average, to analyze your data? Would you change anything if you could?

Yes, definitely. It takes a long time with the data. I would say… All of the analysis wasn't done at the same time, and there was a lot of troubleshooting, but I guess if we did it all at the same go, it would take at least two to three months to do everything.

So at the same time, what do you mean?

What I mean is there were times when we had the first replicate, and we did all the analysis and calling the peaks and looking at the differential gene expression and then we got the second replicate, and then we had to figure out how we were going to pool the data and call the peaks or should we keep it separate and then just look at the overlapping peaks all of those considerations. And then we added more layers of complexity into the analysis. For example, for the peaks we wanted to know where in the genome they are. Are they in the promoter, or are they in the intergenic regions, so we used some R-packages to do that. We also wanted to call like snow what the gene is closest to the peak, and we used also another R-package for that.

Great.

We wanted to also know motives, so we integrated that and everything.

What tools or resources do you use? Are these open access (free) resources or do you pay a fee?

No, we had stayed away from using things that would require a fee. I think part of the reason is that as a new bioinformatician it has been hard for me to convince my PI that that is an expense that we need to make. No, but everything else we have been using are open access.

But, you are aware now that we have to pay for Partek Flow for example for RNA-Seq data analysis?

I see. I didn't know about Partek Flow. I was at the seminar. Was I at the seminar?

The last one we had was less than one month.

I probably wasn't at that one. I was in the one before that. I don't remember the name… Qlucore.

We are going to have the license for Qlucore.

Okay, Okay.

So, you might be able to use it.

Qlucore is definitely something I will use. I will definitely check out the Partek. I'm told I have been using, and one tool that I've been using and thankful for is the IPA to look at the pathways and everything has been very helpful for sure.

Great.

What type of hardware do you use? Do you have any problems with this hardware?

Yes, right now, I use a Dell computer, and it is not working well at all.

We are going to start piloting a high performance work station here at the library.

Okay.

If you have to run some works and stuff, you can leave the computer running overnight and install some things in the computer.

Okay.

That would actually be very helpful.

That would be good especially with running it overnight because I have issues with that if I started running something on the cluster here and then I just pack my computer and go back home maybe there is some kind of problem it is slower through VPN doing it overnight. It is just harder.

If you leave five or six running, you just come back the next day and pick it up. It is something we are going to buy the high performance workstation more than what you have at the lab right now, so I can give you the details later.

Yeah, definitely.

This is exactly for you and the first time we are doing it. I have an interview with a bioinformatician…researcher.

Do you work with a bioinformatician? If so, what is your interaction with them like?

Yes, I do.

What interaction? You are the bioinformatician?

I am, but I work with bigger bioinformatician. So now it is much less infrequent. In the beginning of the project, I was trying to meet him at least once every week or once every two weeks to make sure we are on the same page with how the processing is done. The interaction was very frequent. I know what the biological questions are and the processing that is needed but that doesn't always get translated very smoothly with bioinformatician.

That is something that people have mentioned-the communication between biology and bioinformatics.

So it has been helpful for me to learn a lot by myself. Then, I know what I am doing. There's no confusion.

Talking about this training needed. I didn't know that you have some knowledge of R and stuff. There is a huge need in learning R. What people would appreciate is somebody taking leadership in providing some of the basic training in using R. Like R specific analyses like things like that. Kind of like basics.

I have been approached by students who are doing bioinformatics as part of their project, but they don't necessarily come from a bioinformatics background. And I feel the need is very basic.

It starts from how to download the fast queue files from the server and bring it into either their local machine or directory or even their cluster. Yeah, it's UNIX programming all of those things. I feel there is a very basic need. So I found the class very helpful that Jean Noonan teaches for bioinformatics, but I think it should be taught in the fall, in the summer.

Are we talking about the course or the class?

It is a course.

It requires some more time commitment and stuff.

That is true. I can tell someone who is coming to me just use WGET- to download everything to the cluster, or this is how you write the thing for Top Hat. This what your command is for running Top Hat of Cuff Links, but then they cannot troubleshoot any further because they don't have more knowledge around that.

They get stuck there. Even though, they know some basics.

It' just not possible for me to troubleshoot with them.

Yeah, of course. I understand.

I have my own projects, so…

Would you be willing to provide some basic training, maybe? Because it is a real need I can tell you?

Yeah, yeah.

No, I would rather provide training to group at the same time, rather than dealing with one-on-one.

Yes, that is something that we would have to talk about later.

What are your biggest challenges in analyzing the data? Is there a bottle neck in your data analysis? If so, where and what is it? How do you try to get around that bottleneck?

I think you mentioned storage was one.

Yes, storage was one, which we sorted out by buying the hard drive.

I think you mentioned the process of analyzing data.

So, the remote running has been troublesome and that's because personally the machine that I use. I applied for a grant to update my machine that way. Well, I don't know if the challenge is more personal, when I started as a grad student-like three years ago, I didn't have much programming background, so learning that alongside the biology that I was learning in the lab, that was challenging, but you know, you do what you have to do I guess. That was challenging yeah.

Because it was time consuming?

It was time consuming. It was something very new to me that I have never done. Our lab is primarily a wet lab, so there wasn't much help within the lab- I can't do this do you know how to? There wasn't a lot of me asking other people. The course that I took was ultimately very helpful.

Is there a bottle neck in your data analysis? If so, where and what is it? How do you try to get around that bottleneck?

I wouldn't say there is a bottleneck right now, but in the beginning, we worked with two bioinformaticians. So, we worked with one bioinformatician here at Yale. We worked with a group- a whole lab and then finally before we found this bioinformatician that I am working with now who is helping me a lot, but before that, it was really difficult to find someone who could understand what we were asking.

This bioinformatician is part of the team or someone who collaborates with you?

Collaborates with us.

He is part of the team.

He's part of another lab.

Okay.

He's not part of your lab?

No, No.

It was hard to find someone who could help me the way I needed the help.

Do you have data that you have not been able to analyze? Why?

No, No. I didn't have that problem.

What would make it easier for you to analyze your data? (Specific resources, training, personnel)

I think for training we definitely need some training. I'm wondering if bioinformatics is a beast where biologists are trying to do something very computational-it would be nice if there was something who… For example, I am trying to run an R code-R script, and I'm having some issues and if I could come and talk to someone like you know this what I'm trying to do. I'm trying to draw a heat map, and these are the things that I want, but this is not happening. I'm trying to do a hierarchical clustering but this is not what I like. This is not the clustering that I want. How can I change the clustering?

You are supposed to be able to do that. It would be one problem less.

Right, right. Just for an example, if there are people who could help with the coding and the scripting part that would be really nice. Right… need base training would be nice-like if I am fighting with something to work, and it's not working-and then I can kind of drop in maybe something like that I don't know. And the other research scientist in our group he would definitely benefit from this kind of training.

How do you currently share data with collaborators? Do you have any challenges in sharing data with them?

No, we don't right now because how we share the data with them is we have a shared server, and we can just access that. So it has not been an issue right now.

Is there anything else you'd like to add about data analysis or is there anything else we should know?

I guess for example Qlucore or Qiagen has some is own data processing pipelines. I don't know…there is always a question with these is whether they are authentic-if they are doing everything in an acceptable way that will be accepted for publishing in, so I don't know how to measure that myself other than using it versus what I am doing and seeing what the differences and similarities are.

You mean you don't know what is going on in the background?

Yeah. I don't know how like…It will be a project in itself to clarify that.

That is with the Partek Flow? I see. Okay, Yeah. If we can do some of those similarity things with the packages that the library is making available that would be great.

Do you use resources at Yale library for your data analysis? Are you aware of all resources that we have at Yale library?

Yes, so there were…I was here for the training for the Pathways that I think you did with us. We went through David that at is good. I will come to that, so I'm using IPA. That is one of the resources, and I like IPA very much. I also came to…there was a worship for UNIX language command line-how to use that.

I came to that one as well, so I think that your training classes are very helpful. I know a lot of other people from other labs who are not bioinformaticians who liked those courses as well.

You use IPA, so know we have IPA.

Right.

Interview 7

Which department or departments are you affiliated with at Yale?

What is your role with in that department?

Graduate student.

How many people work in your lab or team?

2 graduate student and a technician. Undergrad and post doc will be starting shortly.

What type of research are you involved in?

Primarily a proteomics lab. We don't do a whole lot of genomics

Which type of data analysis is most important to you?

Don't do a lot of genomics RNA-seq or whole exome or genome sequencing we do very little of that. Most of our informatics centers around proteomics workflows

What is the sample size in your projects?

It depends on the specific project. Generally, when we do proteomics we have 10 or less samples in a given work flow or given experiment.

Let's talk about how you use data in your research. How do you store your data? Do you use Grace or some other storage option at Yale?

We don't. Right now we use a boot trapped method for data storage where we have 6, 3 terabytes hard drives we plug them into our machine and copy data on our machines then back them up to other ones. We are actively looking at moving to a different data storage system weather that be cloud based through Yale ITS or something. We need to figure that out.

What are your current practices for analyzing your data? How many people on your team are involved in analyzing the data?

At least for most projects there are 1 or 2 people analyzing the data. For example, for my experiments there is usually myself and occasionally another grad student depending on what the specific project is.

How much time does it take, on average, to analyze your data? Would you change anything if you could?

Loaded questions. Depends on the project. Quite frequently we find that we analyze the data we develop a hypothesis will test the hypothesis in the lab then quite often based on that result we will go back and reanalyze the data or look for something else based on the hypothesis and the results of the of those experiments. The first task analysis generally takes me 1-3 days based on the size of the data set. It would be nice if some of our standard work flow could be automated and you just upload and it does all the steps that you would automatically do already. Currently we don't have that capability so I do it.

What tools or resources do you use? Are these open access (free) resources or do you pay a fee?

Some tools we pay for other are free and open access. So we use the Mascot search engine which is a paid tool and we have a shared license with the tech proteomics facility. We have a couple other programs that we use. Free to use but not necessarily open source. We use the Maxquant software from the group as well as the downstream analysis package that goes along with the Max Quant that is call Perseus free to use but not open source. Also some tools from the medical library Pathway Ingenuity Analysis. We also used a little bit of cytoscape and I went to the QluCore training this week and I have not used that extensively but we are aware of it.

Do you think that may be helpful?

I don't know yet. One of the programs we use Perseus has a lot of the same functionalities like making the hierarchical clustering and PC analysis and it has the advantage of doing volcano plots. Where I can see QluCore being helpful is that it was more intuitive on making the PC plots and clusters & editing them, like narrowing what you are including in your cluster and what you are excluding. So it could be useful in that way if we as a first pass use the Perseus software to make a hierarchical clustering and then refine that further the QluCore may be useful. I don't know I would have to spend some time using it before I can give a nuanced opinion.

What type of hardware do you use? Do you have any problems with this hardware?

We generally use PC s because the mass specs software we use only run on PCs.
Most of the limitations of the hardware are RAM limitations or hard disk limitations when we are doing mass analysis. For example, the search software we use is Maxquant is a single shared memory program it does not parallelize onto high performance cluster so it requires a machine with a lot of guts. A lot of RAM and storage. A machine with a lot of storage multiple independent processors. We currently use a box with 30 gigs of RAM and 24 (2 duel threaded 6 core processor) so it has 24 theoretical independent processors. We thought about but have not done it scaling up to what is like a server rack that could have up to 128 gigs ram and maybe 60 individual processors theoretical processors that would speed up our data processing significantly especially when we are doing a lot of mass spec files in the same search.

You are at West campus. We are trying install a high performance work station here at the library for those who you in surrounding lab I think it will be a benefit. If they have to run something that a little bit more hardware. Do you think this is a good idea? This is something we are going to pilot. I want to update the library to these kinds big data. The work station we have is like the last century. What if we offer things like this platform so people could come install and run their program and get their work done and leave it running overnight.

I think that would be a good idea. The labs that you would be catering to are kind of dabbling in this a little bit. Labs that do this for a living already have their own computers that can handle this. We a computer with pretty go guts to do most of our searches. West Campus analytical core has a computer similar to what you are talking about putting into the library that is multi user. Many different labs can use that computer and it has a lot of the hardware specs to be more computational intense programs or processing. I think it would be a good idea because many labs that did not start in that space did not make the investment to have the hardware in their own lab. A core facility where they could go and access that would be great.

Are you a bioinformatician?

No I am not. I am a biologist a proteomics I am not a true as a bioinformatician but out of necessity I have learned. I would not call myself a bioinformatician.

Do you work with a bioinformatician? If so, what is your interaction with them like?

No, I do not work with a bioinformatician.

What are your biggest challenges in analyzing the data? Is there a bottle neck in your data analysis? If so, where and what is it? How do you try to get around that bottleneck?

One of the biggest challenges is we do a lot of proteomics, a lot of the tools, for example ingenuity and some of the other programs were developed for microarrays or RNA -seq. So they treat the whole gene or protein as either being up or down. But when you start to getting into PTM space and phosphor-space that is not necessarily the case the whole protein amount may not change but the amount of phosphorylation at a given site will change and there are not any good programs (they we have found) or solutions for being able to address phosphorylation being up or down separate from the protein itself being up or down. Part of it is because of the background of these programs they did not need to have that level they were not looking at protein phosphorylation or any of the other posttranslational modifications. They did not need to write that into their software. That is one of the biggest challenges I have found in our work is the programs not being able to do that.
As far as I am aware nothing exists to fill that need.
Bottleneck: just what I explained.

How do you try to get around the bottleneck?

We have tried different things, I tried to have someone help me write a custom script that would tell me in 2 different runs if I observed   phosphorylation changing and if I observed peptides from that protein. I ask the question is that protein. Is the protein abundance changing independent of … Or this the
Phosphorylation changing independent of the protein abundance or is the phosphorylation changing   with the protein abundance? This is one of the main questions. Is the phosphorylation on the same amount of protein or is the phosphorylation changing because of the amount. We tried doing that and it worked ok. One of the ways we address it by going back and doing Western

blot so if the site the site has a phosphor specific antibody and there happens to be antibody to the total protein then you can go and do a Western Bolt and see if the phosphorylation is changing independent of protein changing but that is a really really low through put way of doing it and fairly expensive way of doing it. For every phosphorylation you need to buy 2 antibodies because you have the phosphor specific and the total.

Do you have data that you have not been able to analyze? Why?

I have data I have not been able to any analyze at that level because we don't have the antibodies because we don't have the budget and I don't have the time to do all the Western Blots. Essentially what we do is narrow our list of "quote un quote" interesting phophopeptides to a reasonable. amount that I can handle then we can get the antibodies and do the Western Blots. There is a significant amount of phosphor-sites that we would have be interested in but have not been able to follow up with that level of detail because of time and expensive and all those other reasons.

What would make it easier for you to analyze your data? (Specific resources, training, personnel)

A nice interface that a biologist could use someone like me who is not a bioinformatician or not a programmer that I could come up and use that would be very useful for these types of analyses. Having a bioinformatician would also be useful to ask if I am doing the appropriate analyses and if the things I am doing are legitimate.

Is training needed at Yale? Any suggestions regarding the training? Any suggestions how we can help? How do you prefer training to be delivered? Are there others in your research group that could benefit from training? Why?

Training is ok. I have attended a couple of training that the library has put on. (QLucoore, CytoScape) Some of the other programs. Ingenuity has online tutorials that are pretty good. I had the opportunity because my PI paid for it to go to a MaxQuant Persius training but that was very useful in using those programs. Then I became the resident expert and taught the rest of the lab how to use them.

