DeNovoGUI: an open source graphical user interface for de novo sequencing of tandem mass spectra.

Thilo Muth[1,$], Lisa Weilnböck[2,$], Erdmann Rapp[1], Christian G. Huber[2], Lennart Martens[3,4], Marc Vaudel[5]* and Harald Barsnes[5].

[1] Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg, Germany.

[2] Department of Molecular Biology, University of Salzburg, Austria.

[3] Department of Biochemistry, Ghent University, Ghent, Belgium.

[4] Department of Medical Protein Research, VIB, Ghent, Belgium.

[5] Proteomics Unit, Department of Biomedicine, University of Bergen, Norway.

[$] These authors contributed equally to the work.

* Corresponding author: Marc Vaudel, Proteomics Unit, Department of Biomedicine, University of Bergen, Jones Liesvei 91, N-5009 Bergen, Norway. Email: marc.vaudel@biomed.uib.no. Tel: +47 55 58 63 78.

Keywords: bioinformatics, de novo, mass spectrometry, PepNovo+, peptide identification

Abbreviations:

MS	Mass Spectrometry
MS/MS	Tandem Mass Spectrometry

Abstract

De novo sequencing is a popular technique in proteomics for identifying peptides from tandem mass spectra without having to rely on a protein sequence database. Despite their strong potential, the adoption threshold of *de novo* sequencing algorithms still remains quite high. We here present a user-friendly and lightweight graphical user interface called DeNovoGUI for running parallelized versions of the freely available *de novo* sequencing software PepNovo+, greatly simplifying the use of *de novo* sequencing in proteomics. Our platform independent software is freely available under the permissible Apache2 open source license. Source code, binaries and additional documentation are available at http://denovogui.googlecode.com.

Main text

Mass spectrometry (MS) based proteomics is an efficient high-throughput method for the analysis of peptides and proteins[1, 2]. However, in a typical tandem mass spectrometry (MS/MS) experiment a high proportion of the mass spectra remain unidentified when matched against *in silico* generated spectra, derived from peptides obtained through *in silico* proteolytic digest of known protein sequences[3]. Some of these unidentified spectra derive from contaminants and low quality spectra, but the rest is likely to contain unexpected peptides[4]. One obstacle for the successful identification of such peptides is the fact that protein sequence databases are incomplete, as many organisms have not yet been sequenced, an issue that is particularly strongly felt in challenging fields such as metaproteomics[5] or plant proteomics[6]. Another common issue is the presence of unknown or unexpected modifications on the peptide precursors[4]. *De novo* sequencing constitutes a powerful technique to overcome such issues and successfully assign high-quality unidentified spectra to peptides. Moreover, *de novo* derived peptide sequences can be used for the validation of insignificant database search results, for instance proteins backed merely by a single peptide identification, so-called "one hit wonders"[7].

Several *de novo* algorithms have been described and evaluated in the literature[8, 9], including the commercial PEAKS[10] software suite. PepNovo+[11] on the other hand is a powerful, freely available software tool. However, as with most open source *de novo* algorithms, it comes with several shortcomings: (i) it is only distributed with a command line interface, thus requiring advanced computational skills to operate; (ii) modifications need to be configured manually for every search and are not based on the standardized PSI-MOD[12] controlled vocabulary; (iii) the search is not parallelized when multiple cores are available; and (iv) the output of the algorithm is a text file containing only the derived sequences and their scores, thus omitting additional useful information such as fragment ions and spectrum annotation. Because of these issues, user validation of the results is cumbersome, and standardized dissemination of results is quite difficult.

Here, we describe an intuitive, end-user oriented front-end to the PepNovo+ algorithm called DeNovoGUI, which aims to solve the aforementioned issues. Similar to the SearchGUI[13] software for the OMSSA[14] and X!Tandem[15] database search algorithms, DeNovoGUI provides a self-contained and easily adopted solution for convenient and efficient *de novo* sequencing using the PepNovo+ algorithm. The processing of a large amount of spectra has been accelerated by automated and completely transparent parallelization across multiple cores, a crucial feature for modern computers which typically come equipped with two to eight (hyperthreaded) cores.

