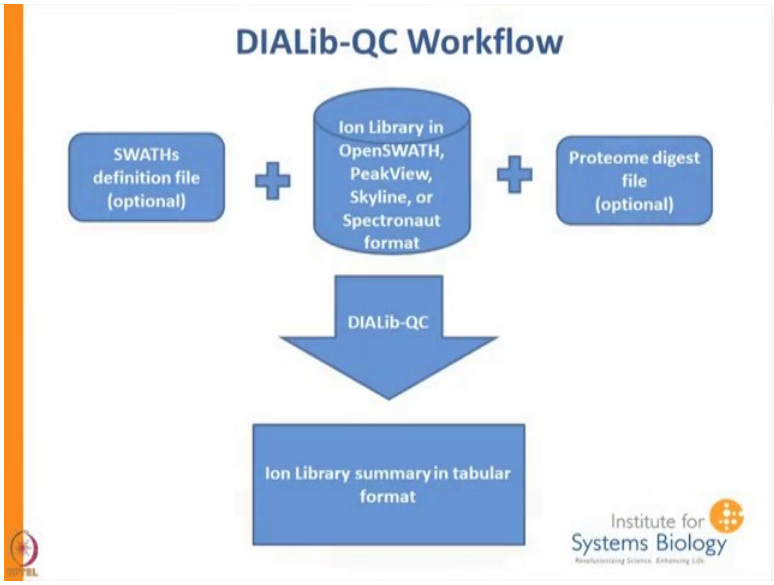
Introduction to Proteogenomics
Dr. Sanjeeva Srivastava
Mr. David Campbell
Department of Biosciences and Bioengineering
Indian Institute of Technology, Bombay
Institute for Systems Biology, Seattle

Lecture - 35
DIA-SWATH Atlas - II

Welcome to MOOC course on Introduction to Proteogenomics. In the last lecture Mr. David Campbell started discussing about SWATH atlas, and various comparisons of DIA versus DDA methods. You were introduced to the concepts of DIA and the softwares and tools available for analyzing DIA datasets. The SWATH Atlas contains high quality ion libraries for use of SWATH or DIA experiments.

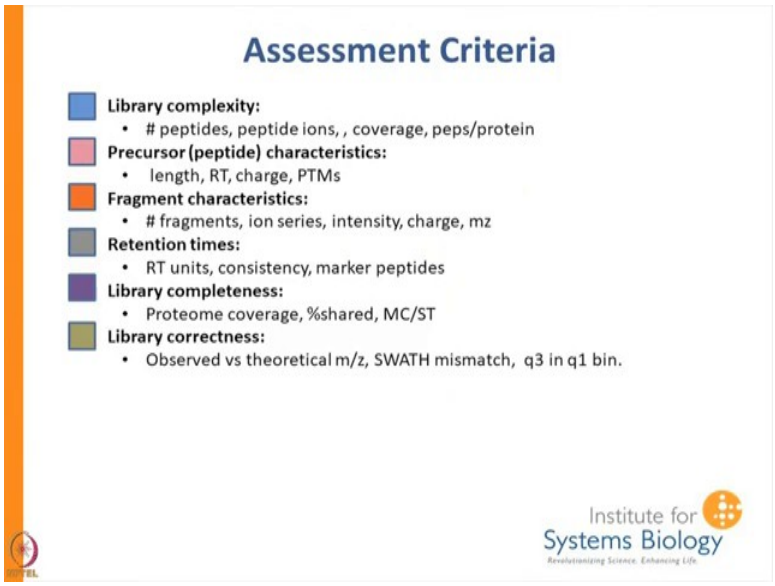
In today's lecture you will provide an overview of the features available on SWATH-MS and how you can utilize this valuable resource for analyzing your mass spectrometry data. So, let us welcome again Mr. David Campbell for his today's lecture.

(Refer Slide Time: 01:13)



So, this is DIALib-QC basically what we have is we have an ion library, we have in a variety different formats.

(Refer Slide Time: 01:43)



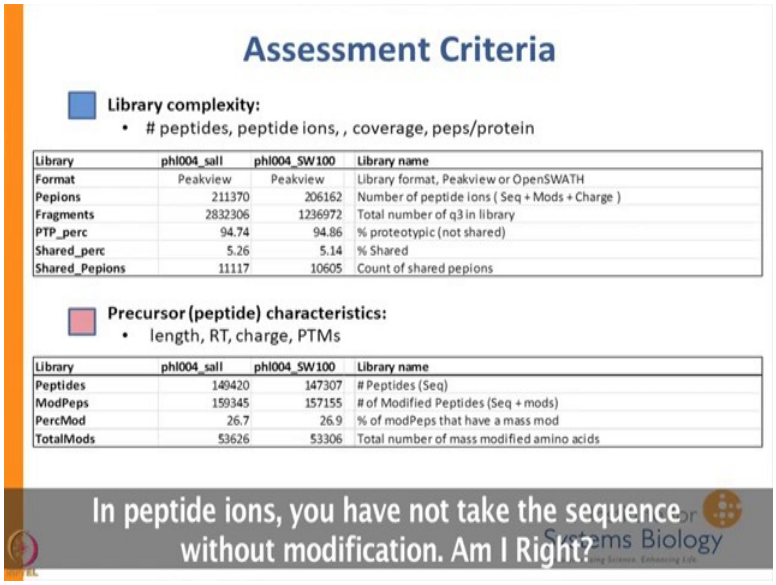
We have a SWATH file and the SWATH file basically says ok, I am going to look from this mass to this mass and I am going to you know basically the width of each bin. And you can also compare to a proteome, so basically you have taken a proteome and you digest it and you have all the tryptic peptides. You basically one this thing and you get this ion library excuse me, ion library sum summary in tabular format.

And so these are the different criteria, library complexity is sort of how big the library is how many peptides, how many peptide ions, there is precursor information so, what is the average length of the peptides in the library, how many modifications are there; what kind of modifications. Fragment characteristics you know what ions we have b or y, how many fragments you have peptides, retention time is a very important part of any sort of targeted or scheduled type analysis, you do not want to look over the entire retention time range.

So, how much does it vary is consistent, do you have marker peptides where these things called iRT peptides. Library completeness you compare to proteome how many of the proteins in that proteome do I cover. And library correctness, so it turns out that there are certain things with relation to the SWATH file or relation to the actual m/z that are being reported or is this library and in fact correct. So, basically I am go through these very briefly and so I have two different libraries depicted here. So, this is the PHL the Pan Human Library and this is again taken from all these different experiments, everything

was run on a pretty modern instrument the AB SCIEX TripleTOF 5600 and you can see that there is some 211000 peptide ions.