I think the training that you guys have been putting on is good. I think the training that is useful are the ones on a specific program or tool. Maybe if there was a bioinformatics for a biologist. Maybe a class? (training or class) something for proteomics specifically (we don't do RNA-seq) or once you have your list of things that are going up or down at a given condition what bioinformatics do you do after that? What statistical test are appropriate? What do you do to make meaningful PCA analyses or hierarchical clusters? Like that kind of information would be great. RNA-seq training would be great if that is something you would be interested in doing. A lot of that I assume you get sequencing data back from the sequencing center and this is how you align it and do quality control. And this is how you do all those other things that are already specific. I think having a generic like bioinformatics once you have this list of things that are going up and down it does not matter if it is RNA-seq or proteomics or metabolomics or whatever. What do you do once you have the list? That would be useful.

How do you currently share data with collaborators? Do you have any challenges in sharing data with them?

Generally, not really. We don’t share raw data so we will take the searched data and it gives an output of a text file or Excel spreadsheet and that is usually what we share. Generally, e-mail is ok I have done Box for a couple of projects. I have only used Box when someone is sharing raw data with me where they did some proteomics or aspect data and they want me to analyze their raw data they do that over Box. It is Box@yale.

Is there anything else you'd like to add about data analysis or is there anything else we should know?

I don’t think so.

Do you use resources at Yale library for your data analysis? Are you aware of all resources that we have at Yale library?

I already mentioned I use Ingenuity Pathway Analysis

Interview 8

Which department or departments are you affiliated with at Yale?

What is your role with in that department?

A liver GI physician

How many people work in your lab or team?

small laboratory consisting usually of one technician and invariably you know one to two postdocs.

What type of research are you involved in?

I have a mouse model of the liver disease that kind of serendipitously or not turned into a bone disorder. So you know it's the knock out model that we have knocks out a gene that's been associated with a variety of clinical disease. That's why I started

I wish I could say Wow.
It's interesting how it transformed into something else
You know actually how it came about? It's a well-known anti angiogenic factor that I had during my fellowship many eons ago and we created a knockout mouse and we thought we are going to get a liver phenotype and there is a little sequencing because it's become as you know affordable in the past few years they've identified this gene that I knocked out for a liver related as a cause of a rare bone disorder. Classic bone disease that every medical student learns about. Osteogenesis imperfecta type VI So it was pure serendipity. .

Which types of data analysis are most important to you?

Well that's again you know, we. try to use as many as genomic approaches that we can use given you know the financial constraints of doing this kind of research in a small lab because
(it is grant based I assume) .
Yeah I mean it's all it has to be or otherwise if you can't afford to keep the mice. Gene arrays obviously become affordable. Having said that and analyzing even basic gene arrays is actually not a small endeavor for a small time operation like mine. I have to get a collaborator actually. And you know that collaborator has his or her own project if you know project so you know I think trying to be able to do these things yourself is helpful. We have metabolomics data which you are for because you know this factor was involved in metabolic dysfunction We use the data for one paper but you know there's just reams and reams of data that we can make sense of actually that we like to analyze as well. So metabolomics ,genomics we don't have RNA seek data but we do have another project where our RNA seek data would be very helpful if we could analyze it. But anyway.

When you bring in a collaborator what types of skills are you looking for?

someone who knows how to do these you know comprehensive gene analysis knows how to go through that gene set enrichment analysis and you know then tell us these are the signature pathways that are up and so forth. But you know when you bring on a collaborator it's a huge time commitment for them and for us and you know sometimes they have their own interests at heart. You know when are at stake and so forth so you know it's hard to bring in a collaborator. And you know it doesn't make for a time when science actually.

What is the sample size in your projects?

Usually you know 6 to 10 mice per group . I will say for our genome array because you know these are kind of backcrossed and you know just when we did the gene array was much more expensive back then we did just five and five just to save on costs. But you know animal data is a lot tighter than human data. So you know it wasn't problematic but we try to keep it in you know generally from five to 10 I would say. Yeah.

Let’s talk about how you use data in your research. How do you store your data? Do you use Grace or some other storage option at Yale?

Well you know they just keep it on my desktop actually and everything is backed up these days. Yeah. Where do you back it up to the I guess has a backup for it because I actually had a receiver try something like I keep it on my external hard drive on the desktop and also keep it on the cloud. And I keep it also backed up on a separate laptop so I have redundancy I would say and do you share with others in your life do they have access to it.

Do you share with other people in the lab?

of course.

Do they get access through either your machine or through another through the ITS server
Yeah they can get access at any time you need to get give them the data whenever they want.

Do you use Grace or some other option at Yale?

I use the backup that they have that actually came in handy. It's relatively new.

What are your current practices for analyzing your data? How many people on your team are involved in analyzing the data?

Yeah I have the data on gene arrays and you know the metabolic So if I'm ok. Yeah. And whenever we get analysis we get it and then we have to do like the biochemical studies in animal studies to confirm that. So that's how I mean as a clinical doc and you know what does translational research I've essentially tried to make it somewhat clinically relevant. So this is not basic biology. This is you know disease oriented research that we do.

Are you using any type of software to do the analysis?

you know the previous Gene array for instance we did those genes that enrichment and also which is available on the web. I went to a course here about ingenuity pathway analysis that you guys have a license for.

You know it's it's ok but you know you've got to be someone who does this full time to be comfortable with it.

Frankly, I you know even though they're supposedly somewhat intuitive and somewhat user friendly I mean I tried playing with a little for a clinician who still has to do a clinic and try to manage you know overseas and who does kind
of very you know just bread and butter type of science you know staining blotting you know just add this kind of genomic stuff and to analyze it through IPA. It's not really feasible.

Am I correct in saying that there's too much of a learning curve involved to spend the time trying to become comfortable with that?

I don't think the learning curve is probably that difficult. I think it's just you have to be completely committed to one of these techniques and you have to use it all the time to be comfortable and to be confident about the data analysis and so forth I think that's what this is.

How many people are involved in your team in analyzing the data?

Well I mean the technician, the post-Docs, I analyze the data everyone analyzes the data. Remember it is a small time operation and it's not like one of those mega labs where you have a separate biostatistician this is not it.

How much time does it take, on average, to analyze your data? Would you change anything if you could?

We had the microarray data for or knock out mouse so we looked at the liver's from the five animals versus wild type to get Gene set enrichment analysis and to confirm it through multiple other you know data analysis programs. I would say it took a year to get you to feel confident about it. And so that we you know we double checked it and we confirmed that biochemically as well. So I a factor of all that time into it

Would you change anything if you could?

I mean it's obviously way too long. I don't want to rely upon a collaborator to take a year to get us through good analysis and for us to double confirm these things.

What tools or resources do you use? Are these open access (free) resources or do you pay a fee?

We've been going to you know the of these seminars that are held here. There were essentially an overview. I mean you know when I went to the IP analysis I thought I was going to get a really in-depth kind of like understanding but you know you do it once with the preceptor at the podium and you know that you're going to forget if you don't use it right away.
Do you pay for the resources to do analysis do you stick with the collaborators?.

No I pay for metabolomics I mean that's expensive to run and I pay I pay the tech to run the gene arrays and those things. But the analysis, I mean these days getting data is not the problem actually. It's data analysis that's the problem.

For a small lab like mine we need a resource who can help us with the analysis given what we know about the animals and biochemical studies. It is very difficult for us to do this alone. We need a collaborator or biostatistician who we can pay a nominal fee.

What type of hardware do you use? Do you have any problems with this hardware?

I use a Mac. No, no problems with hardware.

Do you work with a bioinformatician? If so, what is your interaction with them like?

Supposedly there's one in our section and you know what. From what I understand she is partly paid by a section like 25 percent so imagine getting you know time available for her so you know that's why I actually worked with a collaborator instead.

I never worked with them I've never worked. I have actually that's not true. We did a clinical study like for you three or four years ago we did use this guy And one of his you know people below him actually and they were helpful actually but again you know they were amenable to a collaboration of a time. Plus they don't analyze like genomics or RNA seek. They do kind of biostatistical measures I don't know if they do like you know this these new broad based genomic analysis.

What are your biggest challenges in analyzing the data? Is there a bottle neck in your data analysis? If so, where and what is it? How do you try to get around that bottleneck?

Getting the data is not the problem these days. Interpreting that data is. Is requires someone with. real expertise know who does this all the time.

Do you have data that you have not been able to analyze? Why?

Yeah, I sent one of the grad students who here we have this metabolomics that is essentially untapped we brought it here (the student) saw a library seminar and wanted it analyzed. Still a work in progress.

Difficult to analyze Why?

Because metabolomics it's a lot of material. It's not as much as genomics you know. So you know there are a few thousand metabolites that are being measured by these commercial sites plus the identifiers are relatively new. Metabolomics relatively new compared to genomics gene sampling and so forth. Even getting into one of these programs take a little bit of expertise.

What would make it easier for you to analyze your data? (Specific resources, training, personnel)

Ultimately it comes down to money. And somebody with expertise you can hire. That's in an ideal world. And that's really just sustainable for like a big plan.

Is training needed at Yale? Any suggestions regarding the training? Any suggestions how we can help? How do you prefer training to be delivered? Are there others in your research group that could benefit from training? Why?

I actually don't.
In an ideal world if I that instead of just kind of one hour kind of his show and tell type of thing. I mean that perhaps would allow you know media actually gain real expertise. But right now you know this is essentially I know the lingo. I know kind of what these programs do but I don't know how to run it. I don't know how to run the data through the software.

Multiple sessions? In person?

It's essentially like I have created an abbreviated or accelerated class for someone who can dedicate their time to that. But then you have to find time to do that.
So and what kind of training do you prefer in person online webinars?.
Well whatever works and I don't know what works for you now.

How do you prefer training to be delivered? Are there others in your research group that could benefit from training? Why?

Some of these young people who actually could probably understand so much faster than I could.

How do you currently share data with collaborators? Do you have any challenges in sharing data with them?

I email them the data. I send it to them if it's big I send it through file transfer. If it is small, I send it by e-mail.

Is there anything else you'd like to add about data analysis or is there anything else we should know?

I think the library actually has a lot of resources and ingenuity pathway for instance is actually fairly up to date. It's used by all the major med centers and even pharma uses IPA
you have you provided cutting edge tools. But again, knowing how to use the tools is a different step.

Do you use resources at Yale library for your data analysis? Are you aware of all resources that we have at Yale library?

yes

Interview 9

Thank you for helping us.

We have asked you here today because the Library is interested in how Yale researchers analyze data and the resources they use to do this type of work. We hope that by understanding these practices, the library will be better able to supply researchers with the resources and training they need to best analyze data relevant to their research. Your individual input is important and will be used to improve our services.

This interview should take about ½ hour and we will be recorded. Your name will not be associated with your answers and will be removed during analysis. You can stop the interview at any time.

Do you have any questions before we begin?

No.

Which department or departments are you affiliated with at Yale?

What is your role with in that department?

I am an attending and an Associate Research Scientist.

How many people work in your lab or team?

It varies- it depends- We have a lot of students that come through at various times. The people that are here full-time there are at least 4.

What type of research are you involved in?

Mostly basic science work in brain chemistry.

Which types of data analysis are most important to you?

EG Analysis, Proteomics with Mass Spectrometry. Those are probably the major ones.

Let's talk about you use data in your research. Let me go back. I missed one.

What is the sample size in your projects?

It ranges- The research with the rats can range anywhere between 4 and 12 with the rats per group. The humans can go up to several hundred.

Okay, Interesting.

So, we're going to talk about data now.

Let's talk about how you use data in your research. How do you store your data? Do you use Grace or some other storage option at Yale?

I store my data on my computer.

Do you have backups?

I do. There is a saved One Note drive that we also save our data on.

Do you use any Yale storage option?

No, we use the Yale computer, but no.

What are your current practices for analyzing your data? How many people on your team are involved in analyzing the data?

In terms of…

I guess one way would be what kind of resource do you use?

Some people use Excel, or they use really specific data analysis online tools.

Sure, I use Excel. There are some websites also that do some statistics. There is a great one called Social Statistics… I don't know if that's the website. I can show it to you if you are interested.

Sure.

It's designed for simple like me, so everything is very simple. It's laid out in a very user-friendly format. You can do T-tests in ANOVA and all kinds of other tests. It's basically a calculator, so you put all your data in and it recalculates everything for you, which is really nice. Then, there is another program that a colleague of mine uses. I am trying to remember the name of it, but he runs a lot of the statistics for me as well. It is not signal plot. Signal plot is the program that I use to graph most of the things that I use. It is a PC only program. I can't remember the name of it.

It's okay. If you think of it at some point…

How many people on your team are involved in analyzing the data?

All of them.

How much time does it take, on average, to analyze your data? Would you change anything if you could?

Maybe there is not an average, but…

A lot of time. I spend a lot of time analyzing data. It depends on the project.

Months?

Probably, months.

So, from the time you have the data to the time you are finished analyzing it and are ready to incorporate it into some kind of paper or...

I'll say maybe a week.

Okay.

Would you change anything if you could?

No, I mean the question sometimes is just what is the best way to go about analyzing data? Sometimes we have the answer because sometimes it is a very easy thing to compare different groups. We go out with a set question. It is a lot easier than some of the proteomics. We get hundreds of different metabolites and things. Then it's the questions of what is the best sort of way to go about analyzing the data. Sometimes it is not that straight forward. Sometimes we have to talk with a statistician about it, and at other times, we sort of figure it out on our own. We have several different people with different areas of expertise. We have a bioengineer that works with us, and he has his own ways of analyzing things. Then, the main person in the lab is a MDPHD that has a lot of experience in the area-sort of his own ways. We all sort of have our own unique ways of looking at the data. It helps.

You mentioned using Excel and Social Statistics or something that's similar…

What tools or resources do you use? Are these open access (free) resources or do you pay a fee?

The ones that I use are free. The one that my colleague use, which I will don't remember the name for at some point, I believe his does cost money.

What type of hardware do you use? Do you have any problems with this hardware?

Like what?

Is it just a PC with a CPU or…?

For conducting the experiments or…?

For the analysis.

Just for the analysis…?

Just computer-either PC. I prefer Macs-one or the other

Do you work with a bioinformatician? If so, what is your interaction with them like?

No- Statistician occasionally-the one in our department.

What are your biggest challenges in analyzing the data? Is there a bottle neck in your data analysis? If so, where and what is it? How do you try to get around that bottleneck?

In analyzing the data, a lot of it is a lack of expertise on my part, especially when it comes to statistics, so sometimes the answers that we get are different than the questions that we originally asked.

Interesting

So a lot of times we are left with a lot of interesting data that we don't know the best way to go about analyzing or even discussing with a statistician. He sort of looks at it and says, you know it would have better if you had discussed with me in the beginning phases designing the experiment… And the answer is always well this isn't the data we sort of set out to find. It sort of changed along the way. So I'd say that's probably the biggest challenge, and then obviously interpreting the data. Once you have the data, the data is the data, and like I said, a lot of times, it's not necessarily sort of what you sort of set out to ask. Then, once you have the data, it's about trusting the data-knowing that you did the experiments well. And how can you interpret it and explain it in a way that makes sense and answers the larger part of the question.

Interesting-okay.

Is there a bottle neck in your data analysis, Do you continually find it takes more time than anything else? If so, where and what is it? How do you try to get around that bottleneck?

No, I would say maybe, not even in terms of the process of analyzing the data. When we do are doing microdialysis-studies that we do, we are left with a lot of samples. Basically, we look at brain fluid and brain chemistry and I think that delimiting data, delimiting stuff better because sometimes we have thousands of samples. It is just a matter of running the samples on mass spectrometry, which we can only run 96 samples at a time. Basically, we can run one plate a day. That's like an all-day kind of thing. That is pretty much the bottle neck. It is pretty much from the time we get the samples to running all the samples. Once you have all the data that is another step from there to figure out what we want to do with it and figure out what kinds of things we want to look at more carefully. But I would say running the samples is probably the biggest hurdle.

Do you have data that you have not been able to analyze? Why?

Data that I've not been able to analyze? I don't think so. No, I don't think so.

What would make it easier for you to analyze your data? (Specific resources, training, personnel)

A statistician is helpful. We have one of those in our department.

