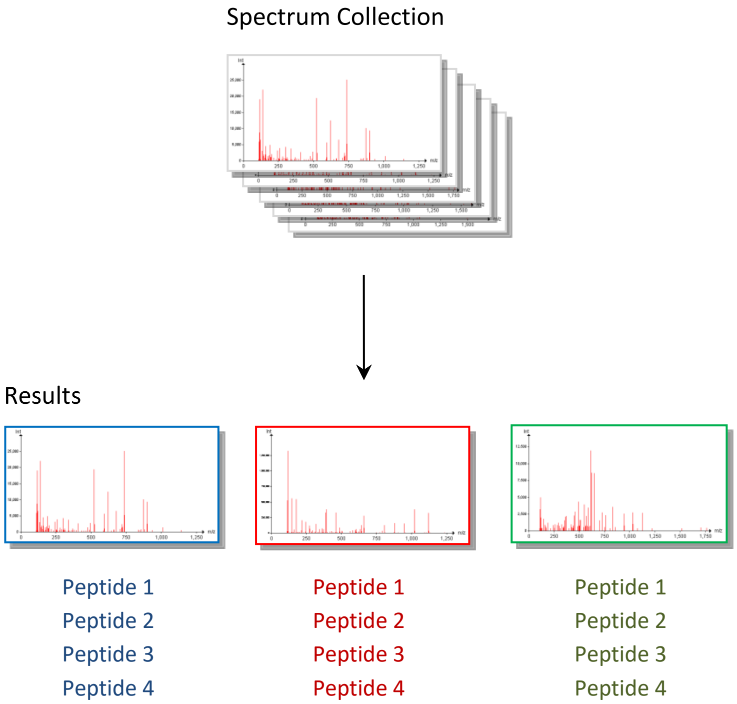
De Novo Peptide Identification

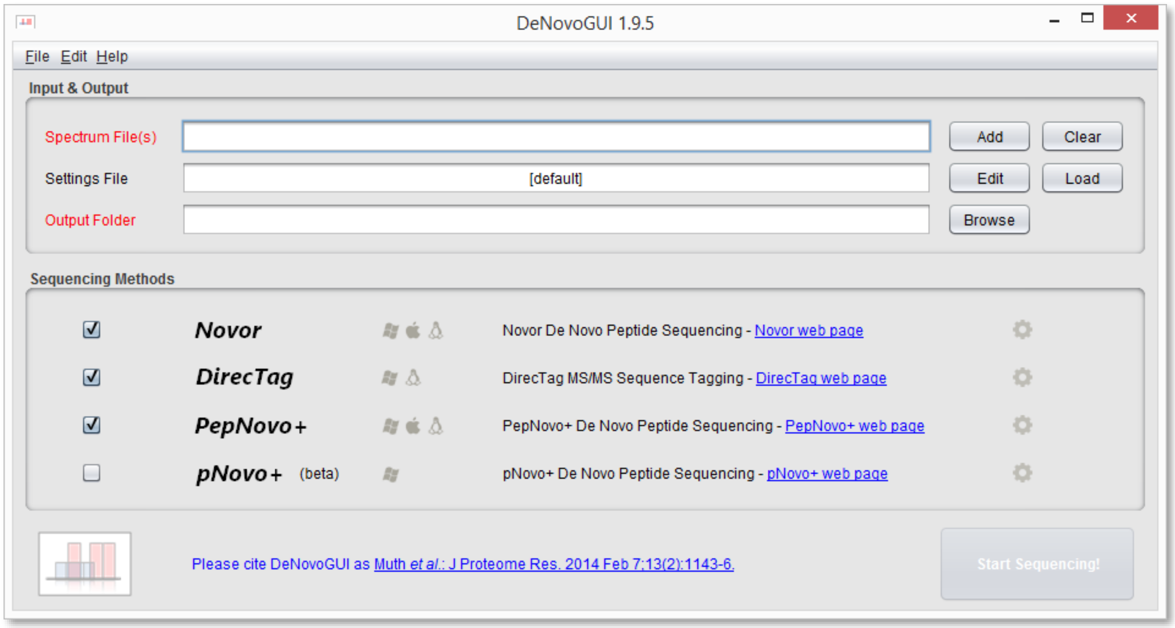
When a suitable sequence database is not available for your sample of interest, identification of your acquired fragmentation mass spectra becomes quite difficult. Indeed, the widely-used database search engines cannot be used anymore because they crucially rely on a predefined list of possible protein (and after *in silico* digest, peptide) sequences. In the absence of such a database however, we instead need to rely on the sequence information that can be read directly from a fragmentation spectrum. This procedure is called *de novo* sequencing, and special software tools have been built for this purpose.

We are again going to use the mgf file obtained in the "Peak List Generation" chapter, but instead of searching it against its database (as we did in the "Peptide to Spectrum Matching" chapter), we will attempt to extract as many sequences as possible from the spectra themselves. To do this, we will make use of freely available *de novo* engines. The necessary spectrum file can be found in the resources folder.



Novor[1], DirecTag[2], PepNovo+[3] and pNovo+[4] can all easily be used *via* a graphical user interface called DeNovoGUI[5]. DeNovoGUI for Windows platforms is provided in the software folder together with Novor, DirecTag, PepNovo+, and pNovo+. For Mac and Linux versions, please see the DeNovoGUI web page http://compomics.github.io/projects/denovogui.html. Start DeNovoGUI by double clicking the file DeNovoGUI-X.Y.Z.jar (replace X.Y.Z with the current DeNovoGUI version number).