(Refer Slide Time: 03:04)



Library	phl004_sall	phl004_SW100	Library name
Format	Peakview	Peakview	Library format, Peakview or OpenSWATH
Pepions	211370	206162	Number of peptide ions (Seq + Mods + Charge)
Fragments	2832306	1236972	Total number of q3 in library
PTP_perc	94.74	94.86	% proteotypic (not shared)
Shared_perc	5.26	5.14	% Shared
Shared_Pepions	11117	10605	Count of shared pepions

Library	phl004_sall	phl004_SW100	Library name
Peptides	149420	147307	# Peptides (Seq)
ModPeps	159345	157155	# of Modified Peptides (Seq + mods)
PercMod	26.7	26.9	% of modPeps that have a mass mod
TotalMods	53626	53306	Total number of mass modified amino acids

So, again a peptide ion is a primary sequence plus modification plus charge and there are you know almost 3 million fragments. So, the other thing I have is I basically took this phl and I applied our SWATHs so, now, because they used they had a when they did DDA, they were looking at wider m/z ranges. In addition we were looking at these 100 SWATHs and so we do not want any, any fragment ions in our precursor windows.

And much more we actually took the top six fragment, so we want to specifically take just the top 6. And so you can see that going from here to here, we did not lose that many ions. So, by applying maybe some more stringent mass filters we did not lose that many ions, but if you look at the number of fragments, we have way fewer fragments almost you know something like 35 or 38 percent only of the fragments. And, that is because now we only have up to 6 fragments per peptide, and actually I have got ahead of myself here.

Student: So in peptide ions, you have not taken the sequence without modification. Am I right?

So, so.

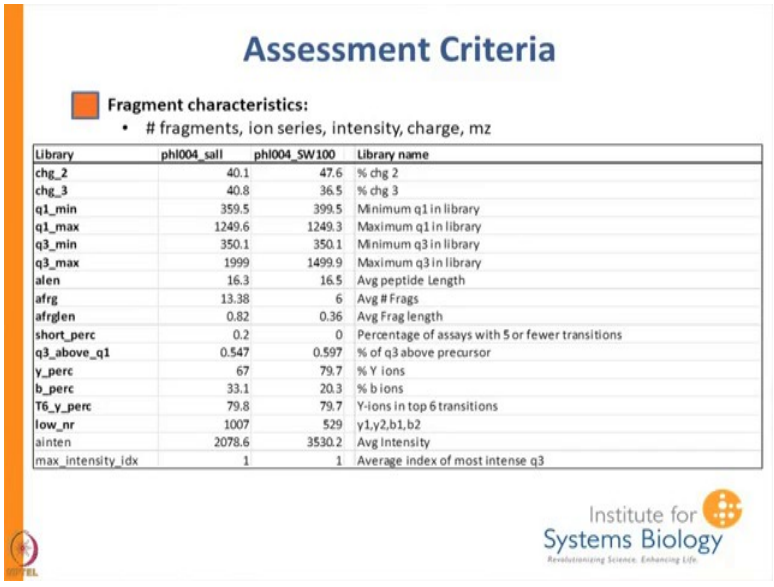
Student: You have mentioned sequence plus mod mods plus charge.

Yes. So, I have confusingly I have put it down in the section. So, this is the number of stripped peptides we would say, so this is just the number of total peptides and again these two numbers are pretty close. But if you look at this number 149 up to 211, that are those are things that we saw say both oxidized and un-oxidized or multiple charged states of the same peptide that is actually pretty common especially plus 2 and plus 3.

So, yeah so again the peptide characteristics are how many peptides did we see, what sort of modifications, what percentage of the peptides are modified things like that. This is a big one, so basically this is in the number of fragments and as I told you; so this is the PHL the average number of fragments is about 13 whereas, the new library its exactly 6. And that is because we wanted to know we were comparing these different softwares and we wanted to ensure that all the different software used the same peaks to do their analysis.

And so most of them use 6 fragments by default and so we limited our library to 6, so that they would all be on footing, but you can see the percentage in charge 2 and charge 3.

(Refer Slide Time: 05:49)



Library	phl004_sall	phl004_SW100	Library name
chg_2	40.1	47.6	% chg 2
chg_3	40.8	36.5	% chg 3
q1_min	359.5	399.5	Minimum q1 in library
q1_max	1249.6	1249.3	Maximum q1 in library
q3_min	350.1	350.1	Minimum q3 in library
q3_max	1999	1499.9	Maximum q3 in library
alen	16.3	16.5	Avg peptide Length
afrg	13.38	6	Avg # Frags
afrglen	0.82	0.36	Avg Frag length
short_perc	0.2	0	Percentage of assays with 5 or fewer transitions
q3_above_q1	0.547	0.597	% of q3 above precursor
y_perc	67	79.7	% Y ions
b_perc	33.1	20.3	% b ions
T6_y_perc	79.8	79.7	Y-ions in top 6 transitions
low_nr	1007	529	y1,y2,b1,b2
ainten	2078.6	3530.2	Avg Intensity
max_intensity_idx	1	1	Average index of most intense q3

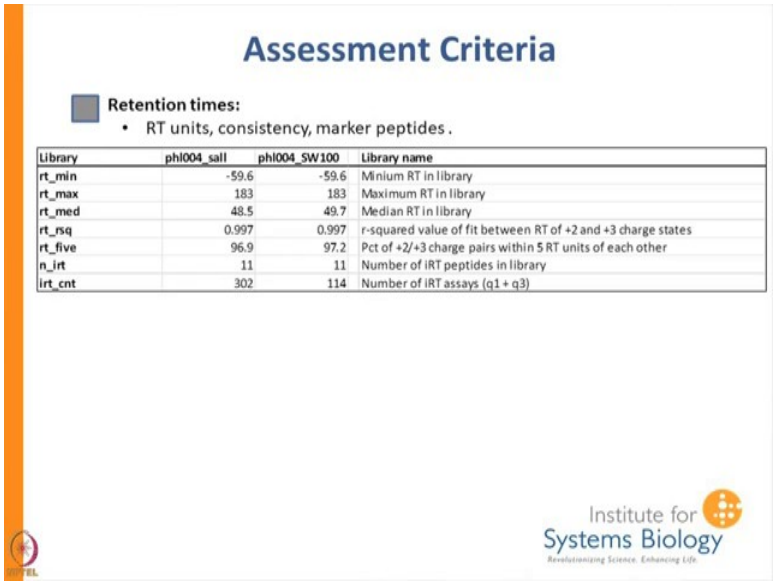
And you know sometimes going from the all to the modified library, you do actually see a swing in the percentage and you know I do not really know why that would be, but you know you might expect you see fewer charge 3, because you know basically m to z

might fall below now the lowest m to z value that were looking for, because the something divided by 3 is obviously smaller than something divided by 2.

But, this is it is a useful sanity check to compare to what you are expecting or the mass ranges in my library what I would expect. So, here is a difference, so you can see that in their analysis they had Q 3s fragment ions up to 2000 whereas we only took fragment ions up to 1500, because that is what were looking for in our SWATHs if you not that much above it, so we do not bother with it. We just we just look up to 1500, because again we want to do our cycle time, so that we can sample across the peak.

