Instructor's space

Instructors, please use this space to post materials for May Institute.

Please feel free to contact me at bsearle@systemsbiology.org or @briansearle on Twitter to get answers to any future questions you have.

The slides are "open source" with a Creative Commons license (with attribution), so email me to get the PPTs if you want to use specific slides in your talks.

Slides
Recent article: Acquiring and Analyzing Data Independent Acquisition Proteomics Experiments without Spectrum Libraries

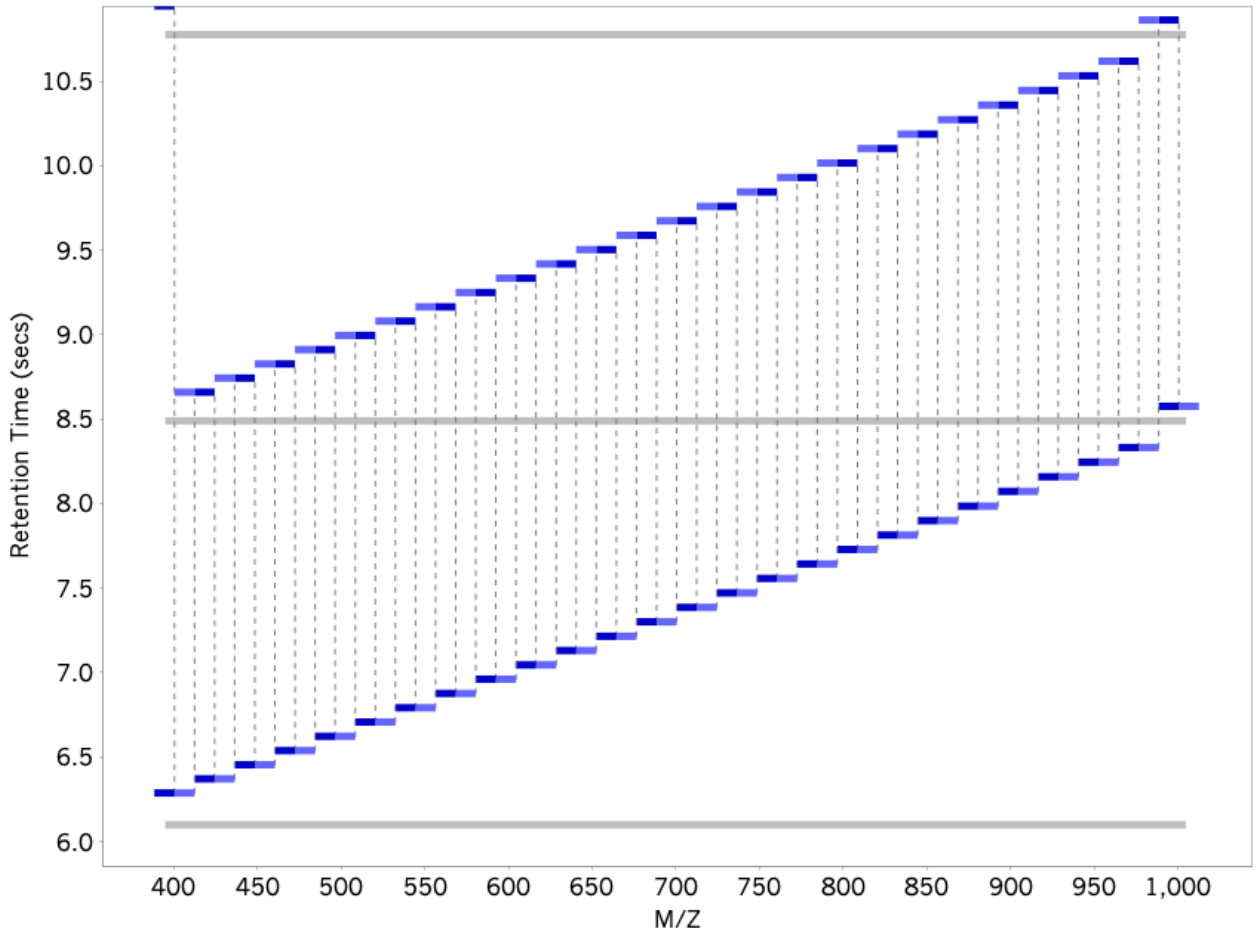
---

Participant's space

Participants, please use this space to ask questions in advance, or during the presentation. Please do not delete posts by others.