You will then see the following dialog:



Settings File [default]

☑	Novor		Novor De Novo Peptide Sequencing - Novor web page
☑	DirecTag		DirecTag MS/MS Sequence Tagging - DirecTag web page
☑	PepNovo+		PepNovo+ De Novo Peptide Sequencing - PepNovo+ web page
☐	pNovo+	(beta)	pNovo+ De Novo Peptide Sequencing - pNovo+ web page

Please cite DeNovoGUI as Muth *et al.*: J Proteome Res. 2014 Feb 7;13(2):1143-6.

The input fields you need to complete are straightforward: provide one or more files with fragmentation spectra, configure some settings, and provide an output folder.

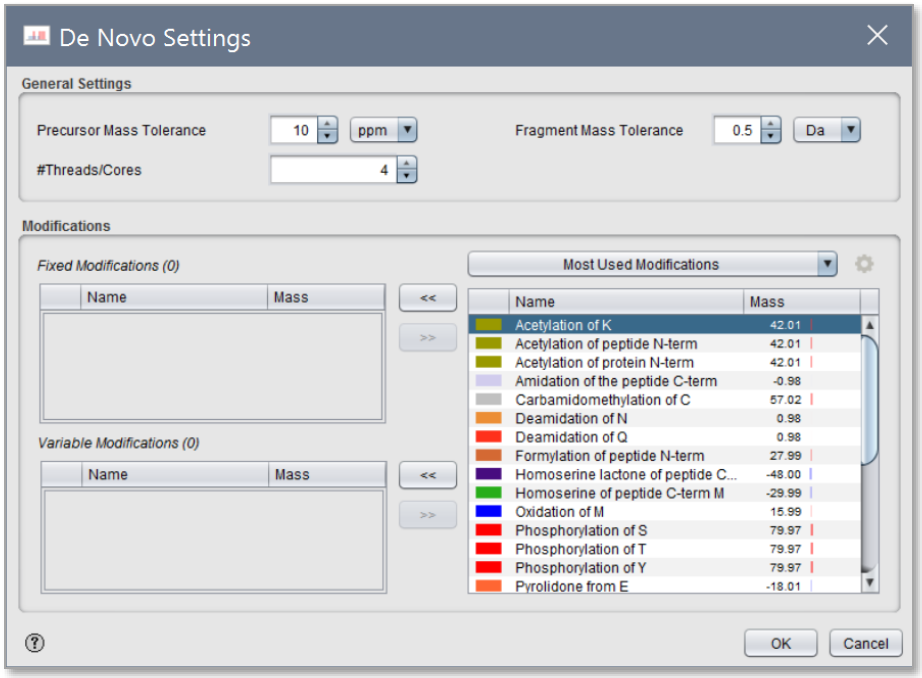
The next section allows you to select the *de novo* tools you wish to apply to your data. You will notice that Novor, DirecTag, and PepNovo+ are already selected. In fact, keen observers may already have noticed these tools in the DeNovoGUI resources folder. This means that when you have downloaded the DeNovoGUI zip file and unzipped it (which comprises the entire installation procedure), you have also already installed all four *de novo* tools along with it!

In order to perform the *de novo* identification, we need to provide the spectra and experiment dependent search settings. Load the mgf file qExactive01819.mgf created in the “Peak List Generation” chapter (also available in the resources folder).

Tip:

Note that you can load multiple mgf files and even entire folders.

We are now going to set the search settings in the Search Settings dialog. Click the 'Edit' button after the 'Search Settings' text field.



De Novo Settings

General Settings

Precursor Mass Tolerance: 10 ppm

Fragment Mass Tolerance: 0.5 Da

#Threads/Cores: 4

Modifications

Fixed Modifications (0)

Name	Mass

Variable Modifications (0)

Name	Mass

Most Used Modifications

	Name	Mass
🟨	Acetylation of K	42.01
🟨	Acetylation of peptide N-term	42.01
🟨	Acetylation of protein N-term	42.01
🟪	Amidation of the peptide C-term	-0.98
⬜	Carbamidomethylation of C	57.02
🟧	Deamidation of N	0.98
🟥	Deamidation of Q	0.98
🟫	Formylation of peptide N-term	27.99
🟪	Homoserine lactone of peptide C...	-48.00
🟩	Homoserine of peptide C-term M	-29.99
🟦	Oxidation of M	15.99
🟥	Phosphorylation of S	79.97
🟥	Phosphorylation of T	79.97
🟥	Phosphorylation of Y	79.97
🟧	Pyrolidone from E	-18.01

In which ways is this settings dialog similar to the one presented by SearchGUI (or in fact, any search engine)? In which ways does it differ? Does this make sense? [1.7a]

Because the data we are using is derived from a high-resolution instrument, we leave the precursor mass tolerance at 10 ppm, and set the fragment mass tolerance to 0.02 Da.

Do you think that it is useful in de novo identification to have smaller mass tolerances? Which mass tolerance would be the most important to minimize? [1.7b]

When using *de novo* it is often possible to set the maximum number of solutions suggested. The default in DeNoboGUI is 10 solutions per spectrum.