See, yeah there is all these different things and they probably I mean I am showing a theory, guys in the back cannot even see what it is, but if you trying to assess the validity of your library, it is actually very helpful.

(Refer Slide Time: 07:00)



And we have had a number of occurrences where, a this has helped us a lot. So, again retention time is very important for analysis. So, we are using something called iRT peptides, in iRT peptides are peptides you are meant to put in what is a pretty much all of your runs and they help you to register the retention times between a runs.

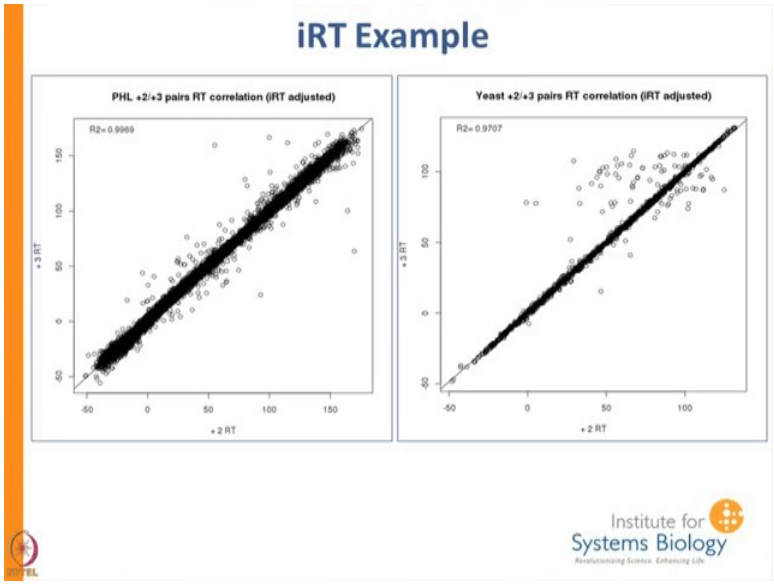
And it just so happens for some reason, they there they go from negative to positive. And so the reason that these show up as negative value from negative 60 to 180 is basically, because they have been converted into what we call iRT space, so that is a little unusual.

But it is a dead giveaway if you are looking at a library that you know, if these ranges do not make sense, then it is a red flag to look at something.

So, one other thing there is a measure of consistency. So, as I am as I mentioned sometimes you see multiple charge states of the same ion and it turns out when you make a spectral library, you if the software treats everything independently. And so you end up getting potentially different charges for your plus 2 and plus 3 of the same peptide, now that does not make any sense from the perspective that.

When your peptides eluting it is not like your plus 3, elutes here and your plus 2 loops here, really the peptide elutes; and then you do your ionization and the ionization is a either causes a plus 2 or plus 3. So, you would hope that your plus 2 and plus 3s would be very close to each other, but it turns out that they are not.

(Refer Slide Time: 08:48)

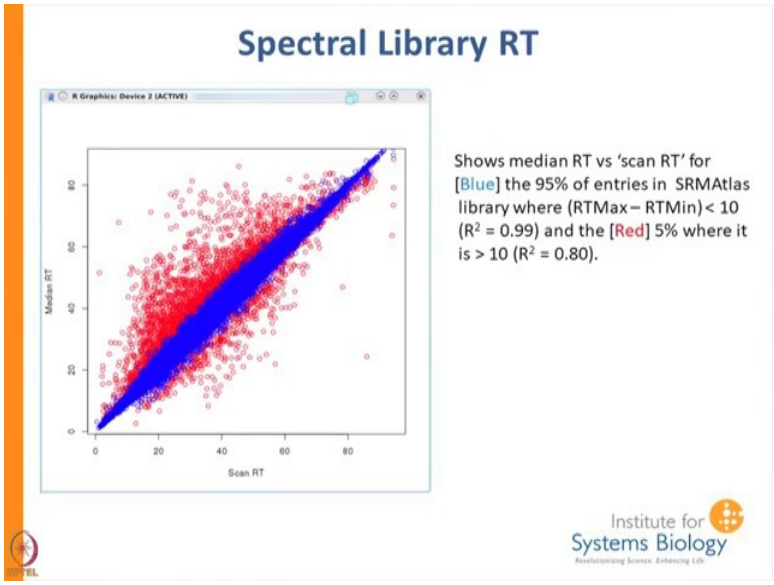


And, in some cases we found libraries where these numbers are very different. So, here two libraries that are in the SWATH Atlas and basically oops, sorry. And, so looking at is plus 2 and plus 3 charges and so this one has a very good correlation. So, having used the iRT peptides everything is pretty much right along the diagonal. So, this is basically we taken all the peptide ions that you have the plus 2 and the plus 3 for exactly the same modified peptide and plotted them against each other.

So, you would expect in a perfect world they just lie down on this axis, but you see that there some outliers and the width of this is kind of you know kind of big. So, you know you could look at this and say well you know that is actually you know plus or minus 5 minutes or plus or minus 3 minutes something like that, so that can help you to decide how wide of a; wide of a sort of window that you look at when you are trying to look for these things.

So, this is a yeast library and actually its got a pretty strong correlation as well. Everything wise, a little bit better along the axis, but we have sort of more pronounced outliers. So, this is just an example of how iRT can vary and it illustrates why we might look at this, in our libraries.

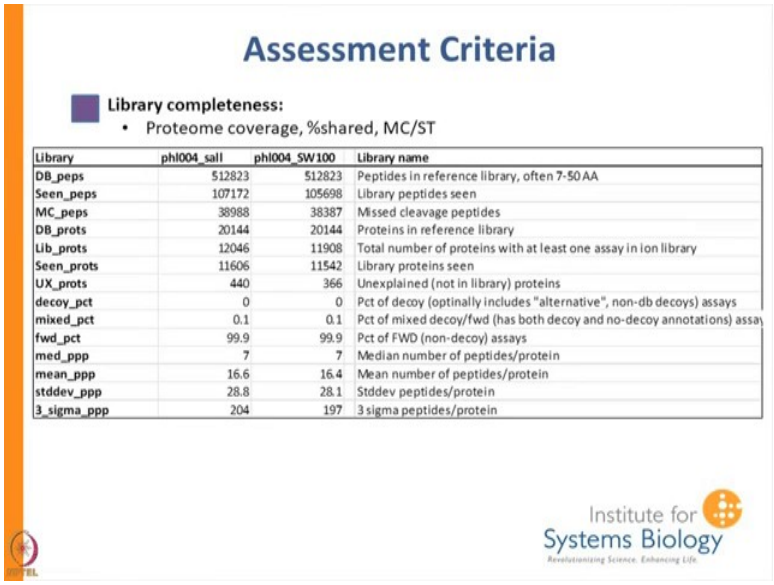
(Refer Slide Time: 10:02)



So, this is sort of another thing. So, basically in addition to storing the retention time these libraries have what they consider to be the best spectrum. So, we have taken you know between 1 and a 100 different spectra and made the best consensus that we could. So, spectraST will pick one peptide, then either has the highest signal or has the lowest signal to noise depending on how you have it set up.