Q1 how to estimate peak width from

BCS: I recommend using EncyclopeDIA to plot characteristics of DIA files. For example, it can generate plots of the structure to show how long each cycle is. For example, this experiments' cycle time is about 2.5 seconds:



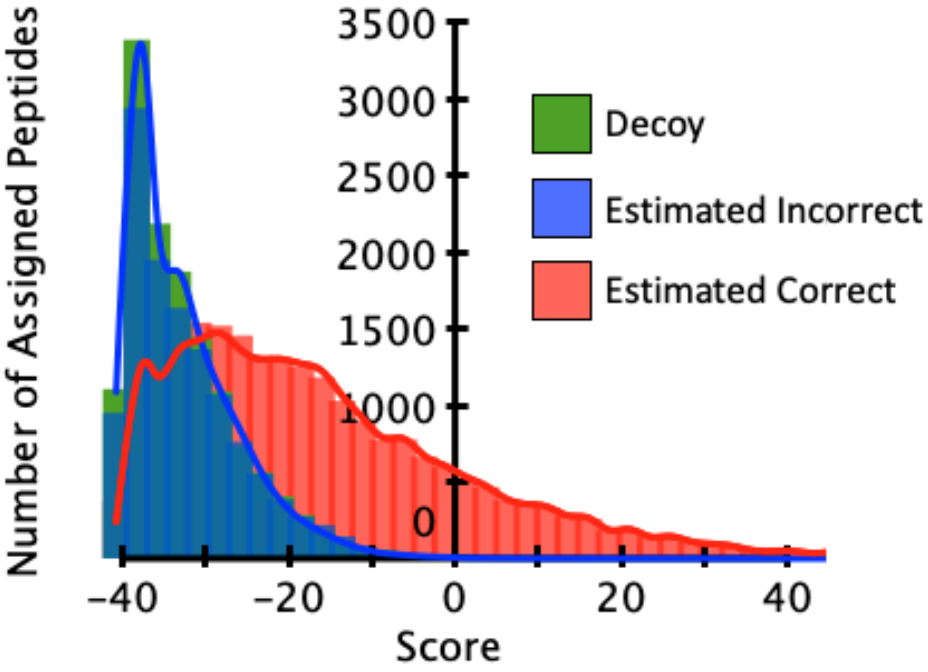
You can combine this with looking at precursor data from DDA or DIA runs to estimate how wide peaks are on your instrument and calculate how many points across the peak.

Question from youtube: How many SRM experiment will be performed by this DIA software sir?

BCS: The number of peptides you can monitor depends on the sample complexity. In typical DIA experiments of cell lines, we can often monitor 5-6,000 proteins. However, in serum we may only be able to monitor 300-400 proteins.

Q3: Can you repeat what the purpose of a “decoy” spectrum is?