Is it reasonable to expect so many possible sequences? Is this unique to de novo identification, or do you think it affects traditional search engines as well? [1.7c]

Leave the number of threads at its default; this setting will default to the maximum number of cores on your machine for maximal speed.

The next step is to specify the modifications to consider. As fixed modifications choose Carbamidomethylation of C, and as variable modifications choose Oxidation of M.

Tip:

'CTRL + Mouse Click' allows you to select multiple entries.

Note that only the most commonly used modifications are listed in this dialog. There are more modifications available in DeNovoGUI, and you can also set up your own modifications. To see all the modifications, select "All Modifications" in the drop down menu above the modifications table.

To see the modification details or to add your own modifications click the settings ions next to the drop-down menu. (The modification details are also available in the main DeNovoGUI frame, Edit menu > Modifications.) The procedure to manage modifications is exactly the same as in SearchGUI, so we refer the interested reader to the corresponding explanations in the “Peptide to Spectrum Matching” chapter.

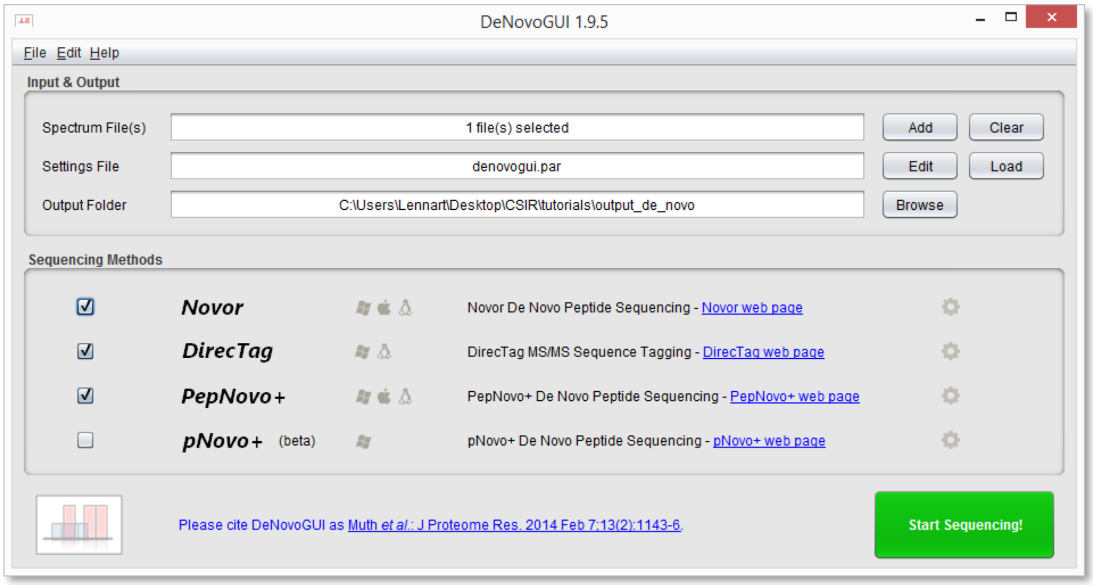
All settings are now filled in. Click the 'OK' button and save the settings for future reuse. Choose a meaningful location and name for the parameter file in the dialog. The next time you want to use the exact same search settings you can simply select this file in the main DeNovoGUI dialog.

Tip:

A well-organized library of parameter files can save a lot of time!

In the main DeNovoGUI dialog you will note that three of the four tools are selected. This means that one can easily run multiple tools and get result files for all of them at the same time. We will keep the default selection.

Finally, select an output folder. You should now see the following:



Spectrum File(s): 1 file(s) selected
Settings File: denovogui.par
Output Folder: C:\Users\Lennart\Desktop\CSIR\tutorials\output_de_novo

☑ Novor — Novor De Novo Peptide Sequencing - Novor web page
☑ DirecTag — DirecTag MS/MS Sequence Tagging - DirecTag web page
☑ PepNovo+ — PepNovo+ De Novo Peptide Sequencing - PepNovo+ web page
☐ pNovo+ (beta) — pNovo+ De Novo Peptide Sequencing - pNovo+ web page

Please cite DeNovoGUI as Muth *et al.*: J Proteome Res. 2014 Feb 7;13(2):1143-6.

Tip:

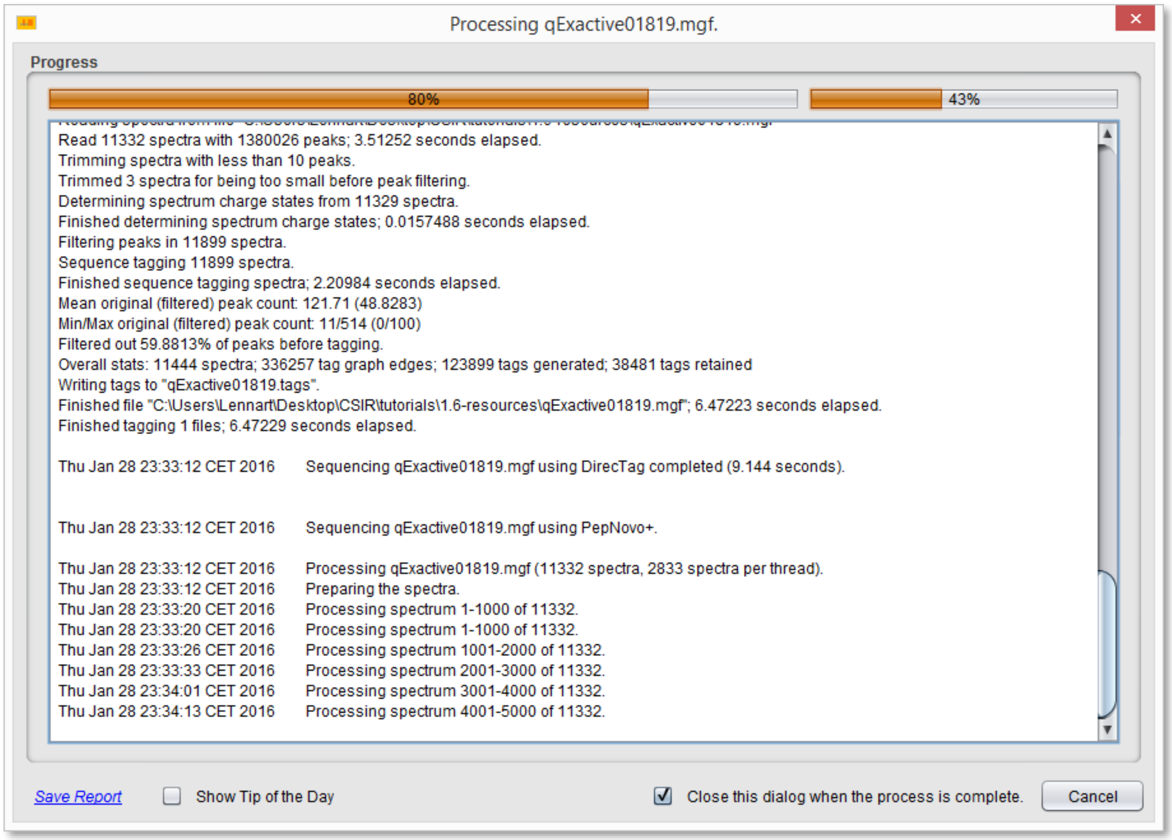
Advanced settings are available for each tool by clicking the settings icon to the right of each tool.

Note: changing these advanced settings are recommended for advanced users only!

Tip:

Always use an empty folder for the search output as this simplifies post-processing!

Pressing the 'Start Sequencing!' button will launch the selected *de novo* tools. A progress bar and scrolling text will keep you informed on the progress of the search.



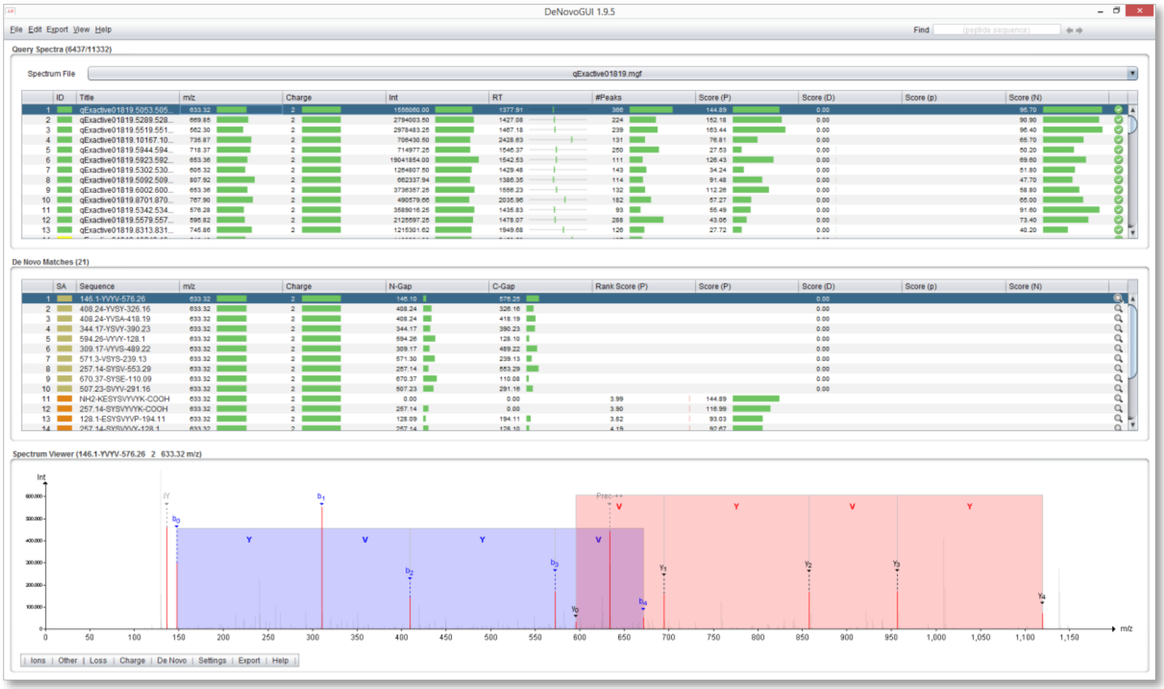
Processing qExactive01819.mgf.

```
Read 11332 spectra with 1380026 peaks; 3.51252 seconds elapsed.
Trimming spectra with less than 10 peaks.
Trimmed 3 spectra for being too small before peak filtering.
Determining spectrum charge states from 11329 spectra.
Finished determining spectrum charge states; 0.0157488 seconds elapsed.
Filtering peaks in 11899 spectra.
Sequence tagging 11899 spectra.
Finished sequence tagging spectra; 2.20984 seconds elapsed.
Mean original (filtered) peak count: 121.71 (48.8283)
Min/Max original (filtered) peak count: 11/514 (0/100)
Filtered out 59.8813% of peaks before tagging.
Overall stats: 11444 spectra; 336257 tag graph edges; 123899 tags generated; 38481 tags retained
Writing tags to "qExactive01819.tags".
Finished file "C:\Users\Lennart\Desktop\CSIR\tutorials\1.6-resources\qExactive01819.mgf"; 6.47223 seconds elapsed.
Finished tagging 1 files; 6.47229 seconds elapsed.

Thu Jan 28 23:33:12 CET 2016     Sequencing qExactive01819.mgf using DirecTag completed (9.144 seconds).


Thu Jan 28 23:33:12 CET 2016     Sequencing qExactive01819.mgf using PepNovo+.

Thu Jan 28 23:33:12 CET 2016     Processing qExactive01819.mgf (11332 spectra, 2833 spectra per thread).
Thu Jan 28 23:33:12 CET 2016     Preparing the spectra.
Thu Jan 28 23:33:20 CET 2016     Processing spectrum 1-1000 of 11332.
Thu Jan 28 23:33:20 CET 2016     Processing spectrum 1-1000 of 11332.
Thu Jan 28 23:33:26 CET 2016     Processing spectrum 1001-2000 of 11332.
Thu Jan 28 23:33:33 CET 2016     Processing spectrum 2001-3000 of 11332.
Thu Jan 28 23:34:01 CET 2016     Processing spectrum 3001-4000 of 11332.
Thu Jan 28 23:34:13 CET 2016     Processing spectrum 4001-5000 of 11332.
```