And so all this shows is that if you pick pairs of peptides in the blue, where basically the retention time the difference between the minimum retention time, seen for this ion and the maximum retention time, seen for the ion is pretty close; then you get a much better correlation between the median retention time and this best replicate spectrum.

(Refer Slide Time: 10:59)



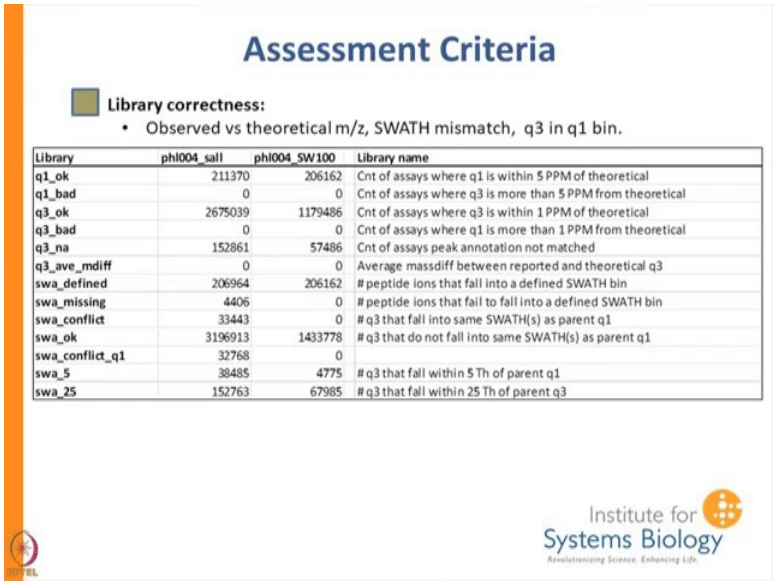
Library	phl004_sall	phl004_SW100	Library name
DB_peps	512823	512823	Peptides in reference library, often 7-50 AA
Seen_peps	107172	105698	Library peptides seen
MC_peps	38988	38387	Missed cleavage peptides
DB_prots	20144	20144	Proteins in reference library
Lib_prots	12046	11908	Total number of proteins with at least one assay in ion library
Seen_prots	11606	11542	Library proteins seen
UX_prots	440	366	Unexplained (not in library) proteins
decoy_pct	0	0	Pct of decoy (optinally includes "alternative", non-db decoys) assays
mixed_pct	0.1	0.1	Pct of mixed decoy/fwd (has both decoy and no-decoy annotations) assa
fwd_pct	99.9	99.9	Pct of FWD (non-decoy) assays
med_ppp	7	7	Median number of peptides/protein
mean_ppp	16.6	16.4	Mean number of peptides/protein
stddev_ppp	28.8	28.1	Stddev peptides/protein
3_sigma_ppp	204	197	3 sigma peptides/protein

So, library completeness is basically taking a proteome. So, most of you if you are working with some sort of organism, you have a proteome or a genome that you want to compare too. And this just looks at those things and says ok, I have of the things that I might expect to see, of the peptides I might expect to see, of the proteins I might expect to see, how many of them do I actually see in this library and it gives you an idea of how complete the library is.

And for most of these things, that the two different libraries even though we have cut this one down, do not really vary that much. One thing that is kind of interesting to notice is the average number of peptides per protein is 7, but the median peptides per protein is like 17. So, it turns out that the PHL is pretty over represented in some proteins are over represented you have up to say 100 peptides for a protein; including some things that are semi-tryptic or have missed cleavages.

So, normally for SRM you would not want to use that type of peptide, you want it look sort of the best representative peptides for each protein, so that is one thing we might want to do with the PHL is narrow it down a little bit, and I will get some more evidence of that later.

(Refer Slide Time: 12:05)



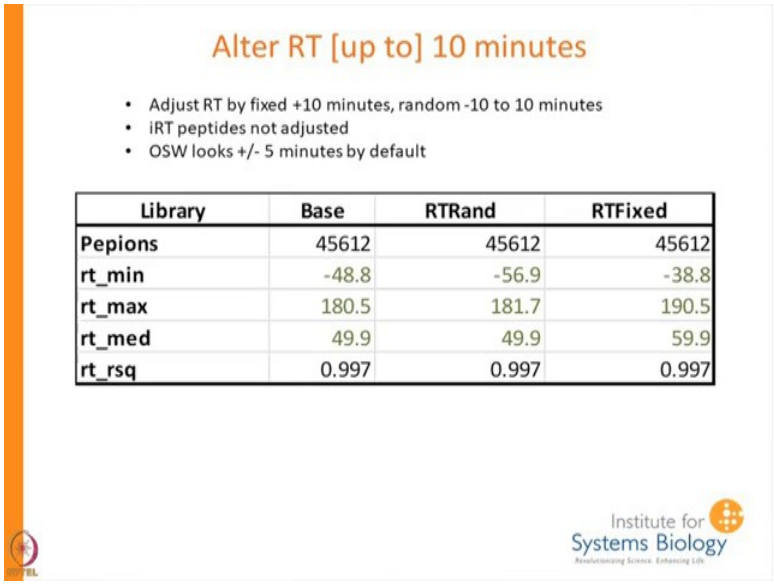
Library	phl004_sall	phl004_SW100	Library name
q1_ok	211370	206162	Cnt of assays where q1 is within 5 PPM of theoretical
q1_bad	0	0	Cnt of assays where q3 is more than 5 PPM from theoretical
q3_ok	2675039	1179486	Cnt of assays where q3 is within 1 PPM of theoretical
q3_bad	0	0	Cnt of assays where q1 is more than 1 PPM from theoretical
q3_na	152861	57486	Cnt of assays peak annotation not matched
q3_ave_mdiff	0	0	Average massdiff between reported and theoretical q3
swa_defined	206964	206162	# peptide ions that fall into a defined SWATH bin
swa_missing	4406	0	# peptide ions that fail to fall into a defined SWATH bin
swa_conflict	33443	0	# q3 that fall into same SWATH(s) as parent q1
swa_ok	3196913	1433778	# q3 that do not fall into same SWATH(s) as parent q1
swa_conflict_q1	32768	0	
swa_5	38485	4775	# q3 that fall within 5 Th of parent q1
swa_25	152763	67985	# q3 that fall within 25 Th of parent q3

So, actually one caveat; so, basically this is the number of tryptic peptides, this is the number of perfectly tryptic peptides that match the perfectly tryptic peptides here. So, 20 percent of the entire proteome it covered in this library, we know that there is about 12000 out of 20000. So, we know that we have only covered 60 percent of the proteins. So, those proteins were just not getting perfect coverage.