Yeah. He is very helpful. We don't use him probably as often as we maybe should. I think a lot of the basic science stuff that we do the statistics are a lot more straightforward, so for the human stuff that I do it involves a much larger patient population that's where it gets a lot harder to be able to control for different variables and things. does a great job. That is way beyond my level of expertise. So, either maybe training for somebody like myself whose expertise in statistics is limited or just having maybe better access to somebody who does have that expertise.

Is training needed at Yale? Any suggestions regarding the training? Any suggestions how we can help? How do you prefer training to be delivered? Are there others in your research group that could benefit from training? Why?

You mentioned ideally you could be trained as one possible way to deal with analysis.

What kind of training would you be interested in specifically?

That's a good question. I mean I have taken statistics. I have event taken a statistics course here as part of the PhD program. Did I feel like it was useful? It was for sure useful. Did I feel like it gave me all the training that I needed to be able to do anything-for sure not. I guess it would be helpful to at least know better what the resources are and sort of how to access them, even somebody like Fang… I knew that we had a statistician, but I didn't sort of realize how accessible he was and the types of things he did until somebody gave me the idea to go out and seek him but the same thing with you actually. I didn't realize what kind of things you do also and that we had a librarian. I didn't really understand what that meant, until the first time that we worked together on that review. It was a huge eye-opener.

I mentioned you at one of the interviews recently.

This is sort of a side thing.

Good.

But I mentioned you at one of the interview recently to a large group because they were talking about some of the resources and things like that.

I mentioned both you and as an example that sometimes you don't know that you sort of need that resource until after you've used it, and then you realize, oh my god that was amazing!!!

Then it's something you don't have to be trained in because somebody can do a lot of it, and save your time for the things…

That's true. But I mention as an example, the systematic review I had mentioned that I have written the first time and you wrote back and said you don't want to do a systematic review find somebody who can help you do it. Then, we ended up getting it into a much better journal than the original journal that I had applied for.

Oh, really. I didn't know that.

Yeah, yeah, yeah.

Where did I initially submit it? I think it might have been to the, which is not a bad journal, but we got it into a better journal-a higher impact journal. The same thing was try with one of the clinical studies I had written. I basically did all the statistics and everything myself. It all made sense to me, and I submitted to the and again blasted me and rejected me, and then I spoke with Fang and completely re-did the statistics, and right now it is in a revision with I think they are going to take it up…hopefully soon. At least they liked it enough to ask me to revise it, so I have revised it. Again, it is one of those things I would have never sort of looked at unless I kind of failed the first time and realized that I needed the help and after the fact I realized how much help he actually was.

He's great.

How do you currently share data with collaborators? Do you have any challenges in sharing data with them?

So the One Note thing, actually I even have it open-this is something that we've been using just recently.

It looks like it's a Microsoft product?

It is.

Okay.

Yeah, Microsoft One Note, which I don't know if you've used it before-but this just an example.

No, I have not.

It's really good. This is actually one of my studies. It is password protected and shared with people that…only people who have password protected access or who have been given access to these programs. It has all the different data that we have, and you can put a lot of different things on here. The nice thing is once I add something in here, everybody gets an email that says that I have updated something, and they find out pretty much immediately, and they can go and access it. The other nice thing is that they can actually give feedback. They can write things. I am not sure how to do that. I know that one of my colleagues does it. I never actually tried to do it myself. It's a great thing. He kind of writes things on here writes things on here-take away this graph. Change this color. Do this. Do that.

You get that message?

I get it, and everybody else gets it immediately.

So, is this shared through the Cloud?

It is shared through the Cloud. It is password protected. It's shared through the Cloud.

That's good.

It's very helpful.

Do you have any challenges in sharing data?

It sounds like you don't…

We were using something before this, but I can't remember what it was. One Note seems pretty good. There are still some things on here that I have not really figured out. It is sort of user-friendly, but it's not perfect.

Is this free through Yale, Microsoft Office?

It's part of Office. If you have Microsoft Office, then you'll have this program.

Is there anything else you'd like to add about data analysis or is there anything else we should know?

Anything else you should know-like what?

I don't know.

I don't know either.

You think… I hate when I do my analysis or this works really well…or. You've touched upon a lot.

I don't think there is anything else as far as I can think of.

You may have already answered this to some degree.

Do you use resources at Yale library for your data analysis? Are you aware of all resources that we have at Yale library?

Do I use Yale resources-like what kinds of resources?

We actually have, if you go to the library's website…

Do I use any of the things on the library's website?

Our Bioinformatics Librarian has actually licensed or made available a number of data analysis tools.

No, I have not used that.

So, if you go to Basic Science Researchers in the middle of the page, so he has a number of different resources-like Biopsyche and things like that.

Okay. That's interesting.

Okay.

It's probably one of those thing that I don't really know exists.

You would have to have been wandering through a page, which I don't think people typically do.

I mean I go to the main page. I usually go for either PubMed or UpToDate or something like that.

Right, things we know are already there.

I don't think I have ever looked at any of the stuff…Clinical… I've never looked at the clinician ones.

That's alright you can call me.

So, those are all my questions.

[Interview ends]

Interview 10

Which department or departments are you affiliated with at Yale?

What is your role with in that department?

How many people work in your lab or team?

I am the PI currently one other person and another person from time to time (3)

What type of research are you involved in?

Molecular aspect aging. I am looking at genomics studies all kinds… and the motivation is to try to harvest some information from those meta-analyses to potentially come up with a therapy or some type of medical intervention in the field of aging.

What type of data analysis is most important to you? For instance, high input data, RNA/DNA sequencing, signaling pathway, gene regulatory

All of the above! Transcript analysis which is analysis of transcription factor binding site things that are related to generals that are conducting transcription armies. Can later be related to drug hunting and drug design and also to analyze things related to network analysis and analysis etc…

What is the sample size in your projects?

It varies, it runs minimum say 12 or 18 samples all the way to 40,60 It depends on the study, the specificity and what you want to do.

Let's talk about how you use data in your research. How do you store your data? Do you use Grace or some other storage option at Yale?

On my computer with backups. I backup on . I back up to an external hard drive which I use.

anything cloud based?

ANS: No I prefer to keep things close to my chest.

Do you use Grace or other storage at Yale?

ANS: I think so some of my sequence data is there in a facility.

What are your current practices for analyzing your data? How many people on your team are involved in analyzing the data?

Initial stages we look at the raw data, to put into simple terms we "normalize "it and try to standardize it. Once that is done we can look at it using bioinformatics tools.

: What types of tools do you use? Software

ANS: yeah

Which do you prefer?

ANS: they change quite frequently. I rely on those I recall the people who buffer between the biology and the bioinformatics really more mathematicians. They are not really, most of them are not biologists they bring with them computer science and bioinformatics.

Do you hire someone to do that?

ANS: We have some people around. Most of them Yale issue funding. 6:48 In order to push things through in a stingy way you collaborate with the people that have that expertise. They can assist you with normalizing data.

How do you identify who they are?

ANS: you are in the field and you know the out of the field.

So you bring some people in for some projects

ANS: The issue is noisier than what I told you. You are dealing with people so there are things beyond the science there are characters, personalities vary as well as the level of expertise. Also the level of biological knowledge beyond their expertise as to what software to use, what aggregator to use, how do we process these? So it is beyond just the mathematics because if they don't understand what the idea is behind the experiment all kinds of things may happen. In addition, the most basic is I think is the human interaction. If you have someone you can flow with you know then things can work. If you have to deal with people who bring personal issues it is not going to work.

How many people are involved in your team in analyzing the data?

So it sounds like you don't have 1 person to go to for data analysis. You may have to bring in a different person depending on the project.

ANS: Yes, and also kind of a learning curve that you go through all kinds of interactions to identify who would be the best person to do this. It is kind of an art; it is not really well established

I want to say it is primarily me. For instance, when I work with Orlando it is me and another lady but it is primarily me. I have to be on top of things and sometimes I will drive Orlando nuts because I will say no,no,no, what you just did does not make sense. Let's compare this to this and not the other way around. And really, you need to pay attention because all kinds of interesting things may come up. If you don't pay attention don't blame anyone else.

How much time does it take, on average, to analyze your data? Would you change anything if you could?

I would say weeks.

Is there anything you would change if you could?

Oh yeah. If there was some type of way to have data standardized. Let's put it this way in a A university service or something like that. I would be like some kind of cover for that. It will help me bridge the gap between me and Orlando. There would be another missing piece. Right now I am finding my solutions with other options but let's say the library would have something like that.

What tools or resources do you use? Are these open access (free) resources or do you pay a fee?

I am not paying but the lady I am using she does. Basically I am using transfect analysis stuff that Orlando Metacore, IPA, and other things I can't remember the name.

What type of hardware do you use? Do you have any problems with this hardware?

PC or MAC

University provided?

ANS: no this is mine.

Problems. Well this is getting old. The good thing about this is it has a configuration of an internal backup in addition to, so it has twin hard drives here. In addition to that I have external hardware. It is getting old and sometimes I feel that constraints coming from the University, about what kind of hardware I can purchase and I have had serious issue with that including banging heads with our business office. The University and SiQuest has some type of agreements with It was an agreement that was done once and it doesn't get updated frequently and it comes to the point now let's see now we are close to the end of the year and all those companies drop their prices dramatically and we cannot take advantage of It while we are at the University so we have to go to SiQuest and guess what? The pricing of SiQuest is more expensive than you or I could go privately put our credit card and get the same machine for less. So this is something extremely annoying and when it comes to the scares of funding it does not help.

Do you work with a bioinformatician? If so, what is your interaction with them like?

Yes

Is that what you were talking about when you have someone come in after the data is normalized?

Yeah

What is your interaction like? I guess you have talked about that they need to be more that a person who is good at statistics and they need to collaborate at work.

ANS: They need some biology background. Otherwise, just looking at things from the mathematical point of view can create all kinds of abnormalities. Not because they can't figure out biology, I don't think so. They do not have that information and knowledge and they don't know somethings happen this way and not that way. For instance, let me give you an example that really happened to me. I was working with mathematician's way back when to write a description of bone marrow and kind of elementary way just writing algorithms that represent the proliferation of stem cells. They wrote algorithms basically something that was kind of strange they said that cells can divide into n number of new cells. I asked "What does this mean?" They said "They can divide into 2 cell, 3 cells,4 cells, five cells". I said "They can only divide into 2 cells". They said "This does not make any sense." They asked "Why?" I said "Mathematics is not limited. Here in our limited field of biology this is reality." And you know we laugh about it now, but you can't believe how difficult was for them when they think very (I would say) elementary. We are laughing about it now but this was an issue. They had a hard time to go back and fix it.

What are your biggest challenges in analyzing the data? Is there a bottle neck in your data analysis? If so, where and what is it? How do you try to get around that bottleneck?

I would say the bottleneck is standardizing data. Currently.
How do you get around that?
ANS: Well you go to the person that you are with and there is a cure. Because the other person is using the same person. So it is based on first come first served.

Do you have data that you have not been able to analyze? Why?

No

What would make it easier for you to analyze your data? (Specific resources, training, personnel)

I know you mentioned a few things, having a person normalize/standardize data more than one person.
Yes this can help

Is training needed at Yale? Any suggestions regarding the training? Any suggestions how we can help? How do you prefer training to be delivered? Are there others in your research group that could benefit from training? Why?

Yes, I think the classes that Rolando is bringing I and all kinds of companies like IPA, MetaCore, ect. (19:06) The big problem that I see is the very limited time options. I really wanted to sit in on some of those class but I really couldn't make it I had to go to a conference or work. (other issue you can't anticipate) I would like to see more time options/time slots. In many cases, it is overbooked. I was on the waiting list for MetaCore and he did not sneak me in.
Was that enough training or would you prefer more?
ANS. A couple of sessions would be better but with Rolando basically after taking the training session they were more like basic introduction and after that you cannot drive by your own. Then I had some crunching time with him. (hours) and even then I know I am not going to perform it smooth as he. But I understand what he is doing which is even more important.

How do you prefer training to be delivered? Are there others in your research group that could benefit from training? Why?

I do not like Webinars in general. Webinars are aiming at the average everyone. If you know B, C and D why not skip it? Instead of starting at the beginning you should start with the information relevant to you. It is much more efficient using my time. This is something in general. With Webinars, if they are not recorded (Laugh!)
Are there others in your research group that could benefit from your training?
ANS: Yes

How do you currently share data with collaborators? Do you have any challenges in sharing data with them?

Very tricky question. It depends on the nature of the collaboration. Depends on what do we need to share? It is not an issue of trust; you trust your collaborators but there could be some things that there is no point in sharing because it won't make any difference on the other side what it is you are showing and definitely the end product would be shown. For instance, analysis of will show A, B and C and we are making it as a slide and I can share it by sending it as, or I can do Skype or screen sharing time. Show it and explain what did they see and how did they get there? They can ask me about the and the schools. Usually if you collaborate with someone (rule of thumb) you trust them, that they know what they are doing. That is why you collaborate with them. Otherwise strange things will happen.

When you are sharing data in the lab is everyone using the same gene?

ANS: No we are not sharing the same data. In the lab it is not like that. It is not really sharing like that it is more like you know processing assignments you if something needs to be processed it is kind of like an assignment. It is not like we are throwing the ball from one end to the other. It is nothing like that.

Is there anything else you'd like to add about data analysis or is there anything else we should know?

Yes, it seems like sometimes the access to medical or IPA or other things is limited to a number of users. Then you basically submit analysis and nothing really works. I know there are preferably better times of the day or evening. If we could get some type of license that would be more effective. That will accommodate more users I don't know how many users or how much demand. Maybe Roland knows the best.

: Much is based on cost

AND: A lot of it is cost based analysis. How many users does it make sense to invest in this if no one is using it? Only me? (laugh) I am sure there is some type of counter that tells you guys how many users there are, each person how much did they use through the year through the month.

Do you use resources at Yale library for your data analysis? Are you aware of all resources that we have at Yale library?

IPA and MetaCore. I would like to see more software out there. I don't know how much is overlapping with what is already there at the library but it may be possible to persuade some of those software companies to provide us with at least some kind of a trial. For example, I was using analyses which is a genomic started from the age of MicroArrays then it went through all kinds of development. It is kind of IPA but not an IPA that predicts interactions.

My biggest problem with those packages is that they are based on bias information, meaning they have their own archives and their archives are based on whatever the company people decided ???????

When I ask people IPA, medical I ask "what are the criteria's?" So for instance they are choosing articles for the top notch journals and I don't know if you know but there was some kind of article that showed the inverse correlation between the stability of the data and the exposure factor of the magazine. As sexy as the journal is, it seems like people are really trying to generate something really shiny and sexy in order to attract patients so they get in to publish…. It is kind of a way You will see in your analysis only things that they plugged in. So I don't know if this is the only way to go it is much better in the case of MetaCore when they plug in information about drugs all kinds of chemicals, drugs and inhibitors … this is good because this is kind of providing you with what is available. Some of it was approved by the FDA. These are things that will definitely help in our push forward.

Interview 11

Which department or departments are you affiliated with at Yale?

What is your role with in that department?

Clinician at, about one day per week, rest of time conduct research on genetics

How many people work in your lab or team?

Myself and three others.

What type of research are you involved in?

genomics

Which types of data analysis are most important to you?

Calling the variants from high throughput sequencing data, but this has become much easier with bioinformatics Core Facility. Now challenge is filtering those variants for one that are relevant for disease (DNA variants from whole genome sequencing).

Once you filter those variants, making sense of them from a systems point of view- network analysis , pathways analysis.

What is the sample size in your projects?

Around 300 families and like 600 controls (Families= 3 people)

Let's talk about how you use data in your research. How do you store your data? Do you use Grace or some other storage option at Yale?