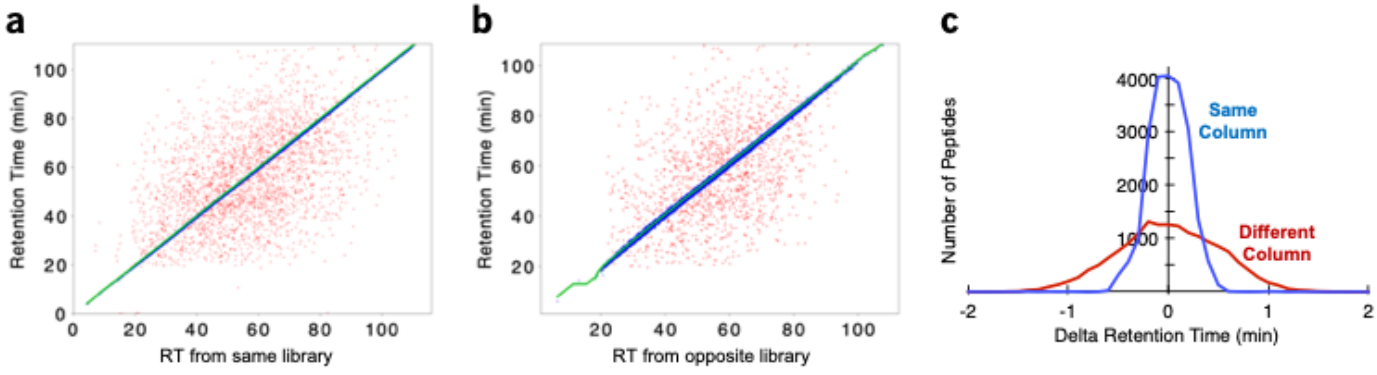
BCS: As with DDA, “decoy” peptides help you estimate how confident you are at detecting “target” peptides. Decoy peptides help you determine a null distribution to estimate the percentage of target peptides that are correct above a certain score, versus those that are incorrect. You can use this to estimate FDR with the equation %FDR=#Decoy/#Target



I recommend reading this excellent short paper by David Tabb on the subject: What's Driving False Discovery Rates?

Question from youtube: how to sort nLC reproducibility without using iRT peptides

BCS: nano LC typically produces reproducible peak ordering when using the same column, even when the actual retention times shift. However, care should be taken when you need to switch columns. In this case, if you're using the GPF-DIA technique for generating DIA libraries, we recommend collecting 6 new runs of the pool on each column. While you can reuse the DIA library, it won't be as accurate:



Can anyone recommend additional resources for learning more about DIA and analysis of DIA data that are tailored toward beginners in this topic?

BCS: Lindsay's new paper is a good first read:

https://www.mcponline.org/content/early/2020/04/20/mcp.P119.001913

what is the state of DIA experiments without project specific libraries now?

BCS: If you empirically-correct Prosit predictions with the 6x GPF-DIA approach, you can produce results just as well, or better than project-specific libraries without collecting any DDA. For more details, see our paper here:
https://www.nature.com/articles/s41467-020-15346-1

how would this work for peptide discovery?

BCS: The above paper on Prosit/GPF-DIA performs very well for discovery methods. In that paper we generate a DIA-only *Plasmodium falciparum* proteome comparable to fractionated DDA experiments with much less effort.

is there a list of "good" peptides somewhere for human proteins? Not just in terms of signal, but in terms of behavior, e.g. not many PTMs or chemical mods and also reproducible digestion, or inversely a list of "bad" peptides?

BCS: Peptides behave very differently depending on the sample workup and matrix, so it's hard to have a list of peptides that are "good" for every situation. Better yet, Lindsay has a great paper on how to relatively easily generate your own list for your own protocols/matrix! Check it out here:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7175947/

How about comibing both peptide centric and spectrum centric approaches, I think Spectronaut does it, any other tools apart from that? or doing sepeartly and then combining results? how should one go about it..

BCS: Personally, I am very cautious about strategies like this because they produce different types of false positives and peptides that score differently. For example, performing spectrum-centric approaches produce acquisition-specific "libraries" that are infinitely accurate because they came from the same data. Combining this with DDA-based spectrum library searching will produce two types of peptide matches with two distinct types of score distributions.

But using too big library will also be not ideal, rigth?

BCS: You want to tread a fine line, here. If your library is too small then you may mistake a peptide you care about for signal from another peptide you hadn't considered. Your library needs to be large enough to represent all of the types of peptides you're interested in characterizing, without having such a large library that it's impossible to make significant peptide detections. This is one of the reasons why the GPF-DIA library building strategy works so well -- you can use a huge library to start because it's easy to detect peptides with GPF-DIA, and then pair the library down to only peptides that are detectable in the pool.