So, yeah you might think that this number is pretty small compared to the 500000 yeah, you may well be right, may be a function of sort of quality criteria that were applied here, but I mean it is a very rare mass spec experiment or set of experiments that you can see a 100000 different peptides from human. So, these things are actually things that are along with the library and you might not think that it is possible, but it turns out that you can actually have improper m/z.

Because, were accepting things at SWATH outlets we want to make sure that people submit a database and were then providing it for the community, we want to make sure that there is not something wrong. And so by and large these numbers are all ok, these are things comparing the SWATH files you provided with the actual library. And so generally speaking, you look at all these things and there is like 4 or 5 of them that stand out to you and. You know, as somebody who actually play works with these libraries quite a bit, I have I have used this a number of times. Anyway, so I know that is pretty boring here is a little bit of description of why we might want to have good libraries.

(Refer Slide Time: 13:54)



So, as a test basically I took this library that we are talking about. And, these are the and it is its only sort of a subset of the PHL. And so these are the minimum, maximum and medium retention time. And so I did two different things. So, first of all I took randomly added or subtracted up to 10 Daltons from every peptide. So, maybe I added 1 and subtracted 3; and added 10 and subtracted 2, so basically it is randomized. And you can see with that the minimum does go down by about 10, the maximum does not really go up and the median stays the same, because I have done this randomly and so sort of added this much as, as I subtracted.

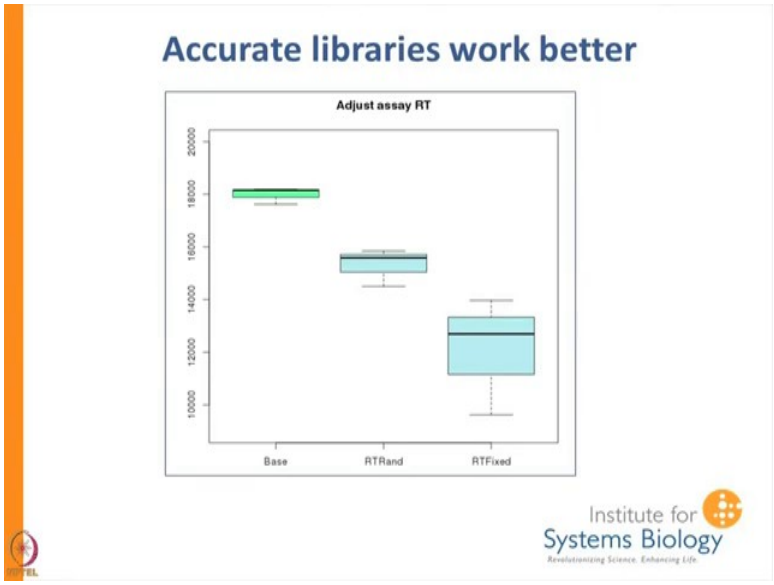
In this case basically I took and did a fixed RT. So, basically I said every peptide in here I am going to add 10 minutes, and then I am going to use this library for a SWATH analysis and I am going to see how the software can handle it. So, between these two things, so this is the sort of the baseline library, I have done a random thing where each individual change is probably less, but it is plus or minus. And in this case I have added 10 to everything, so everything in here is moved by 10 minutes.

So, which one do you think gave the worst results? The one where I randomly added plus or minus in this, in the magnitude is probably averages around 5 or the one where I added 10 to everything. Who thinks that the random one would be the worst? Who thinks the fixed one would be the worst? Who has no idea? Ok, no one's going to play.

Student: Fixed one is a problem.

Fixed one is the problem that is correct.

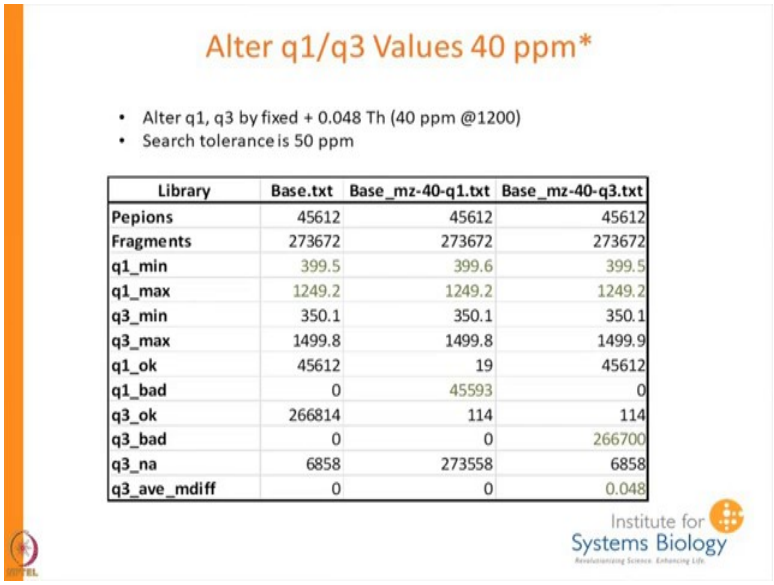
(Refer Slide Time: 15:42)



So, basically this is a so called whisker plot, box and whisker plot. The box and whisker plot basically this is the median of your data, this is the first quartile, this is the third quartile and this is the minimum and maximum, although there is some outlier detection done and so sometimes you can actually get some things below the minimum or above the maximum.

But basically it shows that the library that we started with has very narrow tolerance, I mean everything we are getting, so this is 15 different samples were seeing how many basically HeLa peptides were seeing. And in this case, each of our analyses came up with about yeah 18100 very tightly grouped. When we did these two retention time perturbations, we actually saw a significant decrease in not only the number that we see, but also the consistency across the different replicates that we see.

(Refer Slide Time: 16:38)

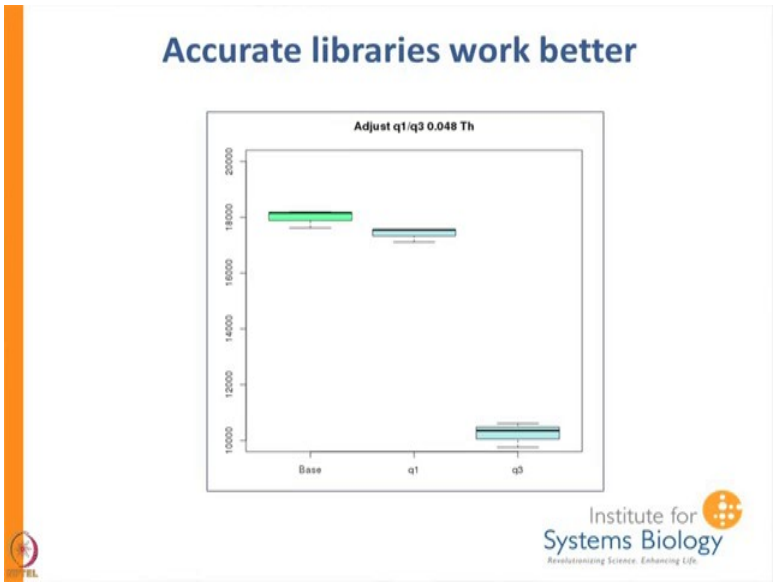


So, in a similar way basically we took the q 1 and q 3 values and we added basically modified by 40 ppm the Q1s or we modified by 40 ppm the Q3 is. So, this one which would be which would have more of an effect were taking these big Q1 chunks, and then were basically fragmenting them all, and then analyzing the Q3 versus modifying the Q3 which is actually what were reading out. Who thinks the Q1 will be a worse thing?