A real challenge up until recently. Was storing it on network hard drives but recently has moved to the archive storage at Yale (forgot name of it). The Yale research storage- forgot the acronym. Have different tiers of storage prices and now is much more affordable. Needs to store 100 terabytes of data at a time. New pricing plan about three months ago and can now afford to store data with them.

Using the network hard drives as a backup only but eventually will exclusively use the Yale storage because it is backed up automatically. It is connected to all of the servers he does his actual data processing on.

If you store it on an institutional repository it cost money?

Yes (see above) (Have different tiers of storage prices and now is much more affordable. Needs to store 100 terabytes of data at a time. New pricing plan about three months ago and can now afford to store data with them.)

What are your current practices for analyzing your data? How many people on your team are involved in analyzing the data?

Doing gene variant discovery. Variants that make people predisposed to a disorder. Do all of their sequencing at Yale Center for Genome Analysis. They deposit their data on the servers they use so he FTPs into them. Usually use a pipeline for whole genome sequencing (developed by at Yale). His pipeline incorporates a variety of tools but just need one command to use them all. The researcher does the analysis but wrote the pipeline.

All four people on team analyze data.

How much time does it take, on average, to analyze your data? Would you change anything if you could?

All of the time I'm here. It depends. It is a moving field and there are always new ways to analyze the data. With variant discovery, doing the sequencing and making the variant calls is kind of the easy part. A hundred family project, with three people per family, could be analyzed in about a day, for instance. It is mostly automated. Can be run over night. Sequencing time takes about a couple of weeks to get the samples processed. To do the project and get the variants can take a month.

Which variants are relevant? Always new ways to filter variants. Which ones are in genes? Straight forward. Compare them to other studies in the literature. New data sets come out all of the time so contestant process of filtering your variants. You decide when you want to start. Having a tool to help with that is helpful.

What tools or resources do you use? Are these open access (free) resources or do you pay a fee?

Ingenuity Variant Analysis- he uses free trials of this. Uses library's subscription of pathway analysis but that's a step further down the analysis pipeline. Ingenuity variant analysis gets variants from high throughout sequencing and then filtering them for different criteria against different databases.

Uses SNPs and Variation Suite from Golden Helix. Purchased a license with them for past six or seven years. Very useful but very expensive so may need to drop it. Around for the year for a single seat. A nice graphical user interface but doesn't do anything he can't do for free with command line tools but packages everything nicely into one tool. Feels like he is now comfortable enough to use the command line tools which are free. Took time to reach that comfort level.

Would like us to have SNP and Variation Suite. A very nice and easy to use tool.

I should say that right now we're doing variant discovery and gene discovery, but once we find a variant linked to a disease, the next question is what do these variants do to gene expression. Getting into RNA sequencing and proteomics will be next. Hasn't done anything with that yet but having tools for that available will be useful in the future.

Recently tried Partek Genomic Suite. A desktop application (not online). Needed it for methylation array analysis. (Has chips that look for methylation) There are a lot of bugs in it but does a pretty good job. Has some features different from Partek Flow which we subscribe to. Not free. Had a free trial to try it out.

What type of hardware do you use? Do you have any problems with this hardware?

I use my desktop computer but I'm usually working through the command line that's tapped in to the Yale servers. I use Ruddel, that you use for high throughput sequencing computations. To do the variant calling you need high performance computing. All of the downstream stuff can be done on the desk top though.

Do you work with a bioinformatician? If so, what is your interaction with them like?

Not closely. Jim Knight (see above) is a bioinformatician, and can call upon him but this researcher has learned to do his own analyses.

What are your biggest challenges in analyzing the data? Is there a bottle neck in your data analysis? If so, where and what is it? How do you try to get around that bottleneck?

Storage was a big one for a while. Now can get affordable storage at Yale.

I spend a lot of time trying to figure out ways to filter variants. Having someone who knows how to do that, or tools to make that easier is a big thing.

Do you have data that you have not been able to analyze? Why?

What would make it easier for you to analyze your data? (Specific resources, training, personnel)

Tools that I mentioned would make it easier. I noticed a lot more training offered through the library and the high performance computing folks. Having the software, then someone available whose skilled in using it would be helpful.

Is training needed at Yale? Any suggestions regarding the training? Any suggestions how we can help? How do you prefer training to be delivered? Are there others in your research group that could benefit from training? Why?

Webinars that are recorded are the best. Everyone in the research group can benefit from training. If there are relevant training sessions we all go.

How do you currently share data with collaborators? Do you have any challenges in sharing data with them?

Used to be a challenge depending on what platform they were on. A challenge because they were sharing band files that were large, hundreds of gigabytes of data. They now use Globus, a parallel way to share data. Goes through their servers.
Very expensive license but west campus just licensed it to Yale. Hooks up to Ruddel server. All data can be transferred to the Globus service. Person you're sharing with needs to be on that service too, but it is becoming much more popular.

Is there anything else you'd like to add about data analysis or is there anything else we should know?

Can't think of anything. Never enough time in the day to do all of the analyses you want to do. I've taken on the role of bioinformatician and clinician so very busy. Would be helpful to have someone a little more hands on to help analyze the data. But you do have to know what you're looking for so hesitant to advocate for having someone else do it. Having the tools available, maybe having someone familiar with running the data, is helpful.

Do you use resources at Yale library for your data analysis? Are you aware of all resources that we have at Yale library?

Uses Ingenuity Pathways Analysis and Metacore.

Interview 12

Which department or departments are you affiliated with at Yale?

What is your role with in that department?

How many people work in your lab or team?

Two (three if you count

What type of research are you involved in?

Molecular biology, some of it is genomics.

Which types of data analysis are most important to you?

Low through put amounts to graphing things on graph pads and doing some sort of statistical analysis. We'll call one of the statistics guys if it's complicated. In part of a grant for the Skin Cancer has a bioinformatics section. Bioinformatics uses for analysis.

Know what he's doing with a t test or p value. Anything more sophisticated than that require assistance. Would love help to understand which statistical models (valid tests) to use for specific info.

high throughput stuff- doing whole genome DNA sequencing. Go through YCGA. They give me raw sequence data and then the question is what do we do with it? We are paying part of the salary of someone from YCGA to do the alignment to the genome and then do various kinds of test on the data. We're making that up as we go along. Funding for this person comes out of the. Also interested in some bioinformatics analysis of existing expressionary data and genomic sequencing data which is already public. Tried to teach himself some bioinformatics to do this work (including R). Wasn't doable so working with someone in California to analyze this data.

In regards to learning "R": Need to spend a lot of time working with R to be really good at using it. Can't do it part time because you need to memorize code and code changes. Finds Python hopeless.

I'm stymied. "R" may be a good choice for the library to focus on in regards to training. If you spend some time learning it, you can actually use it (esp. postdocs). Hoped there would be a bioinformatics package for amateurs, maybe like GenesSpring? Which is kind of like that. Don't have a strong sense if it pays off or if you throw up your hands and hire a biologist.

What is the sample size in your projects?

When looking at different tumors, about 100 that are sequenced. For our own genomics project, looking for DNA damage accumulating in a genome over years, about 10.

Let's talk about how you use data in your research. How do you store your data? Do you use Grace or some other storage option at Yale?

I actually don't know. The raw data is stored at YCGA. Most of our stuff, (we've been running a 10th of a lane) I think I have some of it on my computer and my postdoc has some. No backups except for some of it on an external hard drive. Because of HIPAA, there is an automated backup of desk top computers. My computer is so old that it can't handle the software that does that back up. Not sure if the department will purchase new computers.

What are your current practices for analyzing your data? How many people on your team are involved in analyzing the data?

Not even using SAS. When runs the data, I have no idea what he uses or does for the analysis. We're paying the guy to write his own code.

Only one person, in addition to me, analyzes the high through put data. Results are delivered via Excel from guy.

How much time does it take, on average, to analyze your data? Would you change anything if you could?

Before we paid the guy, it used to be three or four months. Now it takes weeks to a month. If we alter our process as we go along, it takes another few weeks but at least it moves.

What tools or resources do you use? Are these open access (free) resources or do you pay a fee?

The guy creates the software, which we use for free, but we pay for his time

What type of hardware do you use? Do you have any problems with this hardware?

On the High performance computing array. Some of it the YCGA guy does on his computer as well.

Do you work with a bioinformatician? If so, what is your interaction with them like?

Yes. The experience with bioinformatician is highly individual. Had been working with a bioinformatician assigned to me through SPORE but he did almost nothing. Eventually SPORE paid for someone else.

Had a fairly senior guy who wrote the raw alignment code special for us but then YNHH started analyzing patients' sequence data and this took priority over our research. He did our analysis correctly but then it took months because he was busy with YNHH work.

So then we hired this guy and he does the work, he shows me the data, I look it over and I can usually find some mistakes, not in the code but in the results that tell me there is some mistake. He understands some biology so we can have these discussions.

Yale in general, but med school in particular, has been a spectacular failure in dealing with bioinformatics as compared to say Harvard. There are two kinds of people; statisticians and bioinformaticians. You need three types of expertise, stats, computer science, and biology to know what you're talking about. The likely hood of finding someone with all three of these qualities is zero.

I have a test to find out who I'm talking with. If somebody asks what my model is, I know I'm talking to a statistician who doesn't know any biology because my model is what I'm trying to find out. Sometimes you can find someone who happens to know what DNA is but typically isn't a computer scientist. You meet with a faculty member and their postdoc. You'll spend an hour talking about your work, the post grad will spend a few hours on it (because they're working on a lot of other projects), you'll have another meeting, spend the first hour reminding everyone what you talked about last week; the grad students will give you results that typically having nothing to do with what you were trying to find out, you'll explain it all again and this will repeat. In about six weeks, if the bioinformatician doesn't see a Nature paper coming out of this work, they'll move on (). Everything takes more than six weeks because if you're only to work one hour a week on it, it takes six weeks. It's maddening. (Jim Noonan knows the biology. You should talk to him. He also understands the programming and computers. Used to teach a course on bioinformatics. In pathology, you might want to speak to)

What are your biggest challenges in analyzing the data? Is there a bottle neck in your data analysis? If so, where and what is it? How do you try to get around that bottleneck?

Hard to get someone to sit down and analyze the data for you. Those who are able to do it are in high demand.

Do you have data that you have not been able to analyze? Why?

Yes, I have some data form two and a half years ago and we finally gave up and sent it to a guy at BU and I haven't heard back from him yet.

What would make it easier for you to analyze your data? (Specific resources, training, personnel)

If you have a grad student or a post doc, for whom they were going to spend half of their time doing this work, then the training on certain resources might pay off. But for somebody who is going to analyze this sample now, and the next in March, it kind of doesn't pay off. Only solution is more bioinformaticians to help.

Is training needed at Yale? Any suggestions regarding the training? Any suggestions how we can help? How do you prefer training to be delivered? Are there others in your research group that could benefit from training? Why?

There is some training at Yale. You need to know how to talk to a bioinformatician. You need to have someone know what code looks like, what optimization means. Help researchers understand what the limitations are and what to ask bioinformaticians or statisticians. Training that would allow researchers to understand what statisticians can do, to allow them to communicate better with them, would be helpful. For instance, should be able to understand if empty cells are labeled as zero or left blank.

I like to be trained in person but younger guys seem happy with webinars. Problem with webinars is that you only learn the things the person with the webinar thought to include.

How do you currently share data with collaborators? Do you have any challenges in sharing data with them?

The guys I currently work with like to put things in Box. I hate using Box because you need to use a password to get in (too many steps). I want a PDF of their graph or output. Can use skype to

have a discussion about the results. Find out what their underlying assumptions were and if they were correct.
With people in the lab?: We usually sit around and work together.

Is there anything else you'd like to add about data analysis or is there anything else we should know?

Data analysis is the rate limiting thing for the 21st century and Yale has got to solve it somehow.

Do you use resources at Yale library for your data analysis? Are you aware of all resources that we have at Yale library?

I don’t know what you have as far as data analysis goes. I get emails from but I’m far too busy to check in to them.

Interview 13

Which department or departments are you affiliated with at Yale?

What is your role with in that department?

I am an associate research scientist working with my mentor What we do is basically human immune cell analysis. We try to understand how the cells in the blood function with respect of different types of stimuli that may be originating from pathogens or maybe from our own body and how that process of threat evaluation and this crisis resolution by the immune system is functioning with the aging. So we try to understand aging. So, basically, we try to do lot of omics data and we generate large volume of data. And we rely mostly on informatics core facilities to go through it, solve it and give us few nuggets so that we can use it.

How many people work in your lab or team?

Team of three, the core group, which sometimes expands to five depending on the visitors or MD-PHD fellow students which… temporarily.

What type of research are you involved in?

Correct, we routinely recruit human subjects through, People center, and the volunteers are consenting using our IRB protocol and they allow us to use their cells from the blood as well as genomic material to conduct basic research. We do not use the data to treat patients or to diagnose the patients. It is just for our understanding and interpretation of the hypothesis that we built upon.

So, basically, between genomics, transcriptomics, proteomics and biomarkers you would be more into? We almost do everything you've mentioned.

Everything?

Yeah

Which types of data analysis are most important to you?

We routinely generate flow cytometry data, which is probably not in the list of applications library provide. It is conducted in the laboratory and we process the samples using a cytometer provided at and this generates the surface expression or intercellular expression of different molecules, which can be potential biomarkers depending on their dynamics of expression. So, flow cytometry data, second is genomics data, proteomics data which are mostly done by outside core facilities either in the West Campus or in collaboration with other institutions.

What is the sample size in your projects?

We routinely recruit about one hundred human subjects every year during the October to December recruitment season because our interest to group is the prime recruitment site where we enroll our subjects.

Let’s talk about how you use data in your research. How do you store your data? Do you use Grace or some other storage option at Yale?

Data is usually stored by a central server, with specific users. And the data generated at the recruitment site is stored in the Red Cap so we have no access to it. Mostly it is operated by one who is clearance. He sees the core side identification of the subjects because we do not want to know about our subjects. Because our interest is not to treat, our interest is to understand the biology. So what we get from the subject is barcode ID. And, that barcode ID propagates to type of data that are generated from the recruitment site. So, when someone wants the demographics or any other information for publication or other, we go back to the known person who is controlling this, and then collect.

Ok, so you don’t use storage option at Yale like Grace, you just use a central server which is different.

We might go to Yale servers. It is not yet available I think. We were way back into 2006 when we started. Very recently it has expending the Yale data thing. We are still in old school but it is encrypted, controlled within the Yale network. No one can access it without a Yale net ID, and access to that network is also restricted to maybe two or three people like me, my mentor and collaborators in the informatic school. So, that is pretty restricted, it is in the Yale network

What are your current practices for analyzing your data? How many people on your team are involved in analyzing the data?

The flow cytometry data is analyzed by only one platform that is provided by. It is. It Is …..It comes build into the instrument. And the data stays in a file format called FCS (Flow Cytometry Samples). It is nothing but a text file, but it can be interpreted in different platforms of …. or Java or anything. But we don’t utilize those skill search, we rely on a sister program which is called FlowJo. It is a… Seattle-Washington based company that provides a very user friendly platform to generate high graphics, decent images of the data, as well as the tabular format to do statistics and utilize those output files to publish or communicate or whichever. So, the flow cytometry aspect of the study is processed by as well as FlowJo. And output files are PDF, or Powerpoint and a table Excel file.

How many people are involved in your team in analyzing the data?

Currently two. It is a highly-specialized software, unless you are interested in conducting large volume cell analysis, you might not be involved in it. We are a core group of three, my mentor knows about the software but he has other assignments, he does not do it but he can understand what we are doing.

How much time does it take, on average, to analyze your data? Would you change anything if you could?

Per file or per group?

Both, how much per file, how much per group do you think?