□ Show Tip of the Day

☑ Close this dialog when the process is complete.

Once the *de novo* identification completes, you will see the following results display:



Query Spectra (6437/11332)

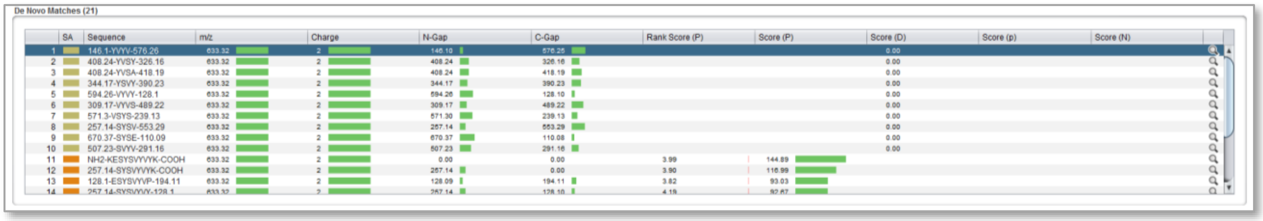
	ID	Title	m/z	Charge	Int	RT	#Peaks	Score (P)	Score (D)	Score (p)	Score (N)	
1	🟩	qExactive01819.5053.505...	633.32	2	1556060.00	1377.91	366	144.89	0.00		95.70	✓
2	🟩	qExactive01819.5289.528...	669.85	2	2794003.60	1427.08	224	152.18	0.00		90.90	✓
3	🟩	qExactive01819.5519.551...	562.30	2	2978483.25	1467.18	239	163.44	0.00		96.40	✓
4	🟩	qExactive01819.10167.10...	735.87	2	706430.50	2428.63	131	76.81	0.00		65.70	✓
5	🟩	qExactive01819.5944.594...	718.37	2	714977.25	1646.37	250	27.53	0.00		50.20	✓
6	🟩	qExactive01819.5923.592...	653.36	2	19041854.00	1542.53	111	126.43	0.00		69.60	✓
7	🟩	qExactive01819.5302.530...	605.32	2	1264807.50	1429.48	143	34.24	0.00		51.80	✓
8	🟩	qExactive01819.5092.509...	807.92	2	662337.94	1386.35	114	91.48	0.00		47.70	✓
9	🟩	qExactive01819.6002.600...	653.36	2	3736357.25	1556.23	132	112.26	0.00		58.80	✓
10	🟩	qExactive01819.8701.870...	767.90	2	490579.66	2035.96	182	57.27	0.00		65.00	✓
11	🟩	qExactive01819.5342.534...	576.28	2	3589016.25	1435.83	93	55.49	0.00		91.60	✓
12	🟩	qExactive01819.5579.557...	595.82	2	2125597.25	1478.07	288	43.06	0.00		73.40	✓
13	🟩	qExactive01819.8313.831...	745.86	2	1215301.62	1949.68	126	27.72	0.00		40.20	✓

De Novo Matches (21)