What is success rate in confimation of DIA run by PRM?

BCS: In our limited confirmation analysis of 500 peptides (shown in the slide titled “DIA vs DDA vs PRM with transition refinement”), DIA produced roughly accurate results 95% of the time.

Most likely, DIA will detect less proteins than DDA, correct?

BCS: I disagree. This strongly depends on library quality and depth. In our experiments on chromatogram libraries, all library-based methods (either DDA spectrum libraries or DIA chromatogram libraries) produced longer peptide detection lists than comparable DDA experiments:
https://www.nature.com/articles/s41467-018-07454-w

In our experiments of the *Plasmodium falciparum* proteome, DIA significantly outperformed DDA, and was able to accurately detect and quantify peptides even when the proteome was diluted with human red blood cells (a generally very difficult proteome due to the high concentration of hemoglobin):
https://www.nature.com/articles/s41467-020-15346-1

do you recommand IPs for DIA method

BCS: If you’re using IPs for specific proteins, then you may be better off implementing a targeted proteomics scheme using PRMs. However, if your goal is to determine non-specific or indirect binding for your IPs, then DIA may be helpful.

can i use DDA spectral libraries to analyze DIA data generated on another instrument and having non-matched RT? Like using DDA spectral libraies downloaded from different databases? Or we need to match always DDA for spectral libraries with DIA by using the same instrument, gradient, etc?

BCS: Most DIA tools feature some degree of retention time warping, which lets you use libraries created on different instruments or gradients. However, the library accuracy will be higher if you generate DDA libraries on the same column/instrument or make a DIA chromatogram library using GPF-DIA.

what is the optimum size for the library in terms of number of precusrors?

BCS: This depends on the sample you’re analyzing. I try to not use libraries smaller than about 5,000 peptides or larger than 300,000 peptides.
https://docs.google.com/document/d/1UU8aF-T5EZ7h97ZebuGAtjisLrYS7sLCXUX5yPhv-ek/edit#
If you have a small number of interested proteins, wouldn't PRM be a more useful approach as the analysis would be much less complex?

BCS: Yes, if you have proteins you're interested in targeting, then PRM data is easier to analyze. However, with DIA you don't have to schedule peptides, so DIA data acquisition is easier to set up if you're only going to analyze a few samples.

Could a retention time standard pepmix be used along with sample (spiked) to construct iRT list?

BCS: You can build a retention time standard using peptide mixtures. That said, iRT is a dimensionless coordinate, so as long as you're internally consistent, you can use any peptides (including endogenous peptides) to build your index.

Is SpectraST relevant ?
Or is it too old ..

BCS: SpectraST is great for analyzing DDA data with spectrum libraries, or for generating DDA-based libraries. It was not designed for searching DIA data, so I wouldn't use it for that.

Where does TMT based quantification sit in terms of % quantification in comparison to DIA and label free DDA

BCS: TMT/iTRAQ with MS3 (e.g. only with a Fusion/Lumos/Eclipse) can be quite accurate and can produce deep quantification lists with enough fractionation. These methods begin to break down as you need to expand to multiple TMT/iTRAQ batches (e.g. >16 samples). While it is possible to combine batches (e.g. using a method like this: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5417827/), you are limited to quantifying only peptides/proteins that are detected in all of the batches. This can be difficult if combined with fractionation, since reproducibility between batches can be a concern.
OV: let me add that MSstatsTMT can analyze multi-batch experiments where a peptide is not detected in every batch

When throwing out transitions 'as needed' must that be manual? if so - is that feasible across the whole dataset?

BCS: EncyclopeDIA uses automated transition refinement (without any manual effort) to remove fragment ions that do not agree with the overall peak shape.

Are those settings based on a 2hr LC gradient?