A: It takes a lot of time. The reason behind this is that the desktop platform that we use, the software itself. The data is generated in an instrument that is attached to the computer. And data flows through a server through FTP. And from the server we have access using computers and most of them are desktop computers. Each file size is about one gigabite. And processing such file takes weeks to reach at a reasonable stage of tables, statistics and images. So, if we roughly divide our time in October then submission and data analysis taking January until March, we are almost busy until September to process those hundred subjects.

Because each hundred subjects, they go into multiple assays. So, each subject generates ten tubes of acquisitions in terms of the FlowJo or in terms of the Diva. So, ten times, five to seven assays depending on what we want to ask that particular group. So we can multiply that, seventy times hundreds, that is our output size. It requests a lot of time and I would love to see something in the high throughput level, like cluster computer where you can assign the job and it processes automatically. But the bottleneck is FlowJo and Diva would not allow you to install it in the community server or in a centralized server. Because they generate revenue using a dongle or a licensed key that recognizes this computer, only. I have them but told me "Not it has been misused in the past , we generate revenue selling this USB Key, if you put that in a commonplace, many users will utilize it and we will lose revenue. So, they don't allow us to do that.

These are the two platforms which are routinely used by many researchers worldwide so we cannot just walk away from them and generate a software of our own, which can be done and it has already and there are lots of free softwares in a server.

But they are not being accepted well by the community and they have still bugs, people still….. it. So that is a very limiting step. I totally understand but nothing we can do now, in our level.

Is there a back-up if something happens to do that machine?

Oh our data is in the server. The data is also copied three times. All we are using is user enterprise and when the computational takes place, the graphic is generated, it is sent in the back in the server. We don't process to the hard drive. This cannot handle the hard drive.

Would you change anything if you could in the way you analyze your data?

That's my answering. Ideally, I think my best place to go is not Amazon, Amazon is…quite versatile but. Just 2 or 4 weeks ago, the is their platform.

So, we use 90% of the Microsoft products in academic research. Except Adobe and Photoshop and few others products, most of it is Microsoft. And Windows 10 is the universal platform accepted by everybody.

Azure is essentially you are renting a high throughput cluster computer for minutes, essentially using a credit card and paying by minutes. You are not paying for the whole year or a month, that Infrastructure. And you can take a Unix virtual computer or a Microsoft 10 Windows computer or even Windows 7, they are phasing out Windows seven, and literally mount it on a high throughput computer. And you do the same analysis in a much faster way. I don't have to learn anything, no coding, nothing required. All I do is… my virtual computer in an user account, everything is down there. But the bottleneck is the USB download. You cannot connect a USB key into the remote server, because that is where the problem is. So, I questioned the sells company, that sells Tristar as well as... I own both of their dongle because I use it routinely. I said "You have to allow me, I can give you access and you check my user accounts and how many log-in are there".

Because all I want to know is to process my data faster, I cannot wait. We have just renewed this grant. This grant is a pilot grant. The agreement with the NHS is whatever flows data we generate, it goes to a net server.

Because we are the pilot data center which will be utilized by all other centers who want to replicate our results. So, we need to process faster and also we want to use other six centers data to process. If I am requesting six centers, they have terra and terra bites of data, it is not manageable in a desktop level, it has to go a server level. That is the major thing I want to achieve in a much ideal part.

But I think that what library has done is a phenomenal amount of work. If library talk to these two companies, they would listen. They are not listening to me as a single user, but they would have to.

So, Tristar has started a very similar platform and they are the mediator between the users and Amazon So what they did because they know that data setting is a common standard now, it is globalized. It is not only by… it is users groups. So, they rented of Amazon three servers and they kind of generated their enterprise. this very software and they patch it between the users and the Amazon. And they charge you a hefty amount to utilize that service but essentially you can upload any data using their pipeline, and then do the same analysis.

But unfortunately, we cannot afford to spend that much money in a pipeline. We would rather generate our own pipeline and give access to the other pilot centers. It requests a little bit of involvement and I am trying to communicate among peer groups. We are going to… 22nd 23rd this month, the fast pilot meeting and there is a data driven specialist in this group, so we might put this there.

What tools or resources do you use? Are these open access (free) resources or do you pay a fee?

They are not free resources. It is a one-time fee. The USB key we purchase, some thousands of dollars, that we bought five years ago, and they kind of you keep you in a loop for a substantial period of time. Usually, after 5-6 years and then the person change, they tend to generate revenue and I totally understand their point of view. But they have slowed down a little bit because the public free softwares which are now available in a... What it all requests is one person who knows the coding to come up and then send box everything into a bigger package. Because all the imagining licenses are free and you can use it free and this is for academic purpose. But that person is lacking. If someone knows the code, I want to be that person but I am not a code person.

What type of hardware do you use? Do you have any problems with this hardware?

The user interface is basically a laptop or a desktop computer in every lab. But the data is in a central server and it is connecting through a remote hard drive, you know. Yale network is pretty robust and fast enough. We don't see anything lagging behind because of the storage. And for the safety as well as for the enormous volume, we don't download to a desktop or to a laptop because it is unnecessarily clogging up your instrument and slowing it down. We just rather process it and save it back and forth. All we look at is the screen, whether my plots are correct, whether my stats are correct, did I really apply, mostly checking the errors and things like that.

So, no problems with the hardware?

No, I think we have pretty robust computers. But when we are talking about 5 centers with all their data, we need to apply one common standard for every center to see instrumental variation which sometimes contribute to errors. We are not there yet, because that is where we want the cluster.

And the team where the data can be processed neither Diva nor FlowJo they are interested in. They force us to use an enterprise version which is used by … They don't mind to pay hefty fees and keep their analysis quicker. They are operating. I was told by the Tristar company that "I cannot come to Yale in the next three months because I am in training their internal staff to run our enterprise version". I said " I totally understood, don't bother you!".. He would rather give me a 50% discount one nine thousand dollars' package to buy. I said no. We are not there yet.

Do you work with a bioinformatician? If so, what is your interaction with them like?

We have a core component. It is a pilot project data. We are the site where recruitment takes place. We have a team in and recruitment volunteers and data coordinators. So, they face to face interact with the volunteers, collect their blood sample, collect all their demographics data that is stored in red cap. And then, when we process the data, that is from the flow cytometer through Diva and FlowJo, it comes out to the analysis table that goes to another informatic group.

The informatic team plays a crucial role initially to barcode everything. Then, subsequently, when the analysis of the data comes into the table, they start using big data analysis. There comes the Metacore, Ingenuity Pathway and several other Rolando has put together. of those platform, I love Metacore and that is my fast software that I am trying to learn but I am more consumed by analyzing these raw data. I would love to take 6 month time off and learn how to use these platforms but I don't have that luxury.

But Ingenuity, Metacore and the recently acquired software (Qlucore), that is a fantastic user-centered enterprise. Also, he had acquired last summer a single cell RNA seq analysis platform. I have to check the website, I forgot, I am sorry about that.

What are your biggest challenges in analyzing the data? Is there a bottle neck in your data analysis? If so, where and what is it? How do you try to get around that bottleneck?

The speed. You know we need to achieve quicker. If we are on process… January till September it is a significant amount of time. The running part using the instrument is time limiting, we totally understand, it cannot be processed faster enough. But analysis can be fastened should we have the cluster and these softwares that are available.

Would you say that there is a bottleneck in your data analysis? Is it just the time or is there something else?

Almost, mostly the time.

And how do you try to get around it, is there actually ways to do it or not?

Well, if we deviate from these two software platforms and generate our own platform, this will not be accepted globally. The real problem is that we don't want to do our own analysis and then keep it to us. Our agreement with NHS through the grant is that your data will be analyzed by XWZ platforms so that others will compare. We have no other choice.
But, these two companies can be brought together. I think Diva we don't need, FlowJo we need. There is another one, is coming up, which is a much better analysis platform. They just revolutionized… in last two years. They told me, I have checked with them, they have an … analysis platform like enterprise …… for FlowJo. But they are also willing to send their tech support, come here in the libraries server or in server, they will patch it, they will train us and they will let it run. Should we face any problem? De Novo is always there, 24h/7 days tech support. I just came to know about it six months ago that they have a six week trial period. I have not initiated it because I need really 6 weeks to go into it and try to capture it fully like I know FlowJo and Diva. I am doing FlowJo and Diva since 2001 so I know them inside out. Now, De Novo is a beautiful alternative and it is cost effective and they do not have this regulation that we will not let you do that in your server . They are willing to do it. That needs to be explored. But probably in our lab budget state, we do not want to invest that much.

Do you have data that you have not been able to analyze? Why?

We have a lag phase of…. We are still working in 2015 analysis, 2016 That is the period I am talking afterwards the December recruitment. So we have soft data or dry data rather from 2015. We are generating with data of 2016. We are pretty much finishing. By April, it should be done of 2016. We are now recruiting a new study which is a… project by the drug abusers. Those who have hepatitis A or HIV. We will recruit them throughout the year. Now… diverted because you not only have the analysis, we are also recruiting.

What would make it easier for you to analyze your data? (Specific resources, training, personnel)

Personnel, you need to be highly trained and project specific. We do routinely have visiting students, undergraduate and high school students sometimes, in their final years. They partly help in the analysis wing, not in the recruitment wing because of human subjects involved, we don't let anybody handle it. I think… , can do it. For example, if I have five desktops and five students coming this summer. I might process it faster. But that is not effective management of their time. They would be bored doing this work for six weeks in their summer project. I would rather use De Novo software, should the library helps us or we find some means to pay. Then we can assign the job, the students can look into the screen and say "Ok, here is the bug" and we ended it. So, it would be two hours in the morning and two hours in the afternoon using the students' time and enthusiasm and the rest of the time designing great experiments, which is our basic job. We don't want to waste lots of time in data analysis. We want to travel through their basic bench level things which require lot more time and we don't know how to do sometimes.

Is training needed at Yale? Any suggestions regarding the training? Any suggestions how we can help? How do you prefer training to be delivered? Are there others in your research group that could benefit from training? Why?

As I spoke to De Novo and as I've been told by Tristar, they send an expert in the very beginning when they install the software. They let us learn everything in a two-day seminar. Metacore has done it multiple times, Ingenuity even has done it in the past. Training is not effective and one can learn it quickly. And if we have trouble we can contact them. We have 24/7 tech support. They answer your questions at any time. So, that is not an issue.

How do you prefer training to be delivered? Are there others in your research group that could benefit from training? Why?

I think the platform the library operates like special group/ interest group that is a mail circulating particularly from coming to us saying here is our bioinformatics core group interest. Here is our new software coming for demo. I met Qlucore a year and a half ago in Boston in a conference. When I saw it I said "Have you been to Yale? And they said No we got to go there. So, that evening I came back I wrote to Rolando "Can you please invite them because we want to run that. And so the team came, they… it and then they gave us a free license for six months uses and I think we have already started using it. They are very happy with that.

So, the training given by companies here is ok with you.

That's fine. And then, Metacore also runs monthly webinar if you have any special interest. Metacore with the library license has a free account for every user at Yale. And I have specific questions sometimes, I write to them and their scientific support immediately answers. Even they are running a survey : how many people are interested to have a seminar again in the campus. They would send their team. I think that platform is pretty robust and helps to find ways of handling it. And also with the license comes the 24/7 tech support. You can call them if you are stuck and if you have something, a deadline is coming, you want to ….. figures to be done, email them, they swiftly answer it. I have seen beautiful support from them.

Particularly Glucore, even afford me, with just… some RNA seq data, which is not my domain of expertise but I am pretty good to learn because that data is very well labelled. So Qlucore just wrote me back saying "We can sign up a one on one two-hour web based training because you are a beginner". They know that one and a half year ago I started learning and I have not mastered it yet. They said "no you can be a two-hour session and we will provide you custom training, you can start using it".

How do you currently share data with collaborators? Do you have any challenges in sharing data with them?

It is FTP. Usually, it is in an email. That is the you know we follow ITS help. In the secure server, you need to have a net ID or otherwise you can get it.

So, there are no special challenges in the way you share tour data

No, we stayed away from Thumb Drive. The Yale cytometric core used to provide a DVD disk. That was in 2002 or 2003. That is discontinued. Most computers are set in the network. If it is not, then they allow their sister server, which is in the network. Most of the time the instrument is provided by a company which is not a Yale computer. For these analysis software, the company owns the computer and Yale does not want that computer to be in the network. Yale's basic standard is Windows 10. In a very special circumstance you get an excuse to run a Windows7. Most of these software run with Windows 7. So, in that amendment, they have another computer which is Yale owned, the data pipes into that computer and then flows to ours.

Is there anything else you'd like to add about data analysis or is there anything else we should know?

I want to see De Novo.

Got that.

Do you think that there would be other researchers who could use it then as well? Does it translate to other types of research?

A: To be honest, this data analysis is done by only. It is not that robust. If I compare it to RNA seq or gene array, it is humongous. 90 % of the Yale researchers use either gene array or RNA seq. In… center, they all use this. West campus has a core and a lot more facilities providing this. Flow cytometry used to be a small data size.: 1 mouse, 10 mice, maybe a small set. We probably… were running hundreds and thousands… by five or six years now. I don't think we will find may interest groups.

Ok, just wondering

A: We might be deeper because tomorrow if you survey asking how many of you want to use de Novo, nobody would say yes because we are so used to FlowJo. And we know so much about FlowJo,. We don't want to change. And De Novo is just two-years old. And their all… were not that impressive, that is why people got into FlowJo. Something changed. Windows 10 took like 3 years to come into market and finally Microsoft said: "Sorry we won't support any old version". But this what the pain of Microsoft keep supporting 98 to XP to Windows 7 to Windows 8 but finally saying "No, change is coming be prepared, we are moving away. I don't think De Novo is there yet to replace FlowJo but the user friendliness and what it can provide for data setting and data analysis, DeNovo is real. But library why they would invest for 5 users? They may not. But I personally think, we have a   flow core where data analysis is not part of it. The data acquisition is provided by the core immunobiology, my department of medicine is actually the provider of that core. But analysis are… individual researchers and it has not been interpreted that it requests a high-throughput cluster. No one understands that yet because no one analyzes this big size data. So it might take time and eventually it will come. Why it would come? We used to do cell analysis in a group, all blood cells as a group. But now the research suggests that we need to do cell by cell. So, essentially we have 2 to 3 million mononuclear cells in our blood. If we group them by T cells and K cells, we are talking about sizeable amount of cells. Very few interest groups are doing that now at Yale, it is called index sorting. So every cell has just one phenotype and that data set is humongous. It is enormously large size data. And then, we are trying to do genomics on every cell, which is going to be extremely large data set. When that comes, De Novo is the solution. We are going there, within a year or two, we have to be there.

Do you use resources at Yale library for your data analysis? Are you aware of all resources that we have at Yale library?

Yeah Metacore is my favorite and the next favorite is Qlucore and IPA.

And you are pretty aware of the resources that are here?

Oh yeah. I mean before I think there were some service available but he has made it to everybody. You know, we know now this is available. We know we can write to and get an appointment. He helps us If we have still some levels of difficulties, he really connect to the tech support of the provider side…then solution is done.

Interview 14

Which department or departments are you affiliated with at Yale?

What is your role with in that department?

I am a Professor of Medicine and I am the.

How many people work in your lab or team?

I have no clue. I have around 55 faculties and I don't know how many researchers.

What type of research are you involved in?

Which types of data analysis are most important to you?

We do a lot of RNA seq, we do a lot of other high throughput stuffs. We do a lot of integrative and deep learning approaches.

What is the sample size in your projects?

It is hard to say. Usually a hundred.

Humans?

Or humans or animals it depends. Of course the animal experiments are smaller.

Let's talk about how you use data in your research. How do you store your data? Do you use Grace or some other storage option at Yale?

We do you use Grace and we have a sort of hard drive for other types of data.

What are your current practices for analyzing your data? How many people on your team are involved in analyzing the data?

It depends for what. For simple statistics, I personally use Stata. Other people in the lab use Metacore or Core or whatever. For omics data, I use BRB tools, or some of the R packages. We also have a license for Gene Spring, maybe more.