	SA	Sequence	m/z	Charge	N-Gap	C-Gap	Rank Score (P)	Score (P)	Score (D)	Score (p)	Score (N)	
1	🟫	146.1-YVYV-576.26	633.32	2	146.10	576.25			0.00			🔍
2	🟫	408.24-YVSY-326.16	633.32	2	408.24	326.16			0.00			🔍
3	🟫	408.24-YVSA-418.19	633.32	2	408.24	418.19			0.00			🔍
4	🟫	344.17-YSVY-390.23	633.32	2	344.17	390.23			0.00			🔍
5	🟫	594.26-VYVY-128.1	633.32	2	594.26	128.10			0.00			🔍
6	🟫	309.17-VYVS-489.22	633.32	2	309.17	489.22			0.00			🔍
7	🟫	571.3-VSYS-239.13	633.32	2	571.30	239.13			0.00			🔍
8	🟫	257.14-SYSV-553.29	633.32	2	257.14	553.29			0.00			🔍
9	🟫	670.37-SYSE-110.09	633.32	2	670.37	110.08			0.00			🔍
10	🟫	507.23-SVYV-291.16	633.32	2	507.23	291.16			0.00			🔍
11	🟧	NH2-KESYSVYVYK-COOH	633.32	2	0.00	0.00	3.99	144.89				🔍
12	🟧	257.14-SYSVYVYK-COOH	633.32	2	257.14	0.00	3.90	116.99				🔍
13	🟧	128.1-ESYSVYVP-194.11	633.32	2	128.09	194.11	3.82	93.03				🔍
14	🟧	257.14-SYSVVVY-128.1	633.32	2	257.14	128.10	4.19	92.67				🔍

Note that the top panel lists the spectra that were provided as input, along with summary information for each spectrum. Clicking a spectrum then shows the corresponding results in the middle panel. And the bottom panel, finally, shows the annotation of the selected result on the input spectrum.

Now let’s inspect this in some more detail. Select the spectrum with the following title “qExactive01819.5053.5053.2 File:"qExactive01819.raw", NativeID:"controllerType=0 controllerNumber=1 scan=5053"”. Now have a closer look at the middle, results panel:


De Novo Matches (21)

	SA	Sequence	m/z	Charge	N-Gap	C-Gap	Rank Score (P)	Score (P)	Score (D)	Score (p)	Score (N)	
1	🟫	146.1-YVYV-576.26	633.32	2	146.10	576.25			0.00			🔍
2	🟫	408.24-YVSY-326.16	633.32	2	408.24	326.16			0.00			🔍
3	🟫	408.24-YVSA-418.19	633.32	2	408.24	418.19			0.00			🔍
4	🟫	344.17-YSVY-390.23	633.32	2	344.17	390.23			0.00			🔍
5	🟫	594.26-VYVY-128.1	633.32	2	594.26	128.10			0.00			🔍
6	🟫	309.17-VYVS-489.22	633.32	2	309.17	489.22			0.00			🔍
7	🟫	571.3-VSYS-239.13	633.32	2	571.30	239.13			0.00			🔍
8	🟫	257.14-SYSV-553.29	633.32	2	257.14	553.29			0.00			🔍
9	🟫	670.37-SYSE-110.09	633.32	2	670.37	110.08			0.00			🔍
10	🟫	507.23-SVYV-291.16	633.32	2	507.23	291.16			0.00			🔍
11	🟧	NH2-KESYSVYVYK-COOH	633.32	2	0.00	0.00	3.99	144.89				🔍
12	🟧	257.14-SYSVYVYK-COOH	633.32	2	257.14	0.00	3.90	116.99				🔍
13	🟧	128.1-ESYSVYVP-194.11	633.32	2	128.09	194.11	3.82	93.03				🔍
14	🟧	257.14-SYSVVVY-128.1	633.32	2	257.14	128.10	4.19	92.67				🔍

There are several rows of results, and these are color coded. At the top, in brown, are the results by DirecTag (revealed when you hover over the colored box). Note that the results are all quite short.

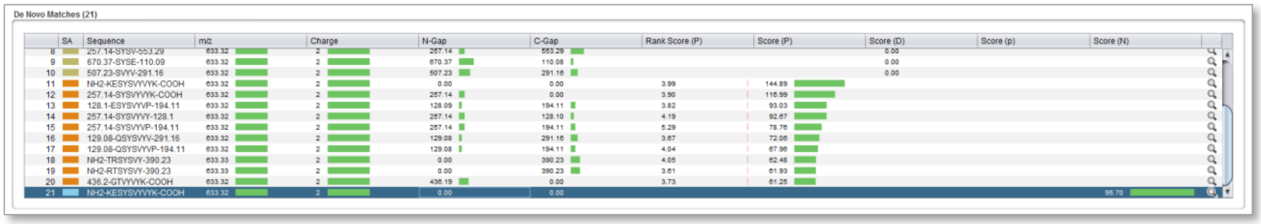
Why are DirecTag results so short? Is there something special about the way that DirecTag works? [1.7d]

The orange box coded results are derived from PepNovo+. Note that some of these PepNovo+ results have N-terminal and/or C-terminal gaps, while others do not (notably, the highest scoring and thus first listed peptide sequence is complete, including termini).

Where do these terminal gaps come from? In what way does a terminally gapped sequence resemble a sequence tag? [1.7e]

If you scroll all the way to the bottom, you can see a single blue box coded Novor result.