Student: Q3 would be worse.

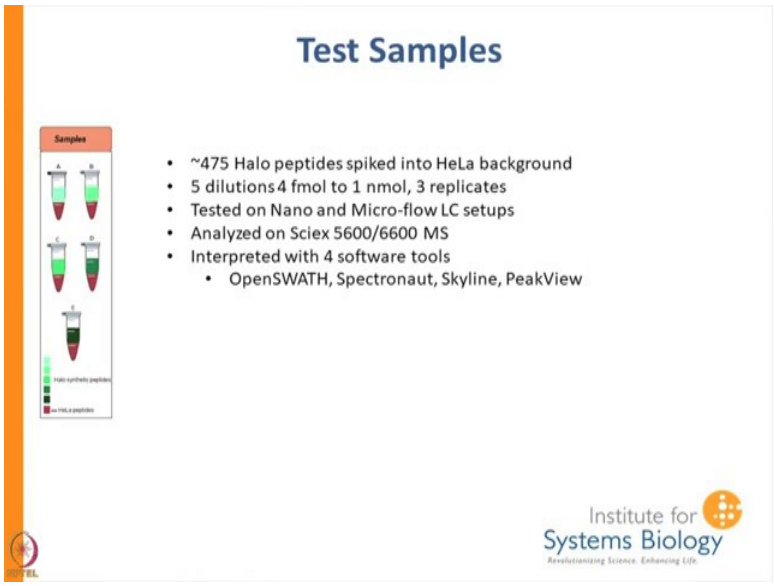
Q3 would be worse and that is exactly right. So, the Q1 you are not really doing much you may be a couple of them you are going to move it from one SWATH to another, but most of them are going to say in the same SWATH and so it does not really matter. You are just looking at basically you are looking at the same ions and it is still in the same SWATH.

(Refer Slide Time: 17:33)



But in the case of the Q3, a change of 40 ppm, again causes a drastic decrease in the number of peptides observed, although it did not affect as much the sort of reproducibility.

(Refer Slide Time: 17:43)



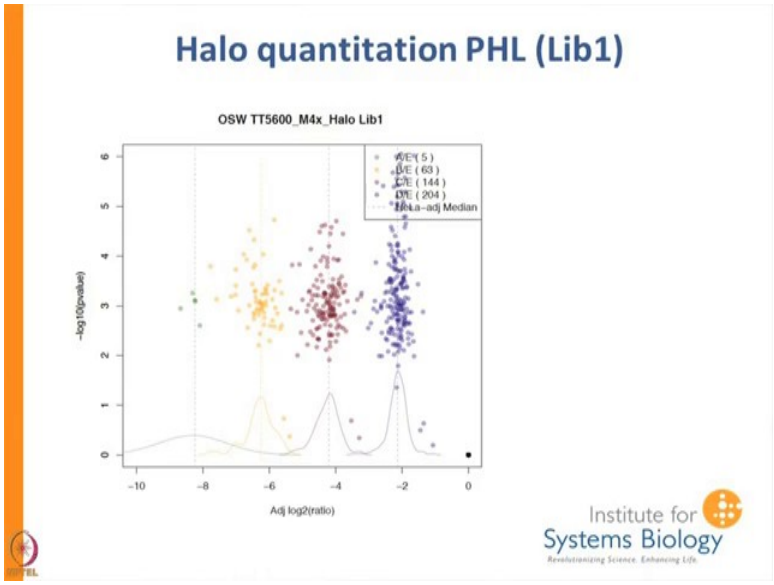
So, when I talk about one sort of real world example. So, basically we took a HeLa, do you guys know what HeLa cells are they are they are a cell line, it is one of the oldest cell lines it was taken from some poor woman back in the early mid 1990s. And it was

grown from a cancer was it sort of the first cell line that became available and it is still in use today.

And then we spiked in these Halo peptides at different concentrations. And there is nothing special about the Halo peptides, it is just a different organism than human. And so what we wanted to see is if we have this complicated background, can we see these Halo peptides we spiked in alright.

So, the 5 different dilutions and in the data, I am going to show it goes from 4 femtomole to 1 nanomole, there is 3 replicates of each tested on nano and micro-flow LC, 2 different instruments interpreted with 4 different software tools, excuse me. So, if you look at all the replicates stands in machines and everything, they are actually 360 different measurements for each point. And so we basically took it and ran against PHL, and this is what we saw.

(Refer Slide Time: 18:52)

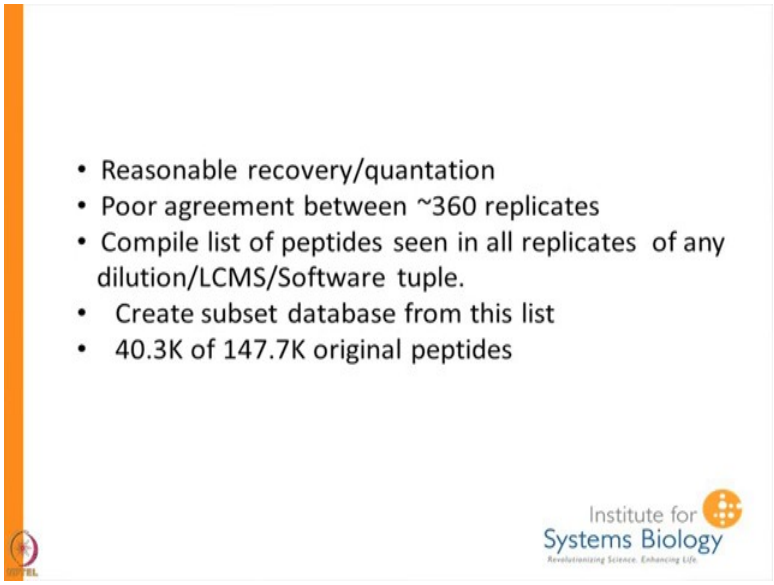


So, basically these are the two fold this is a log transformed, because when you have ratio data you always take the log. And ideally this would be right at 2, and so what we have done is we have taken and so this is sort of the 2 fold were basically taking the ratio of everything to the most concentrated sample.