How many people are involved in your team in analyzing the data?

We have the and in it there are three quantitative people or quantitative faculty. Then we have a database manager and another sort of software technician. And then many of us just do our own data analysis.

Because you know R right?

I train technicians mostly the MD fellows. They are all having everything here... Data analysis is sort of an interactive process. Although I don't generate anymore figures for my part, I usually do a run of the analysis myself regardless of who does it,

How much time does it take, on average, to analyze your data? Would you change anything if you could?

I don't really know. It is a lot of time. You know, this is sort of the unsolvable problem which is data management is more painful than data analysis. That is a very exploratory process.

Would you change anything if you could in the way you analyze your data?

I actually would like it open-ended, so I am less concerned about It though. I think for my point of view the biggest challenge is actually data management and not data analysis.

What tools or resources do you use? Are these open access (free) resources or do you pay a fee?

R is free. BRB array tool is free.

What type of hardware do you use? Do you have any problems with this hardware?

Except the fact that, you know, ITS is what it is. I am the administrator on my computer. I think it would make things much easier if ITS trained people to manage their own computers instead of needed to call. I don't have to but that's an analysis computer. We have another server that they don't handle.

Do you work with a bioinformatician? If so, what is your interaction with them like?

We have many collaborators. All my grants usually have. I have groups of colleagues in development.

How is your interaction with them like?

Yeah it's great. I have no problem. I have done this for years.

Do they all use the same resources you do?

No, no, it is very different actually. The three people that we hired are very different. So one is sort of more of an RNA sequencing person, Jen was more the network genomics genetics association and is more of machine learning. They're very different though. I tend to basically sort of ignore usually where I want to get to. And then there is the group set we develop . But yes, there is constantly a lot of computational people here.

Do they primarily have medical background?

The ones here not actually, they're… In 2002, a year after the fellowship, I already had computer science PhD students in my lab so…

What are your biggest challenges in analyzing the data? Is there a bottle neck in your data analysis? If so, where and what is it? How do you try to get around that bottleneck?

I think that data management is tough because there is no simple tool. It is almost like a limp system. There were years where people where actually spending a lot of money on developing limp system for micro array data. Then RNA seq came and we all forget about it. And that is, I think, actually our biggest risk in some ways because you don't even know sometimes where the data set are. You know the big data set, that's easy, that's a big project, a hundred of patients. But you know if someone runs RNA seq on ten mice and some samples is set somewhere. Actually I have asked the database manager to generate a sort of a simple Java base tool, just so you can see what is there and maybe classify the data. He is working on that and I don't know if he will ever finish it.

See if you might even have metadata associated with these sets.

Exactly. And then the other challenge of course is just the other side of it, which is training. So, I am not a big believer in the formal training but workshops and sort of almost tools usage workshops. I have found myself spending a lot of time, less now because the lab is.., , and other people basically going over how to use tools, where the resources are and all those things. So yes, the training and especially in the side of sort of training MD how to use some of the… Because some of… are actually really good there. They've had computational. But still you need to teach them, I don't know, what is a bioconductor or something like this.

Is there a bottleneck in your data analysis, other than what you've mentioned?

The truth is that the real bottleneck is that there is no good solution. So there is no good solution. For many of the problems we handle it is not a simple solution so …Identifying different gene expression or… is really easy. You don't need to know much. But when you are dealing with some exploratory analysis especially if it is a very therapy.

It sounds like you are able to manage the bottleneck but you wish there were an easier way.

I am not sure I care about it. I am ok with it. I like to struggle. But this is a bottleneck. I think it is harder for people who expect it to be you know "this is the data this is what you do". And I think it is much harder with RNA seq data versus quality control. And cleaning data sets actually even now requires a lot of time. And there is no one place where you could learn it. So I use to teach many years ago. There was a summer course in Jackson Lab actually in… At that time, it was called "practical approaches to micro arrays" and basically what we would do is half a day of computer lab that I would start with a data set explaining to people the concept very practically. They would go from identifying a gene clustering all these things, basically learning to use tools into finding all the resources they could find about the tool and the genome basically. And I found it actually more helpful that many of the courses people do. I think that what we are missing is a very practical user-oriented guide.

Do you have data that you have not been able to analyze? Why?

It is hard to say. You know we have all genome data that have been around multiple groups. But again DNA sequencing is not my area so I am less……

What would make it easier for you to analyze your data? (Specific resources, training, personnel)

I think we all could benefit from using more commercial tools because some of the visualization tools are much better, so I think availability of university wide licenses for commercial tools helps. Workshops in the use of these tools also help, mostly for the trainees. And the third, which is more about a safety feeling is, I know that everything we do is backed up, but actually I think we would feel, and in some ways, it would help the university, is if there was actually a repository of the dataset generated in the university. So, for instance, in theory, if I could go on the website and say ok "Did somebody sequence something?" instead of, you know…But you know we don't have the event databases of the mice in this building. So if you want to know if somebody has a…, you just send an email. In theory this should have been a sort of a simple click, right, one click.

Is training needed at Yale? Any suggestions regarding the training? Any suggestions how we can help? How do you prefer training to be delivered? Are there others in your research group that could benefit from training? Why?

I think short workshops are ideal, you know. Whatever you can come and do a sort of hands-on half day thing. And I think people don't need, you know, it is like… Computation you should learn now from biology so when we do biological assays now, half of the people don't understand the chemistry of the biology anymore right. You just buy a kit and you do it. And if it has been validated you know it works. And there is no reason why at least a first round to analysis should not be the same thing. So as long as you understand what are the comparison groups...So I think workshops help. And the shorter they are, the better they are, you know. And again, I don't know what were the… As long as it is interactive, it does not really matter.

Are there other people in your lab or group that could be interested in training?

Yeah. You know everybody has to learn some quantitative analysis when they are coming to my lab. You know we do now RNA seq, we do the single cell seq, we do cyto-analysis. All of these things are pretty.

How do you currently share data with collaborators? Do you have any challenges in sharing data with them?

So we do the usual stuff, GEO and stuff like this. For some NIH projects, I have actually a sort of data basis that we have just created. We have just created a cohort explorer on one of the dataset. I use a hard drive… So that is my thing. And actually I like the Yale box. I think it is in some ways a really good tool. It is not big enough for real raw data but you can put … So we have this way now for either paper. And basically we generate in the Yale box a sort of folder for raw data, a folder for analysis results, a folder for figures and then sort of the paper in the figures. So we actually have a trail of everything so if the journal wants to see the whole data, we just give them access to the box.

No particular challenges?

There is a lot. This is very difficult you know. And we still send hard drives. For real sequencing data, there is still no good solution.

Is there anything else you'd like to add about data analysis or is there anything else we should know?

I think it would be great if the library takes sort of more of a central role in … Basically both getting some commercial tools for both data visualization and analysis and actually training people. When I was in, I actually had a good relationship, you know, there was a guy called

Yes, and do you know as well?

Yeah, so they were very helpful.

She actually interned with us before she went to Pittsburg. We know her and we have been trying to work more with them because they have such a robust program.

Part of it was my initiative. Because we set up… There was a Yale, programming computational biology. We allowed it for them to basically get funding to do more things. And then came in and he basically just did it, and it is a good program. I think the training you get…The problem was when you put some of this training with a CT side or with a computational, it is a different approach to information.

And they have real science backgrounds like Rolando, they come out of that field so they understand. So we are trying to build something with that kind of approach. I think first some of it may be getting the funding we need to either hire more people or get the resources.

Well the question is now with the thought about the new data center, whether…One of the mandate of the data center is going to be workshops.

I don't know much about the data center.

So there is a new, I was in the committee so I think it is approved. There is data science center within the Medical school, I don't know how they are going to end up calling it. Within the medical school, there was a discussion whether they will open a department of computational biology, medical bioinformatics whatever and they decided to call it the center for something. stands in the…

Do you know if there is somebody Rolando should contact in relation to that center he can maybe start working collaboratively?

I don't really know because, you know Caroline passed away, and you know she was the. I think probably is the person to know but the problem is that his approach is exactly the opposite at mine, which is you know, of these things, these very complex things, very hard to teach anybody, right so..

And you are more "teach it"?

I would just use it. As long as you can get a computational person to look at it later and approve it or not, you know. As long as you don't go out and publish things, you have to learn it. And some of the concepts are like ok so…Does it really matter that an MD fellow does not know what is the Poisson distribution. He just knows that it is not a normal distribution and he needs to do x,w,z to handle it, that's all. So that's like this. Yeah anyway. But Mark may be, I don't know who is running the.. One idea because we had long discussions about who would do the workshops. And actually, that was my suggestion basically to go to the library and have a discussion if there is a way to do it.

Maybe we'll got it down now, so pursue that, so thank you.

Because some natural things which are like actually the extension of data and information are very… natural in the complex of the library and not that very natural in the context of, how do you say, context free analysis, which is the way bio-statisticians are trained.

This is a great opportunity for us to provide a service.

Do you use resources at Yale library for your data analysis? Are you aware of all resources that we have at Yale library?

You've mentioned Metacore.

Yeah, oh well, use this set.

And are you aware generally of all the resources that the library offers for your analysis?

I don't know unless somebody in my lab discovers them exactly.

But not you directly?

I have no clue. don't even know where to find them.

Do you want me to send you the link?

Yes.

But if you have that background maybe you can look at the resources. We have a list and you can always say "You know what, we used this in Pittsburg" and maybe Rolando can make it up. We would love your thoughts on that.

Interview 15

Which department or departments are you affiliated with at Yale?

I am at which is part of the school of medicine.

What is your role with in that department?

I am a postdoc in.

How many people work in your lab or team?

I would say 15, probably a little less, 10.

What type of research are you involved in?

In our lab we are studying the development of and it involves mostly genomics. So the study of RNA seq, DNA seq and Chip seq, which is kind of basic genomics. And we do that on human samples, so this is human genomics.

Which types of data analysis are most important to you?

So you've mentioned RNA and DNA-seq, chip seq ?

Not everyone is doing everything. In our lab we mostly produce at least the libraries and then we sequence so either the RNA, the DNA, the chromatic precipitation and we are also expanding more and more with time to different techniques.

So mostly you said RNA and DNA seq ?

Yes

Because I also had methylation studies?

No yet, it might go to but…. We are mostly on genomics not on proteomics.

And signaling or pathway, gene regulatory sequence, do you have it yet?

Gene regulatory sequence, yes, we can say that we are going to study that at some point.

What is the sample size in your projects?

My samples are from human cells in culture or post-mortem tissues. I am answering for all the lab, not for me personally because it really depends.

Do you have roughly an idea about the size of the sample?

What do you mean by the size?

How many mice, human, and cells you are working on? Do you have an idea of how big it is? It is not a problem if you don't.

Not really, because it really depends.

Let's talk about how you use data in your research. How do you store your data? Do you use Grace or some other storage option at Yale?

Personally, we are working with the Yale sequencing facility, I don't remember the name. And the output of this facility is stored in Ruddle, the cluster. So we mostly use Ruddle as a first way of storing our data. We don't have a way yet, officially at least I don't know, to store our data for the long term. For now, they are all in the Riddle cluster. And we also have access to the Farnam cluster, so we sometimes transfer the data. In our lab, we have two bio-informaticians, one is part-time, one is full-time. And they are mostly in charge of handling the data, the raw data, I would say.

It is not you directly?

I do seat sometimes because I want to trying this part but they are the ones that deal with all the raw data.

What are your current practices for analyzing your data? How many people on your team are involved in analyzing the data?

Here it really depends. For me, I have to say that I am still at the first part of my research, so I am mostly still in biology. I have not analyzed my own data yet but I am training on analyzing the data. And if you look at the lab, I would say of course bio-informatician, know how to work on a cluster with command lines tools. The others are way less trained, they are trained to use the tools that for example proposes. But most of them are not trained on those tools, so they don't really know how to use them or have time to use them. And especially when you have a bio-informatician that can do all of this work, usually you rely on him. You try to.

I can also see that some of my colleagues work with Excel files, this is quite common.

So far they don't use tools such as Metacore?

No, I try to push them to it but still no…

Maybe IPA?

All those tools, I discovered them with trainings. I am trying to advise them when I see that they are struggling. But of course, it is easier to use Excel when you are used to.

How much time does it take, on average, to analyze your data? Would you change anything if you could?

The issue of time for postdoc is really important because especially when you have biological study, when you are doing wet lab experiments, you don't really have time to analyze your data. You don't have time in the week, to put aside an afternoon to do the analysis. So that is why most of us don't really analyze the data. And I started, I have some DNA sequencing that I have to do but it is not really like I am doing this in parallel of training, I can realize that I will never have time to do it properly. Because I am always in experiments, setting-up a time aside, it can take a long long time if I have to do it only on my own. It could take weeks. Just for the fact that I don't have time to do it.

Who would do it then?

On my project it depends. So for the DNA sequencing, it is a collaboration with another lab that is analyzing the sequences.

Is it from Yale?

No it is at the Mio Clinic in……

So it could take weeks, then you can send it to another lab if you don't have time to do it?

Yes, I can. I send them and they are doing it in parallel with me. Me, I do it as a part of training. My plan is to be able to do it in the long run by myself. But the thing is that just to put aside enough time to train with the tools that you need to do. Especially since I am trying to train on the command line part. I am not using such beautiful tools that we have at the library but using the simplest command. Especially since in our lab our analysis don't require only mainstream analysis, classical analysis. But we also have to do some network biology, some more complicated analysis that require to do this by command line.

What tools or resources do you use? Are these open access (free) resources or do you pay a fee?

If I have to answer for me, I plan to use command lines, so all the tools that can be used in command lines. Mostly, it will also involve R the software for statistical analysis and some Python programming if have time. I have to say that now that I can use R, I don't use Excel anymore because the difference is…. And I am also trying to use the pathway analysis tool, IPA. But I don't have my own data yet, so it is mostly training still, it is not…..

And how do you train?

I went to most of the trainings that offers at the Yale library. And I also did some others for programming at the Yale Cluster Center, that offers some training in more raw computing such as Python and R and using the clusters, which is not easy for a beginner.

What type of hardware do you use? Do you have any problems with this hardware?

As I said for data, when you are a bio informatician, you use the cluster so you don't use your own computer but you have sometimes issues to store temporarily the data on your computer. For example, if I work with my own computer, which is my personal computer not my work computer, it is a Mac, and I don't have enough storage. If I wanted to download everything, I could not. Plus, it starts to be old so it would not have the computing power required. Most of the time, I tend to use the cluster power. But other people in my lab, they use their own computer and often they have the issue of the storage. We don't have that much hard drive or have not invested in that yet, to store our data somewhere. And we use also all the temporary boxes that you can put in the cloud. But it is more for transferring data or temporary storage.

Do you work with a bioinformatician? If so, what is your interaction with them like?

You have mentioned two bioinformaticians?

There is one who is associated so he is not really part of the lab. It is more a collaboration. The other one is part time but she works a lot. But is not enough. Half of the time when I speak with my colleague we are waiting for the bioinformatician to do something because they have so many things to do that they don't have time all of use and this is a big part. And also we often have trouble communicating with them. Especially when they are far away, it is not easy when you don't have them, when you can't talk with them on a daily basis, it is really hard to understand what they did with the data, what output they gave. And when you have small things to change and you don't understand the statistical analysis they did, it is really hard to, to process with this.

The following question was: How is your interaction with them like? Is it complicated?

It also really depends on the person and how they do to present bioinformatics data easily. But I have all the examples. The lab with whom we collaborate, they are really nice, but they are far away, so discussing with them can be hard process just because of the technical issues. Because you can’t ask questions whenever you want. And we have also an associate that works in the building and that is never there at the lab. He is kind of doing his analysis on his side and don’t really talk with us, so it is really complicated. And we have one that works directly in our lab and is really nice so it is easier. And she is improving our understanding of bio-informatics. On many issues, we are not trained in our lab in bioinformatics. You can see clearly that half of the time we don’t really know what are the analysis that we are dealing with. We know roughly what it is, what type of results we expect, but we don’t really know, it is still a black box, we don’t really know what is going on inside…

And on the contrary, do they know some bilogy?