BCS: The DIA settings listed are designed LC gradients with 25-45 second peaks. If you have shorter or longer peaks then you may need to adapt the settings to your situation. Suggestions for how to do this are shown here: https://docs.google.com/spreadsheets/d/1A8AQlmLroAkQcAcsiGTNvnGBE2lGpkMwhh0YLTBHXKA/

For overlap, you are sugesting overlapping between cycles. How about overalp within a cycle? Such as 25 m/z windows but having overlap of ~1-2 m/z.

BCS: I refer to small overlaps (<1-2 m/z) as “margins” and 50% overlaps as “staggering”.

What is the fatest staggered method that you have had success with on an Orbitrap?

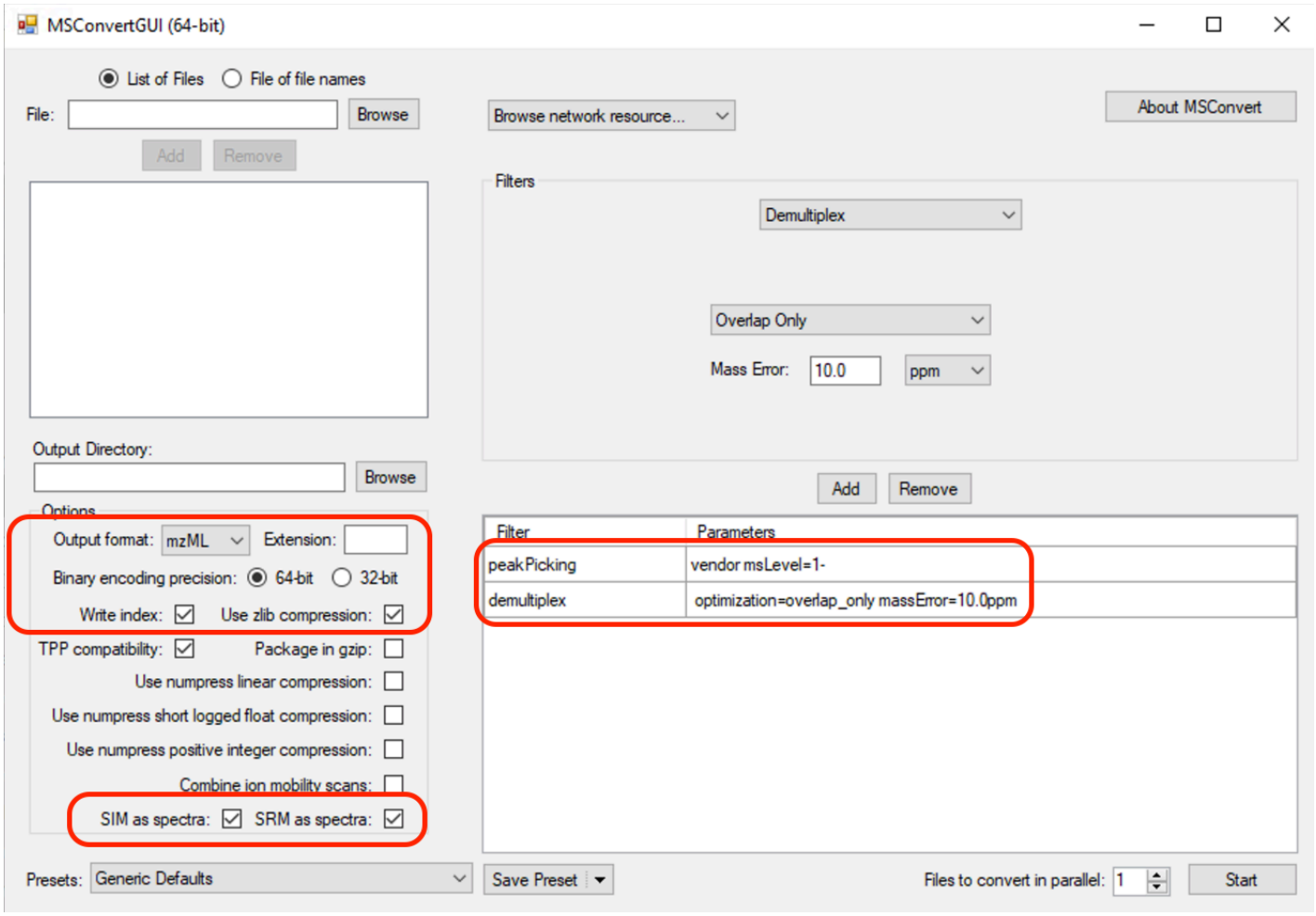
BCS: I think staggering should be used for all Orbitrap DIA methods, since there is little-to-no downside. The staggered method I recommend for Lumos/Eclipse/QE-HFX/Exploris Orbitraps is 8 or 12 m/z, depending on your chromatography speed. The staggered method I recommend for slower instruments is 16 or 24 m/z.