So, this E sample has the most things in it and so we take for and so pretty much anything in D is also going to be an E, and anything in C is also going to be E. So, this is

most concentrated less concentrated, less concentrated down here. So, by the time we get to the very low femtomole sample, we only see 4 or 5. So, actually the number that we see is here, so out of these 475 peptides we spiked in, we are only seeing 204 and actually this is peptide ions. And so there is going to be in excess of 500 possible things, and were only seeing about half of or it really 40 percent.

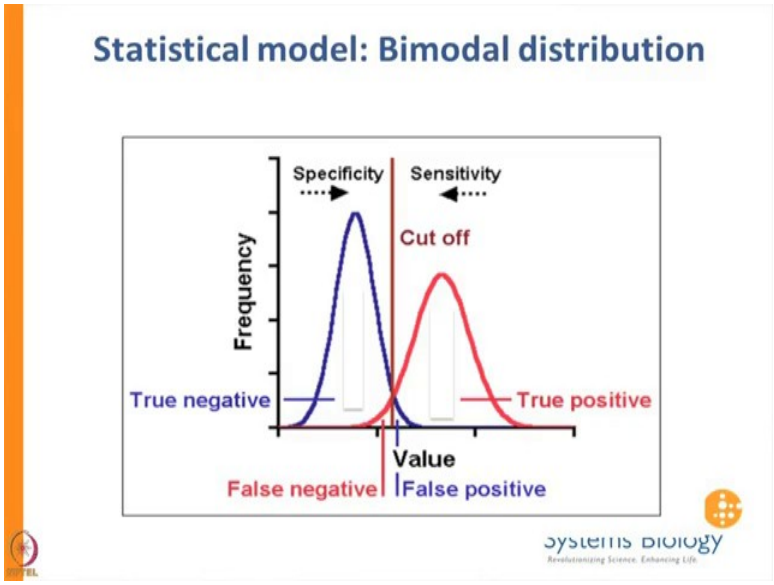
(Refer Slide Time: 19:53)



And so we looked at that we thought well that is not too good, I mean we have reasonable quantitation I mean that everything was lining up on the where it should on those axes, but we had poor agreement between the all the replicates. So, what we did is we looked at all the different experiments; we picked peptides that were seen in exactly. It were seen in all three replicates of any one technique or software what have you, and we can created a subset database from this list.

So, basically the PHL had about 147000 peptides, our new database had 40000 peptides. So, what we have done is taken a repository library and we have focused it on our sample.

(Refer Slide Time: 20:33)

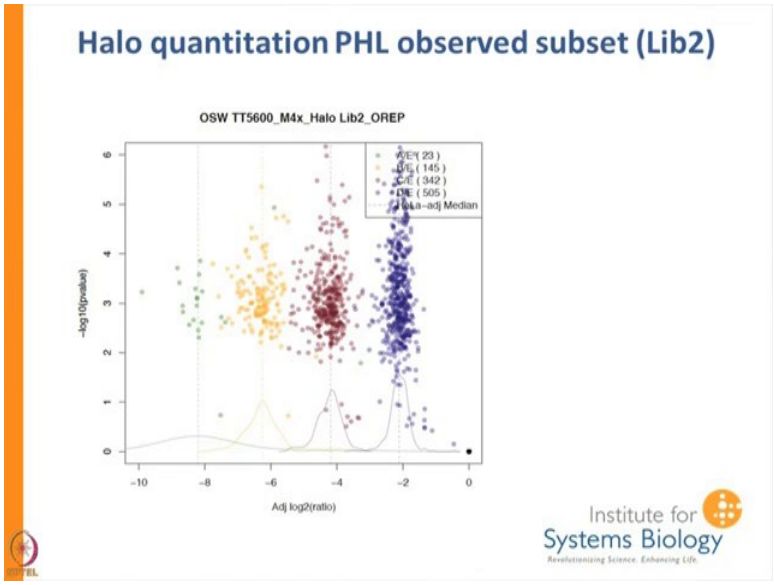


And so the reason for this is basically the way that people do big data nowadays, you take all these measurements and you do not try to threshold, you do not try to decide what is good, you do not try to take a cell and pick a cut off. What you do is you make some sort of statistical model, and you try to separate the true positives from the negatives or the yeah, anyway.

So, as you go here as you push; as you push this way, you get more and more sensitive, you get more and more things you might not have gotten; but as you go this way, you actually lose your specificity, because you start getting these negative things in with your positives. And so what you try to do is pick a point where you have an acceptable FDR.

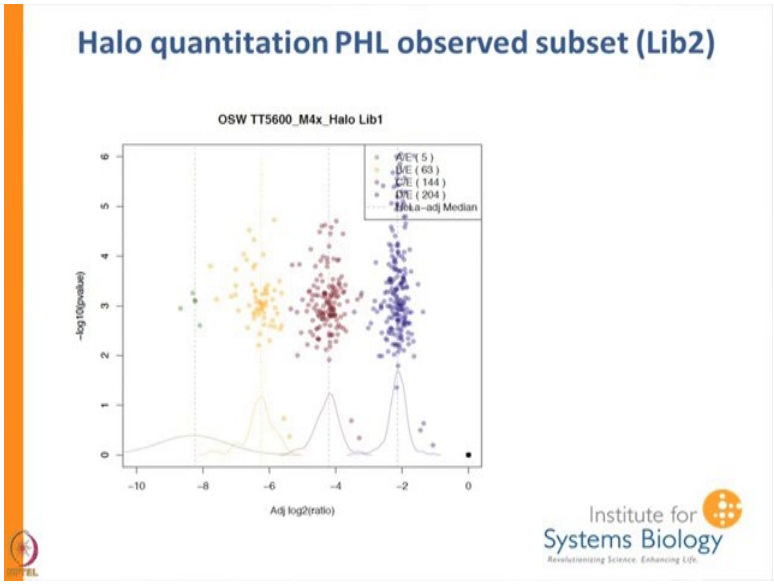
And the thing is with a big SWATH library, you do not expect to see everything in there. And so where all these things use decoys and so you expect to have a 1 to 1 ratio of decoys to properly model this, but if you if you do not see 60 percent of your SWATH library, and you make decoys for a 100 percent you basically have basically you only have about 20 percent true forwards and about 80 percent decoys. So, because you do not see everything in your library, having a big library in if in effect increases your number of decoys and that is how the mark, and so by cutting that down we hope that we have a better result.

(Refer Slide Time: 22:02)



And it turned out that we did, so basically this is the same thing analyzed with the other library. And you can see that now we see in excess of 500 things, they are still pretty close, this is pretty close to plus point to minus 2, this is pretty close to minus 4 and we now see you know 23 of these lowest concentration ones.