Then again it really depends. Usually they know the biology of their specific analysis. They know about DNA. But when they are training on something like Chip seq for example, it changes, it is not the really the same process. And often, they don’t really have the reality of how long it takes to generate a sample for example in biology. So the same goes for us. We don’t really know how it is difficult to generate something.

It is something we have heard a lot, the difficulty to communicate and to understand each other because sometimes you just don’t have the same background, so you just don’t speak the same language.

For me, I would say it is easier because I have enough background in informatics and bio informatics to understand what they are saying. I have also trained a lot this last year to be able to catch up for the analysis. But I can see that for other people, it is not that easy. It is mostly because we don’t have time to train in bioinformatics, not because we don’t want to.

A good training would be to understand the basics of bioinformatics. And for that the courses that Rolando offers at the library were really nice for me.

What are your biggest challenges in analyzing the data? Is there a bottle neck in your data analysis? If so, where and what is it? How do you try to get around that bottleneck?

The biggest challenge?

I would say it would be mostly the time of course and also the synchronization with all the people that would be working on the data. Especially since we generate in our lab different types of genomic analysis. They are generated by different bioinformaticians. So if I conduct some analysis, they conduct their analysis and some other guy conduct some other analysis on other data, putting all of them together to interact with each other might be a challenge, or the biggest issue.

Is it just because of different people, but also because of the data integration and you just can’t integrate them properly?

I think it is both. Both the people are different and have different ways of working together. And the data themselves are different so pulling them together, at the third type of analysis, when you try to put different types of data together to have a bigger answer to your question. Extracting biological information from the data. The analysis by themselves would be really easy because all bioinformaticians are really good. But, it is the step to go back to biology that is usually the hardest.

For the analysis?

Yes, for the analysis. The pathway analysis for example are not… But mostly when you have different types of data to pull together.

Is there a bottle neck in your data analysis? If so, where and what is it? How do you try to get around that bottleneck?

In terms of data analysis, I don't know. Because you have to put your pipeline into place to know where is the slowest part for example. And this I haven't done it yet, so it might not be really relevant for me to answer. It is part of a bottleneck, but it is more on the general whole project. If you are in bioinformatics themselves, the bottleneck, I don't really know which part, will it be the pathway analysis, the….., Yeah, I don't know.

Do you have data that you have not been able to analyze? Why?

Not yet yeah. I think for all the lab usually you don't have time to analyze all your data so there are often things for projects on the side that we have the data, it has been sequenced for example but will never have time to analyze them. At the level of the lab it applies, this question. Sometimes, yes, we don't have the time. I have heard stories about things that we have in somewhere for a year that no one had time to analyze.

What would make it easier for you to analyze your data? (Specific resources, training, personnel)

For me it would be… As I am a beginner, trainings that can get you to work on clusters, like really basic stuffs. Because although there are some, they are not really, they are not perfect yet, either at the Yale cluster facility or here. They are really advanced tools, they are really beautiful but when you want to deal with the cluster or to run a program or software, I had a lot of trouble figuring out how they worked. Of course, it is difficult because there are lots of them, you can do different. There are so many tools that you can't really know which one to present in a course, because some people use this one, some people use that one….

What was the question again?

What would make it easier?

Training resources I guess is the most important part. And also, I haven't checked it yet but it would be good to have some kind of a forum or something where you could have a bioinformatics community at Yale but I have not checked if it exists or not because I am not really part of a bioinformatics lab.

It has not been mentioned yet in what we have come through so…

Because it would be just like…. In my former institute we had those bioinformatics journal clubs for example where each bioinformatician could present his analyses and how he conducts everything, the tools he uses, the tools he develops and stuffs like this. And also on general topics like RNA seq…

Because it was a smaller institute than here at Yale. But it was just some different people from different bioinformatics or even from labs that are mostly doing biology that were just presenting on one topic at one time. And also it would be good to have some kind of online forum where you can speak. Of course, they exist at the level of the international community. But for the Yale community there is no like bioinformatics resources, people that you could talk to if you have an issue, some kind of analysist, some expert on that that you could refer to.

It was once a month, because you don't have time to do more. But the thing is that this way you discover what your colleagues are doing, which ones to talk to if you have this kind of analyses to conduct. And although I know that at Yale it would be difficult. But at least, an online resources where all bioinformaticians at Yale are …

Labelled and you can find them if you have a problem?

Or can discuss you know, some kind of forum, some kind of community. Of course, Ia m not really part of this community yet so I don't really know if it exists.

And it has not been mentioned yet so…And the club you mentioned was at the lab level, it was not in the library?

Well we did not really have a library, but we had someone who was responsible for all the bioinformatics platform. So he was able to organize this kind of events. But I guess that at the level of the library, it could work also. Although it would be hard, because you are not truly bioinformatician here.

But I think that might know some people…

At least, I would say, bringing in the support to make this community work, like an online forum or something like this, this would be a great task.

Is training needed at Yale? Any suggestions regarding the training? Any suggestions how we can help? How do you prefer training to be delivered? Are there others in your research group that could benefit from training? Why?

Yes, definitely. For this, I would say also, try to see if there is some training that are at the level of, for graduate or undergraduate at Yale. They probably have courses in bio- informatics. True courses, and maybe, if we have access to those, to advertise them, for people in labs that could go. For example, to a statistic course or to something they want to learn. Because I guess there is already a lot of courses at Yale, that you don't really have to create them, I mean basic courses. The tools that is presenting during his courses are really advanced, I guess that it is not something he would find in a Yale course. But maybe if you want to learn about bioinformatics in general, there is maybe some kind of training that you could assist to, so…..

You've mentioned that you have been trained in R or Python, did you train yourself?

I trained myself yeah, because I had some time at the end of my PhD, that was my plan.

How did you do that?

I used online courses, on Coursera and EDX and stuffs like these, which are really good. For learning programming, I guess it is the best way. It is better than training courses like these. Because I went to those at the computing center, the training on R, Python and stuffs like these. It is not really efficient in terms of time because you are listening to someone's talk and you are not really programming stuffs, you are not really…. For those kind of courses, I think it is harder to do some kind of training in person. For programming it is really hard. But they are needed, because people don't really know them I guess.

Yeah, and is really hard because we are trying to figure out whether it would be relevant to have some classes here in R because some people said there were some trainings in the CSSSI but they complain because it is too far from the medical school and they don't want to go up the hill. And the second reason is that they said it was not linked with biology. They are giving examples which are not biology related. And they said that it was a shortcoming of the training there so we are trying to figure out whether…

It is true. And I think it would be, for example if you train in R, the training in programming itself is not that long. R is a really simple software/language so you can learn it really quickly during the beginning of a course, like expressions are really simple, at least if you have some knowledge of using a computer. So what may be more interesting is to train Bioconductor packages that are more into bioscience. No training specific on R but on R packages that are useful for…..

I don't remember its name, is it B R B?

The big one package I know is Bioconductor, it has all the packages inside. Yes R, and another kind of training that would be cool is "How to far". How to go from excel to R for example, let's say. But this would be more basic, it is just a suggestion here.

On Excel we had a class just before. A lot of people have registered and there were like 20 people on the waiting list. So I think that there is also a need for Excel trainings.

Excel to R or something else I don't know, it depends. Maybe there is this gap for… Me I don't have it because programming is not an issue, but I guess that for some people it might be hard to see why you should use R when you have Excel where everything is nice and you know how it works.

For you, I f there is a 3-hour class showing how much we can do with R in the Bioconductor field and not coding directly during the training, it would be ok in terms of advertising the tools.

What do you mean?

If we have like short classes, we might not be able to teach how to code in three hours.

No, I think that you don't have the choice of doing it more time. You can't do big classes because people will not show up for four hours, we have other things to do, sadly. So you have to do everything in less than two hours/two hours in a half. Of course, R can be learnt really quickly but it really depends on the person that is coming.

But when you advertise these kinds of things, my main recommendation is to set the level at which would be the training, like would it be really beginners, do you need to know some stuffs before you arrive or is it really advanced. Because there is nothing worse than coming to a training and realize that it is not your place because it is either too complicated or too simple. Especially when you are investing your time to trying to change, it might the biggest issue why people don't show up or don't feel like…

How do you prefer training to be delivered, group, one-on-one, webinars? Do you have any preference?

On-on-one, I won't see the point, because…. Although it is not interesting me, you can do that. Webinars, usually I won't go into a webinar because I think that I could watch a recording of these thing and it would be the same. So…

You prefer a real group of human if front of you.

Yeah, because this way you have some questions some other people might have. Of course, it really depends on the topic. And maybe also in terms of, it is not really training, but having the resources online especially when you are doing a good resources and you come up with a good training material, is to have it available online so that people can train on their own. Because as we said, we don't often have time to go to these courses even if we are interested in it. So having at least basic resources online in parallel to the courses might be really useful. You will have less people showing up at your course because if it is online, we don't like to show up. But for people this is more useful to have them.

Like a powerpoint you would be able to watch after?

Yeah, those kinds of things. If you can watch the powerpoint and having approximately the same material as when you are going. Or even recording of the courses, but this is more address.

Because, usually, sometimes people are used to sending their powerpoint after the courses with an email address, and then this is automatically generated.

Yes, but some people they don't show up because they have an experiment going on, and they want to do it in the evening for example, if they want to invest their time.

And do you think that there are others in your research group that could benefit from training?

Yes, but as I said, they don't have time like me. I barely have time to train myself so I can understand if they… On the training, most in our lab would need to understanding enough of bioinformatics to be able to interact with a bioinformatician efficiently. That would be all those kind of courses. And it also includes things like pathway analysis tools and stuff like these, because sometimes the bioinformatician would just give you a list of gene and say "ok, those are the genes that are under conditions". And then you are in front of this list of genes and I've seen some of my colleagues and they are like: "ok, what do I do with the gene". And you start googling stuffs, you start trying to figure out which gene is important. So, of course those tools like IPA can really be helpful and this is already the case since you are offering IPA, Qlucore and all ….

And maybe also as you've mentioned before, the kind of club to see a bioinformatician and see how they work, like your club.

Yeah, but this would be more advanced. This community I think would be more useful for people that are involved in doing bioinformatics. As I said for a lab that is mostly doing biological stuff, what you need is to be able to interact efficiently with bioinformaticians and what to do when the bioinformatician gives you back as results a list of genes, what are the tools that are available to you and what is the step that you should go to. So for me there is either you are addressing training for people that won't do bioinformatics or people that want to interact with bioinformatics and analyze the results from a bioinformatical analysis.

So the list of tools is for the ones that are not doing bioinformatics and the club is for the ones that have a more advanced level in bioinformatics.

I would say a way to get into the community. For me, it is two separate topics. Because there is really a gap between biology and bioinformatics. And what is good with the beautiful tools that we have like IPA and things like that is that you don't need to be a bioinformatician to use them. With a few hours training, you can understand how it works. You have enough biological background to understand everything so it is ok. Whereas, if you want to go into really complicated statistical analysis, you need a lot of training, you need a lot of command line tools, and how things interact and this is way harder to get when you are...There are two needs. Either you address bioinformaticians or you address biological people that need to ….

The presentation of the tools, such as IPA to see what is available when you want to do your data analysis the format here fits you?

I think yeah. They were good as an introduction of course. When you don't have data like me to apply directly after, it is hard. Maybe, a workshop when you have your own data, you arrive and you try to.. with for example Rolando supervising it. Like everybody is using the tool at the same time and you van have some advice on this tool. Come with your data and we analyze it as a workshop group for example, I don't know.

Come with your data is a nice name.

Is there anything else you'd like to add about data analysis or is there anything else we should know?

was talking about trying to know which type of hardware he could invest in. But I am not really sure this is necessary, at least in our lab, except from some hard drive that maybe some people need. But when you see all the clusters that are available, or all the tools that we already have online like IPA or stuffs like these. I don't really see what in terms of hard drive we would need.

You are saying that the library might not be relevant for that?

I don't really see how. Maybe it is because it is my perspective on it. On that part, I don't really know how to improve things. I am just thinking. I think this a great to have this kind of thing to improve the services offered by the library but yeah, I don't really know what could be improved. I think already making all those tools that cost a lot and are not easy to understand at first available to all the labs here is already great. Then when it is more advanced bioinformatics, there is less people interested so I don't know. It is always the same topics.

Because a lot of people have mentioned Excel and R in the interviews.

I think that the first introduction to bioinformatics would be to R. It is a nice tool and you need to go from Excel to R and it just changes your world because you can do way more easier stuffs. But there is this gap. Maybe this is the only thing that is lacking here.

Do you use resources at Yale library for your data analysis? Are you aware of all resources that we have at Yale library?

Not for the analysis by themselves. Not yet I mean, because the raw analysis is through computer clustering. But for the advanced, for the beautiful images and things like this, I would use IPA mostly. I am not really fond of Qlucore and Parteck Flow but this is because I've got the equivalent in terms of R.

Interview 17

Which department or departments are you affiliated with at Yale?

What is your role with in that department?

How many people work in your lab or team?

Eight people now

What type of research are you involved in?

In general, is immune metabolism. We are interested in knowing the interaction between the immune system and the metabolism so we study inflammation putting the animal on a fat diet or calorie restriction to know what is the role of the immune system.

R: So you use transcriptomics, proteomics, metabolomics…?

Both, transcriptomics and proteomics, and also metabolomics, it is something new we are doing.

Which types of data analysis are most important to you?

Now is RNAseq, we have been doing a lot of RNA seq in our lab, for example from the whole tissue or selected cell type, for example, there is a colleague that is working on adipose tissue macrophages, RNAseq, so transcriptome basically.

Do you use single-cell RNAseq?

No, not yet because we are not trained.

What is the sample size in your projects?

I would say, between 20 and 30 samples. Yes, because we are studying mouse model, we are not in clinical studies.

Let's talk about how you use data in your research. How do you store your data? Do you use Grace or some other storage option at Yale?

I store my data in my computer, we have a Yale computer and we have a Yale share drive. We do not do the initial part of RNAseq analysis so we do not have to use Grace- we have the Yale core basically doing it. At this point in my studies, I am not struggling with storage capacity.

What are your current practices for analyzing your data? How many people on your team are involved in analyzing the data?

For data analysis, we use Graphpad Prism, the one that Yale provide through the ITS library. It is a statistical software. We have also Signa Plot that our lab purchase or SPSS. There are these kind of software that we need at Yale because there are these kind of analyses like SPSS. FlowJo for analyzing the FACS data, and the Excel for sure, if I do not have Excel, I cannot analyze any kind of data. And then, the Ingenuity Pathway Analysis (IPA). MetaCore I am not using. IPA was useful for some of the signaling we were analyzing together. These is for the software. And then, I use some kind of online tool for gene set enrichment analysis, the online version because it is free. There is a part you can use on the internet but there is a part for download, but I was not able to download.

GSEA, I use it for some cell specific pathways. IPA is useful but sometimes it gives you a kind of too broad pathways.

It is good to know because we can maybe have a session on it. I used to teach GSEA in one of my sessions on enrichment analysis.

What I have done with GSEA is to download databases. You can put the name of two or three genes and if you want to know for example, specific pathways related to diseases, specific canonical pathways and then it downloads databases.

How many people are involved in your team in analyzing the data?

Everyone.

How much time does it take, on average, to analyze your data? Would you change anything if you could?

It depends on the amount of data that I have to analyze, just a matter of the time spent to actually complete the process but if it is something new, it takes time trying to understand how to represent the data, how to analyze the data. I think it is half the experimental part and half the analysis process, so it takes time. So, in case of changing anything if I could, maybe, more training on statistics.