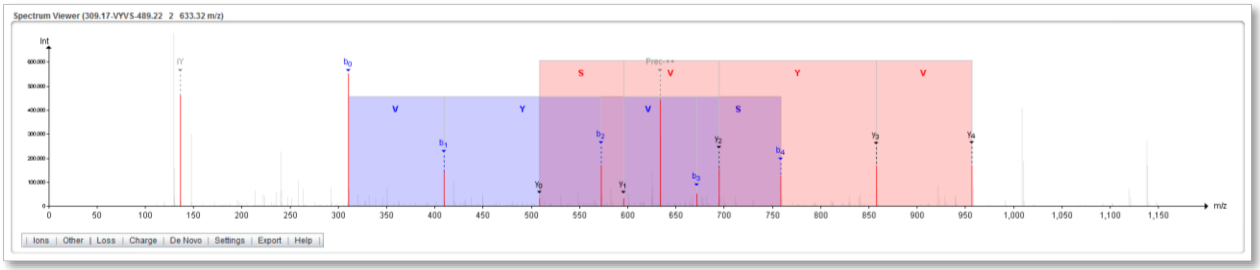
Notice the difference between the number of hits provided by Novor, DirecTag and PepNovo+. Does this make sense? Also, do the three tools agree on the potential sequence for this spectrum? [1.7f]



De Novo Matches (21)

	SA	Sequence	m/z	Charge	N-Gap	C-Gap	Rank Score (P)	Score (P)	Score (D)	Score (p)	Score (N)
8	🟨	257.14-SYSV-553.29	633.32	2	257.14	553.29			0.00		
9	🟨	670.37-SYSE-110.09	633.32	2	670.37	110.08			0.00		
10	🟨	507.23-SVYV-291.16	633.32	2	507.23	291.16			0.00		
11	🟧	NH2-KESYSVYVYK-COOH	633.32	2	0.00	0.00	3.99	144.89			
12	🟧	257.14-SYSVYVYK-COOH	633.32	2	257.14	0.00	3.90	116.99			
13	🟧	128.1-ESYSVYVP-194.11	633.32	2	128.09	194.11	3.82	93.03			
14	🟧	257.14-SYSVYVY-128.1	633.32	2	257.14	128.10	4.19	92.67			
15	🟧	257.14-SYSVYVP-194.11	633.32	2	257.14	194.11	5.29	78.76			
16	🟧	129.08-QSYSVYV-291.16	633.32	2	129.08	291.16	3.67	72.06			
17	🟧	129.08-QSYSVYVP-194.11	633.32	2	129.08	194.11	4.04	67.96			
18	🟧	NH2-TRSYSVY-390.23	633.33	2	0.00	390.23	4.05	62.48			
19	🟧	NH2-RTSYSVY-390.23	633.33	2	0.00	390.23	3.61	61.93			
20	🟧	436.2-GTVYVYK-COOH	633.32	2	436.19	0.00	3.73	61.25			
21	🟦	NH2-KESYSVYVYK-COOH	633.32	2	0.00	0.00					95.70

Now inspect the 6th and 10th result for this spectrum, which are both sequence tags provided by DirecTag: "309.17-VYVS-489.22" and "507.23-SVYV-291.16", respectively. When you select the 6th result, you see the below spectrum matching:



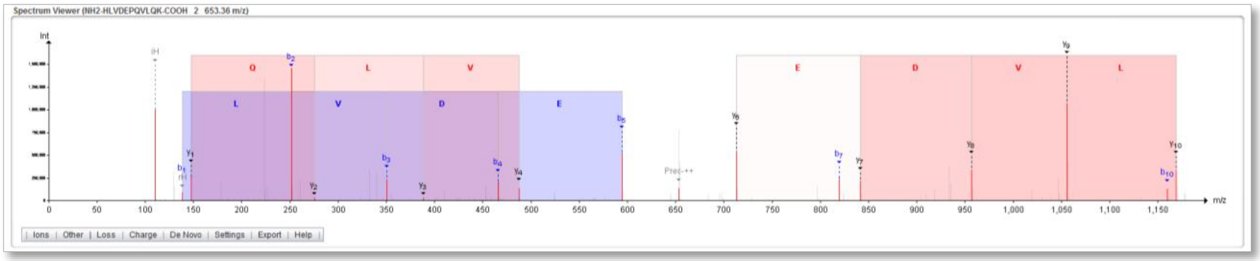
Spectrum Viewer (309.17-VYVS-489.22 2 633.32 m/z)

Why do you see two types of annotation crop up? Which one of these corresponds to the actual tag that we clicked? [1.7g]

Now select the 10th result for this spectrum.