(Refer Slide Time: 22:20)

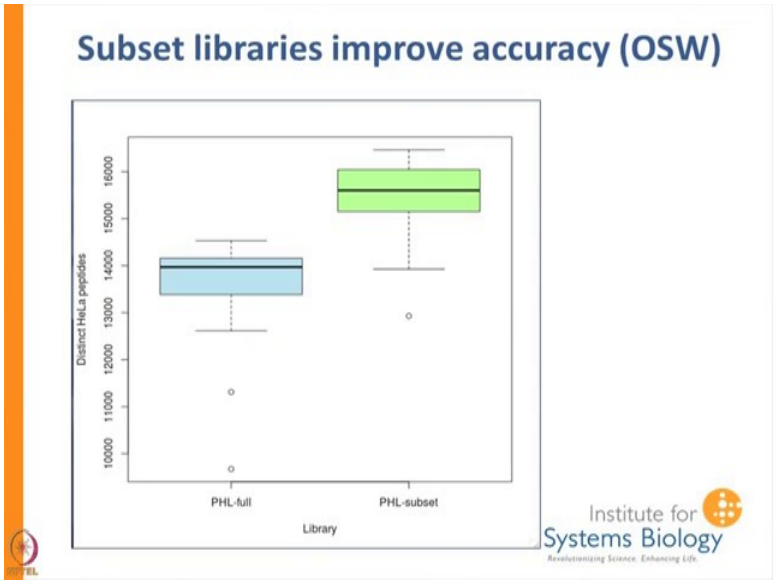


So, I will go back it went from here to here. And, so basically these lines along the bottom are a density plot. So, kernel density basically gives you an idea of how these things are spread out, because if you look at this cloud of things it is kind of hard to tell,

but you can see that these things are a little bit in this shape in and that is an indication that you know there is something a little bit weird about your data. This one has a pretty pronounced shoulder, and so sometimes what we do is we actually cut this one off at a CV coefficient of variation of say 20 percent.

So, we exclude data that is problematic, but by even by doing that we get a much cleaner peak and this density plot gets more nicely shaped, but we still have far more. So, by using a smaller library we have actually achieved, we have seen we have done better quantitation on more peptides.

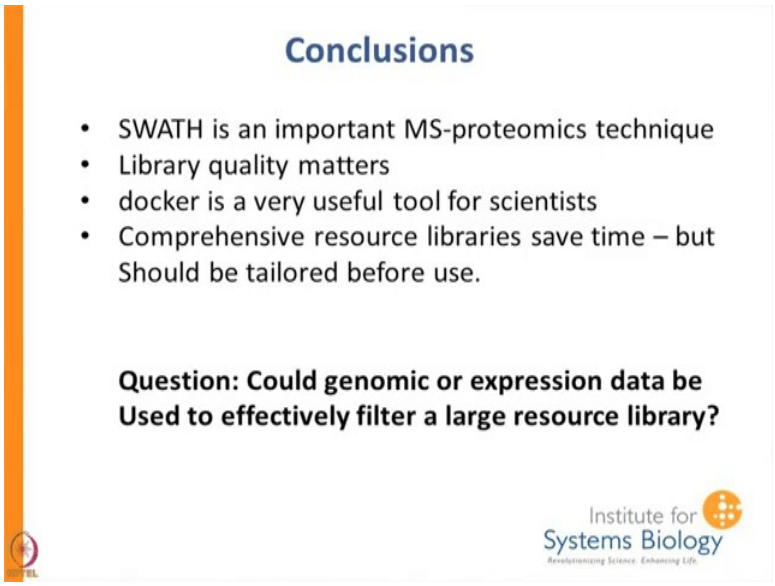
(Refer Slide Time: 23:12)



And this is another way to look at the same thing; this is looking at just the HeLa background. So, these previous one these are all the Halo peptides this is looking only at the HeLa background. And even in the HeLa peptides yeah, here we saw you know on average about 14000, but here we saw over 16000 on average. And as I said you know sometimes this is max and min, but because they do outliers sometimes you have outliers out here.

So, this is an indication that that targeting your library. So, we think that library resources are good, but you still have to do some focusing its good, because you do not have to do all your own DDA, but it does require some customization. And I think that feeds in well with the proteogenomics theme.

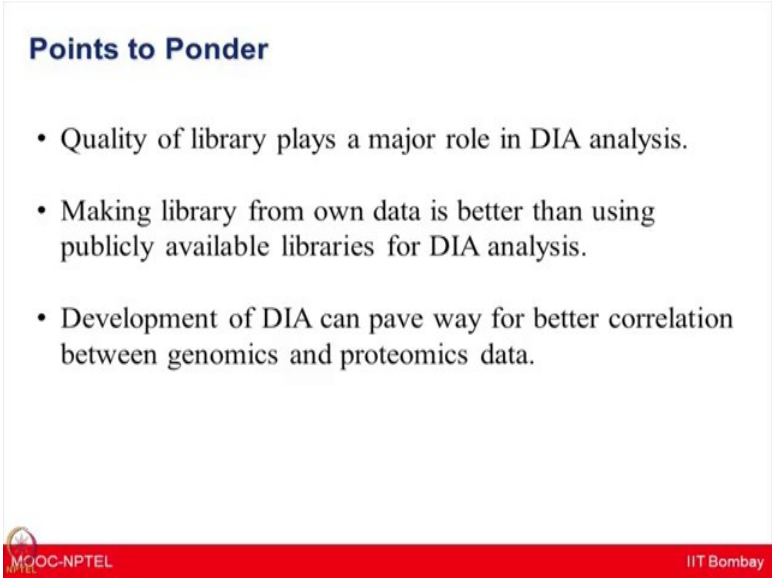
(Refer Slide Time: 23:55)



So, SWATH is an important MS technique. It has the ability to be reanalyzed forever; you get better depth than DDA. And as long as you since you can reanalyze forever as our knowledge of the proteome space and our libraries improve, you are going to be able to reanalyze the same data and get better results. Library quality does matter, I really think that Docker is a useful thing especially for somebody who is not that technically savvy it is pretty easy to install, and then you just have to go browse around for things and run them.

And, comprehensive resource libraries save time, you do not have to be your own DDA, but they should be tailored before use. And, again I think that leads into proteo-genomics, because what if you had some genomic data. And then you took one of these resource libraries and you said ok, I you know I think that in my genomic data, it tells me that I should see all these things. And so now you make a much more focused proteomics library, and you have a better chance of actually seeing what you are looking for.

(Refer Slide Time: 24:59)



In today's lecture you were introduced to the library assessment feature of SWATH Atlas, which provides recommendations to improve the DIA experiment. The DIA library QC workflow considers features such as library complexity, precursor and fragment, peptide characteristics, retention times, library completeness, while assessing and ion library. We hope now you can appreciate the use of tools like SRM Atlas, peptide Atlas and SWATH Atlas in carrying high quality mass spectrometry based proteomics research.

Thank you.