What tools or resources do you use? Are these open access (free) resources or do you pay a fee?

FlowJo we pay. Sigma Plot we bought it, and as far as I know Graphpad Prism should be free for this year- I hope it continues to be free. Most of the programs that we use more often, like Word, Excel, etc. are free. And then, there are some other specific programs that we need for our lab, not everybody uses like for example, FlowJo because not everyone does immunological studies.

What type of hardware do you use? Do you have any problems with this hardware?

I really do not know.

Do you work with a bioinformatician? If so, what is your interaction with them like?

No. We do not have a bioinformatician in the lab. We have started collaboration with other people in the world.

What are your biggest challenges in analyzing the data? Is there a bottle neck in your data analysis? If so, where and what is it? How do you try to get around that bottleneck?

The most difficult part was the transcriptome analysis because I had so many genes that… I do not know the first thing to do, like the Principal Component Analysis for example, I had no idea how to do it. It would be useful to know. So we had it done by somebody else, the PCA analysis. So my first problem was that I had so many genes in different conditions that I did not know how to deal with them. So try to see whatever was…, whatever pathway like in the cell metabolism was downregulated or upregulated. I did not know what genes were involved: for example, in the TCA cycle, in the fatty acid oxidation, in the ones that are regulated in my condition. This is the big problem that I have been facing. So, with all these genes I was looking- at the very beginning- one by one but it is not really.

Yes. Principal Component Analysis is an issue. Also, the other problem was the correlation between the transcriptome and metabolome. We did use IPA but basically, what I was looking was for something that was able to give me the illustration of all the metabolic process in the cell. Like highlighting the genes and the metabolites that are upregulated in my list and upregulated in specific processes. As many genes as I can consider. For example, I have an upregulation of 10% between two different conditions and it is significant, there are for example, I would like to consider all the genes that belong to that specific metabolic process that follow a trend together with the metabolites. I would like to be able to do that so I can have a final figure that it says that this metabolic process is enriched because I have 50 genes that are significant, that follow the trend, and that I have ten metabolites, to have like the whole picture together.

Do you have data that you have not been able to analyze? Why?

For example, if I had the raw data from the RNAseq, I would not be able to analyze those data for sure. My analyses are done by the Core. Before we got the license for IPA, I was facing this problem. So we were giving our data to a bioinformatician, so he started to look at the pathways, and all these stuff. So, at the beginning, yes.

What would make it easier for you to analyze your data? (Specific resources, training, personnel)

Is training needed at Yale? Any suggestions regarding the training? Any suggestions how we can help? How do you prefer training to be delivered? Are there others in your research group that could benefit from training? Why?

For sure training and personnel. There are certain things that I can learn from the Internet but it is impossible for me even because I do not have time. So yes, for me training and personnel to analyze the data is needed. If I had time to spend for two reasons: I do not have a bioinformatics background, basically I am a biologist, I do not go further that the interface that you see on Excel. For me having a person who trains are necessary. As a postdoc, I have to do one million things per day. Most of these are the experimental, analyzing these experiments. When it comes to do these high-throughput studies, when I have done my part, trying to get the biology behind it, it is very complicated.

Any resources?

Not that I can think of.

This is something that I am trying to do. For example, you do not know what is out there that may be of help with your work. Somebody should say, this tool… when you mention the PCA analysis. So it is good to know.

It is for time reason, I mean online you can find everything and anything. Can you imagine that I learned a lot from you about Excel. When you were doing all your stuff from Excel, I also learned a lot from there.

Maybe we should have a training on Excel. We have a person here who is very knowledgeable on Excel. I can talk to him and see if it is interested- because it saves your time.

How do you prefer training to be delivered? Are there others in your research group that could benefit from training? Why?

In person. There are other that could benefit, we are on different projects but the type of experiments is the same. That is how labs go. For example your PI is interested in looking at specific field, it always use a kind of approach, so for us it has been a lot of RNAseq during this last year. We have been changing towards the metabolomics because these kind of high-throughput studies are kind of the future. So, our lab is going towards this direction basically. We use the same tools for answering several different questions in different projects.

Training in person and not huge classes, keep it small. Because sometimes I am shy if for example, there are lectures or seminars, and I have a question, I usually do not ask any question because I feel stupid because there are so many people.

It is good to know. Sometimes it is impossible, for example if it is a vendor, it will not come ten times per year. We have to have big training. But if it is

How do you currently share data with collaborators? Do you have any challenges in sharing data with them?

Now, I have this share folder with this collaborator on Dropbox

I use Dropbox. I have an issue with Yalebox, I am not able to use, I do not know. I tried to use it once but it was not as easy as Dropbox.

Have you used the file transfer tool at Yale?

No.

Is there anything else you'd like to add about data analysis or is there anything else we should know?

No, nothing else.

Do you use resources at Yale library for your data analysis? Are you aware of all resources that we have at Yale library?

Yes, maybe not all of them. I know about IPA and MetaCore because you have been sending several emails about those. There might be things that I do not know. For the things that I am using most, I know that Yale provide these kind of things, as I told you before. For IPA, it is something new and you have been very good at delivering this information. For other types of software, I have not seen this kind of communication. There might be something else, I do not know that we have.

It is important to send it several times. People do not read their emails. Sometimes you have 12 hours working and you have 10 emails from your box, you have to choose.

Sometimes I am surprised that people do not know that we have these tools. I put it on every email, newsletters. For example, is presenting tomorrow in the TAC building. I asked him to mention the tools that we have for these analyses.

Thank you so much for your time.

No, thank you.

Interview 113

Which department or departments are you affiliated with at Yale?

Molecular Biophysics and Biochemistry; Genetics; and Therapeutic Biology

What is your role with in that department?

How many people work in your lab or team?

Five

What type of research are you involved in?

We study gene expression. We are really interested in how ribosomes are made. We use these techniques to answer those questions.

Which types of data analysis are most important to you?

We did a scree where we generated high-throughput data but we used a pipeline that was written out of the West Campus in Cell Profiler. So we used that to analyze our data set. We had a big data set from that. One probably could have analyzed it in different ways if you had time and then as you know we are doing this RNAseq and we are very greatful to have Partek Flow. We sequence DNA all the time, we have some need to do some ChiP experiments but apparently ChiP-seq has a lot of false positives and false negatives, so I am not sure if we are going to do ChiP-seq but ChiP-qPCR. And of course, we pioneer proteomics here we did not do the mass expect here but we did it with a collaborator at the University of Virginia. Yeah, that is probably a problem that the library can't solve.

What is the sample size in your projects?

In this screen, we have roughly 163 hits and RNAseq only 5 of them because of the cost. We are working up a few other to do some more RNA-seq as we get a good data of the first set so… I am not sure of all of them RNA-seq is the right way to go. I do not know of the 163 how many, but there are a number of them that could be analyzed in that way.

Let's talk about how you use data in your research. How do you store your data? Do you use Grace or some other storage option at Yale?

On hard drives. We also do Cryo-electron microscopy (cryo-EM) which is a kind of modern form of structural biology and it is kind of image processing at its highest level. The library can't buy those programs and troubleshoot them. So we get them and we put them on our High Performance Computing Cluster which is called Louise and we image process from there but that is a different thing and we are probably not going to do that much in the future.

It is good to know the things that you are using and have an idea.

And so, we store all our data on hard drives. We have a lot of external hard drives. In fact all the RNA-seq data, we bought a hard drive and put it on so that we have it in our own possession. It must be in some other place.

If you store it on an institutional repository it cost money?

Louise, they have not charged us for a while. There was a time they were charging us for using the HPC and now not so much, we just get charge for the microscope time which is very expensive. … but we store it on our hard drives.

What are your current practices for analyzing your data? How many people on your team are involved in analyzing the data?

The high-throughput data is all done by RNA-seq. We also do a lot of… we run a lot of Northern blot and we quantify them and then we do a lot of statistical analysis and we use a statistical package, is the one that they try to cancel, we use it in all my papers

No, we got some email saying that it was never going away in the first place. I can find the methods… it is called… GraphPad Prism was used to calculate the means of the ratios and follow the standard deviation significance compared to the wild type determined with one-way ANOVA. This is big push from NIH about rigor, and in all our grants now we have to write what we are doing to address rigor and reproducibility. So one way is to do repetitions experiments and quantify and analyze statistically. We are kind of required to do this now so we are doing a lot more than we did in the past. So I guess that is the other kind of thing in addition to the high-throughput is totally dependent on GraphPad Prism. Because that is the only way to make the graphs, it gives you the error bars and stuff. So everybody in the lab does that.

How much time does it take, on average, to analyze your data? Would you change anything if you could?

GraphPad Prism does not take long. The RNA-seq data took, I mean people are still analyzing it but probably in a week took Partek to run. I think it is great. I do not have any complaints.

What tools or resources do you use? Are these open access (free) resources or do you pay a fee?

We use GraphPad Prism, we use Partek Flow, the IPA we have not used because we are using the GO analysis with Panther so far but we could use them. We use a series of nucleolar databases since we are interested in how axons are made and they are made at the nucleolus. People publish papers with lists of nucleolar proteins and we are just comparing our output to those. You showed us how to cluster the list.

We use Cytoscape for other things we use.

What type of hardware do you use? Do you have any problems with this hardware?

We have a combination of Macs and PCs and everybody has their own laptop and so far, no problems. My computer is running kind of slow. When we do image analysis for Cryo-EM, that is a hardware problem, you need a lot of computer time and a lot of space. There is a set of software that takes a lot of computer time called Relyon? that my student installed on the cluster but there is a trick to get it into run a little faster and someone told us what it was, so we were able to use that.

Do you work with a bioinformatician? If so, what is your interaction with them like?

As you know, this is the first time we did RNA-seq, we did Partek Flow ourselves and we gave it to the bioinformatician and he did this wonderful report of ten pages with all this graphs and charts and we could not interpret it. So we asked him to help us and we still cannot interpret it. I think he is a nice guy, he totally knows what he is doing but we much prefer doing Partek Flow because them, we can QC every step and make sure ourselves that it makes sense, and at the end, we of course know what we did because we did it.

What are your biggest challenges in analyzing the data? Is there a bottle neck in your data analysis? If so, where and what is it? How do you try to get around that bottleneck?

I think now it is becoming very easy to generate and less easy to generate results from the data. I think that we can do it because we have specific hypothesis we are testing. You can’t just publish a GO analysis of your gene list. You have to make sense out of the GO analysis. I think it is how do we cross the line between the data and results.

Is there a bottle neck in your data analysis? If so, where and what is it? How do you try to get around that bottleneck?

We need more people so they can do more analysis

I see, the technology generates more data but is only one person to analyze these growing amount of data

Right.

Do you have data that you have not been able to analyze? Why?

We are still in the middle of the one set of RNA-seq… so KF has done a fearless job with the Pacs9 analysis and LO is still in the middle of analyzing some of her data which actually requires… so two hits that we analyze by RNA-seq have the same phenotype on Northern blot, so the hypothesis is that whatever is common between them, might be the key to figuring out why they get that phenotype. So she hasn’t got around it yet but I do not know if that is because of LO or… I am not blaming anybody. It is a matter of having a lot of different things to do.

What would make it easier for you to analyze your data? (Specific resources, training, personnel)

I do not know; you do a good job with the trainings. You were fabulous getting us Partek Flow before… you know…, so we could have it. I think that one of the things you do really well is identifying things that we do not even really know yet that we need because I have never heard of Partek Flow. K, who works across the corridor just hired a bioinformatician, they do a lot of… that is a big price tag. But if that is what you do for living right? Then, that is important. So, we do not have that, we are kind of dependent on what the library can do for us and so far, it has been really great.

Is training needed at Yale? Any suggestions regarding the training? Any suggestions how we can help? How do you prefer training to be delivered? Are there others in your research group that could benefit from training? Why?

I went to your Ingenuity Pathway Analysis (IPA) training and I thought it was really good. Because you know, sometimes I think I learn really faster than most people and sometimes I think I learn slower than most people. One thing about training for sure is that you have to do it yourself. You can't just watch somebody else to do it. So when you say bring your own computer and we'll do. I think it is a good way of training. With the Partek Flow thing we just ------- and we had some issues and they kept calling, the tech support, they got help, that is the only way to do it.

We still have question like we have- I do not know if you can solve this- but we have, they gave us an RNA-seq dat set and they decided they needed a greater read depth for some of the hits so they did some more. We do not know how to compile the two files.

You can get one-on-one support webinar with them if you have a question like that.

My students asked and they did not get an answer, and then they asked the guys at and did not get an answer and there is got to be. You know. He says he is trained in bioinformatics so I told them to go and ask him how to do it, because maybe he knows because they are just in two separated files. They just need to made into one file so they can pass that through the Partek Flow, because you do not want to analyze them separately, because they are the same thing. That is the only thing that keeping L. from analyzing her data. We need to put these two things together. We do not have enough training ourselves because we are using a menu-driven.

I am also trying to get- which is very difficult- bioinformaticians to provide training in the theory of the tools. Because in order to use a tool, you need to know what are the steps that you are going to use, why, what to avoid, what to look for, but there is not enough motivation, there are very busy.

But I watched the video tutorial for Partek and I thought it was OK. The problem is that those are really adds, they are advertisement for the products instead of really training. So I had problems with that because that is not what I need. But you need tech support that can help you, trouble shoot because you know, you always run into a problem, there is always something.

This is before Partek, you see?

How do you currently share data with collaborators? Do you have any challenges in sharing data with them?

OH yeah, there is always a challenge because the files are so big. I have one collaborator in Holland that we do it by Dropbox, I know Yale does not approve that.

No, Yalebox does not work. I use Yalebox for my training grants because my staff won't use anything but Yalebox but it does not work. I use regular Dropbox that I pay for with my collaborator in Holland.

Yes, I use file transfer tool every day for some of my lectures to be posted. For small files, we just mail them back and forth. What people tend to do is to send something in PowerPoint because it makes it smaller, so you look at the figures in PowerPoint and they have the text. Yes, we need some way, because Internet has made collaboration much easier, because you know, we can be in touch with anybody in the world. But it is harder to share the figure which of course is critical.

Is there anything else you'd like to add about data analysis or is there anything else we should know?

You what I wanted to mention to you, it is EndNote. I know you are interested in data analysis but none of us can write a paper without EndNote.

We have thousands of sessions on EndNote.

I know, but I am just saying that is critical. None of us can write a paper without it. The first thing I do when I get an undergraduate- you should use EndNote, figure out how to do it right now.

We have every other week a session on EndNote and if you go to the library and say that you have issuses with EndNote, somebody is going to help you.

He is unbelievable, he is fabulous!

Do you use resources at Yale library for your data analysis? Are you aware of all resources that we have at Yale library?

I told my students today, because I teach medical students. There are always things that I do not know about the medical library, and here [pointing to the computer screen] in highlighted resources, for biochemistry we use exam master as a practice data base, which is still here because you guys still pay for it. We use that, I told them how to get there because it is not totally obvious. Click here, click here, click here. I am certain that there are many things I do not know about.

If you go to Basic Science Researchers, everything that we offer is there, and then all the training are there, you see, Managing your References with EndNote, Basic Unix Command, Enrichment Analysis.

I am having a hell of a time with myNCBI. When I submit a grant I do not have a MyNCBI link, I have to do it with the search function on PubMed. I am also head of a large training grant and so all the pubs from all the students that do not have my name on it are associated to it. None can figure out how to separate things so I can have my own publications from my lab, there are some filters, you see?

You are the PI but somebody publish and they do not tell you, then you receive a notification because you are the PI of the grant -telling you that you are not in compliance.

I know, I have been down this road and I find these people and I send them emails, because we can lose our training grant if they do not get it in compliance. It used to be Denise, but I do not know who to send them to now. We are just about to start that face of the year where we have to get in compliance again.