Did you do fractionation of peptides before DIA acquisition? Or you prefer to have a longer gradient without any fractionation (like the high pH HPLC in DDA we usually do)?

BCS: I do not perform any offline fractionation prior to DIA (only gas-phase fractionation in Q1). All of the gradients I showed were 90 minutes long.

how do you deconvolute the data acquired with staggered or overlapping margin windows?

BCS: Staggering deconvolution can be easily performed in Proteowizard using these settings:



can you elaborate on your preference for +3 now?

BCS: The majority of detectable tryptic peptides are +2H, so it is logical to fragment all peptides as if they were +2H. More recently I find a small increase in peptide detections when in fragmenting peptides as +3H. I suspect this may be because it rescues some +4H peptides, while not dramatically changing the detection rate for +2H peptides. However, the difference is small.

which software can deal with DIA data analysis using gas phase fractionation library?

BCS: Right now, only EncyclopeDIA and Scaffold DIA can build GPF-DIA libraries. However, once built, these can be used in other software programs, such as Skyline.

Can we use EncyclopeDIA for MSE data from a Waters Synapt (QTOF, no ion mobility)?

BCS: I haven't tried, but I would not recommend doing this. PLGS is probably the best tool for extracting information from MSE without ion mobility.

would you still need to spike in iRT standards in your samples and pooled sample when doing a library-free + GPF workflow?

BCS: You do not need any iRT or other stable isotope-labeled (SIL) peptides for the GPF workflow, or any other EncyclopeDIA or Scaffold DIA workflow. These software packages are designed to work without any SIL peptides.

In support of yr point Yates published long that you could park on a precursor and get an interpretable MSMS without ever seeing the precursor in the MS1.

BCS: Yes! David Goodlett's lab showed this with Pacific, as well in: https://www.ncbi.nlm.nih.gov/pubmed/19572557

To clarify, the gas phase fractionation makes a spectrum library composed of the 50 samples in your pooled sample - the generated library could be done in a program like encyclopedia. Prosit predicts all the possible peptides in a FASTA library. You compare the peptides from that library to the spectrum library generated in EncyclopeDIA. You compare this to the data in your individual samples. Is this correct?

BCS: Let me restate it this way: Prosit generates a predicted spectrum library for all possible +2H and +3H peptides in a FASTA. If you pool a subaliquot of your samples and run GPF-DIA with 4 m/z wide windows (2 m/z after deconvolution), you can search those GPF-DIA injections with the predicted spectrum library to identify peptides. Using those peptide identifications, you can create a DIA-based chromatogram library (complete with peakshape and interference statistics) for your experiment. Searching single-injection DIA runs with a chromatogram library typically performs as well as (or better than) searching those runs with normal sample-specific DDA-based libraries. This method is more thoroughly described here:

https://www.nature.com/articles/s41467-020-15346-1

Can I use EncyclopeDIA for Orbitrap Exploris together with FAIMS? In that case what parameters need to be modified? Thank you.

BCS: I have not tried this yet. Email me at bsearle@systemsbiology.org and let's take a look!