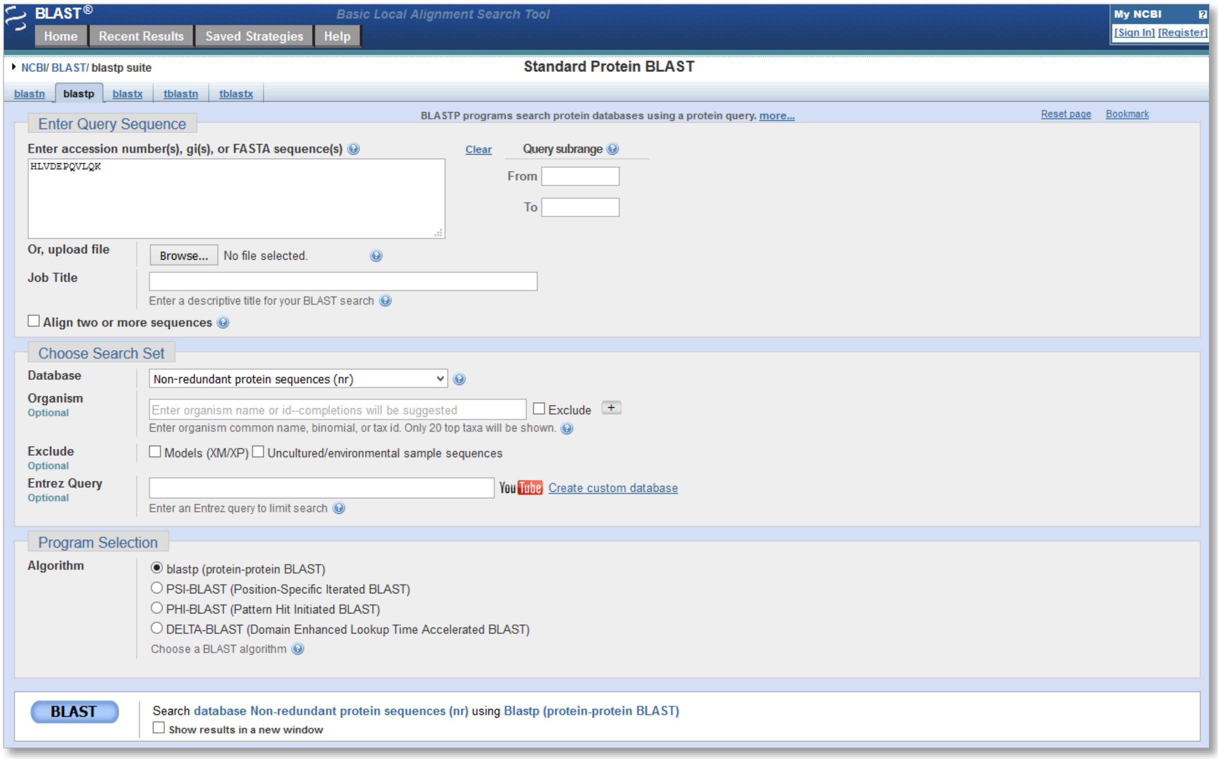
Compare the two spectrum annotations by selecting both the 6th and the 10th hit by holding down the Ctrl key. What do you see? Does this make sense? What does this tell you about sequence tags? [1.7h]

An interesting feature of the Novor algorithm is its residue-specific score. By using the Find feature in the upper right corner of the screen, find the spectrum annotated as HLVDEPQVLQK, and then highlight the blue boxed Novor result in the middle panel for this spectrum. Look closely at the spectrum.



Do you notice anything special about the annotation? What could be represented? [1.7i]

There are several ways to work with the results. For individual hits, a BLAST search can be performed quickly and easily by clicking the magnifying glass icon at the far right in the middle table. For instance, clicking this icon for the Novor result that we have just been inspecting, yields this BLAST page:



Standard Protein BLAST

BLASTP programs search protein databases using a protein query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)   Clear   Query subrange

HLVDEPQVLQK

From

To

Or, upload file   Browse... No file selected.

Job Title

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database   Non-redundant protein sequences (nr)

Organism Optional   Enter organism name or id--completions will be suggested   □ Exclude   +

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude Optional   □ Models (XM/XP) □ Uncultured/environmental sample sequences

Entrez Query Optional   YouTube Create custom database

Enter an Entrez query to limit search

Program Selection

Algorithm

● blastp (protein-protein BLAST)
○ PSI-BLAST (Position-Specific Iterated BLAST)
○ PHI-BLAST (Pattern Hit Initiated BLAST)
○ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

BLAST   Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)

□ Show results in a new window

If you want to perform a more extensive BLAST search, you can also use the 'Export' menu, and then select the 'BLAST' option. This creates a FASTA text file that can be directly submitted to BLAST, in which each peptide is a separate entry. The header for each of these entries contains the *de novo* identification information.

Why is this link with BLAST useful? Is it a good idea to use a regular BLAST query? Are there any things to keep in mind when doing a BLAST with a de novo sequence? [1.7j]

There is a further export option that can be very useful, and that is the 'Export', 'Tag Matches' option. This yields a tab-delimited list of all results obtained from your *de novo* identification.

Bonus exercise for the intrepid reader:

Use the 'Tag Matches' export in the tab-delimited format along with a similar export from the original SearchGUI/PeptideShaker identifications from the same spectrum file, and compare the results for each spectrum, and across the whole data set.

If you encounter any issues with DeNovoGUI, please consult the troubleshooting section at: http://compomics.github.io/projects/denovogui.html.

References

1. Ma, B., Novor: real-time peptide de novo sequencing software. *J. Am. Soc. Mass Spectrom.* **26**, 1885-1894 (2015).
2. Tabb, D.L., Ma, Z., Martin, D.B., Ham, A.L. & Chambers, M.C., Directag: accurate sequence tags from peptide ms/ms through statistical scoring. *J. Proteome Res.* **7**, 3838-3846 (2008).
3. Frank, A.M., Predicting intensity ranks of peptide fragment ions. *J. Proteome Res.* **8**, 2226-2240 (2009).
4. Chi, H., Chen, H., He, K., Wu, L., Yang, B. et al., Pnovo+: de novo peptide sequencing using complementary hcd and etd tandem mass spectra. *J. Proteome Res.* **12**, 615-625 (2013).
5. Muth, T., Weilnböck, L., Rapp, E., Huber, C.G., Martens, L. et al., Denovogui: an open source graphical user interface for de novo sequencing of tandem mass spectra. *J. Proteome Res.* **13**, 1143-1146 (2014).