DeNovoGUI can be installed with minimal effort by downloading the latest release from the tool website (http://denovogui.googlecode.com), subsequently unzipping the downloaded file, and then double-clicking the DeNovoGUI jar file. In order to start the *de novo* procedure, the user only has to provide the spectrum files to analyze (in the standard mgf format), the settings to use, and the desired output folder to store the *de novo* results in (**Figure 1A**). The settings include the fragment and precursor ion mass tolerances, as well as the fixed and variable post-translational modifications to consider. Furthermore, additional settings to fine tune the PepNovo+ algorithm can be specified, including the number of *de novo* solutions to provide for each spectrum. **Figure 1B** provides a complete overview of the available settings. Importantly, the handling of modifications is greatly simplified by the graphical user interface: user defined modifications can easily be created from the Edit menu in the main frame. Note that DeNovoGUI allows all settings to be saved for later reuse or batch entry.

As soon as the settings and input files have been provided, the *de novo* sequencing can be initiated by clicking the “Start Sequencing!” button in the main DeNovoGUI interface. While the PepNovo+ algorithm is running, the user is continuously informed about the status of the sequencing and a progress bar is displayed to indicate overall progress. When completed, the *de novo* sequencing results are stored in the provided output folder in a simple text based format, and the detailed results can be visualized in the DeNovoGUI interface (see **Figure 2**). At the top, the user can browse through all the input spectra in the ’Query Spectra’ table, and through the *de novo* peptide matches for the selected spectrum in the ‘De Novo Peptides’ table. The ‘Query Spectra’ table provides information collected from the original spectra, such as title, precursor m/z, charge and identification state, while the ‘De Novo Peptides’ table shows details obtained from the *de novo* sequencing results: peptide sequence, scores, and terminal gaps and precursor m/z and charge. At the bottom, a spectrum viewer[16] shows the currently selected spectrum with the fragment ion annotation corresponding to the selected *de novo* peptide solution. A sequence overlay is also presented on the spectrum by default, aiding the efficient validation of the proposed peptide solution. The *de novo* results can be validated using BLAST, either by clicking the BLAST option at the end of a given line in the 'De Novo Peptides' table or by exporting a list of matches in a BLAST compatible format *via* the Export menu. Peptide matches can also be exported in a simple text based format from the same menu.

A reference dataset is provided as example in DeNovoGUI and can be opened easily from the main menu. It consists of 30,289 MS/MS spectra from an *Arabidopsis thaliana* whole leaf proteome. The obtained tryptic peptides were separated *via* ion-pair reversed-phase high-performance liquid chromatography on a poly-(styrene/divinylbenzene) monolithic colum[17] using a 5 hour gradient and were measured on an LTQ Orbitrap XL mass spectrometer using high-resolution precursor ion selection followed by CID fragmentation. This reference dataset of a well-established plant model system represents a realistic study case for plant proteomics and is thus ideally suited for the benchmarking of *de novo* sequencing algorithms. For further details on the dataset we refer to the **Supporting Information**.

Due its ability to spread the *de novo* task across multiple compute cores and/or hyperthreads, DeNovoGUI substantially reduces the time required to analyze large amounts of spectra using PepNovo+. Indeed, while analyzing our 30,289 MS/MS spectra took around 7 hours using only a single thread, the running time was reduced to approximately 3 hours using four threads and to approximately 1.5 hours using eight threads. In order to obtain comparable and consistent results, running times were measured on identical virtual machines with the desired number of cores set (Intel Xeon CPU X5660 @ 2.80GHz). These tests clearly show that the multithreading capability of DeNovoGUI results in substantial reductions in processing time on today's multi-threading, multi-core laptop and desktop computers.

Upon downloading, DeNovoGUI comes with the latest version of PepNovo+ included and apart from unzipping the downloaded DeNovoGUI zip file, no further installation is required to run the software. DeNovoGUI is written in the Java programming language and is freely available as open source under the permissive Apache2 license. Documentation, source files and binaries can be downloaded from http://denovogui.googlecode.com.

Acknowledgements

T.M. and E.R. acknowledge the support by Max Planck Society. L.W. and C.G.H. appreciate the support by the Austrian Science Funding Agency under the project W_01213. L.M. acknowledges the support of Ghent University (Multidisciplinary Research Partnership "Bioinformatics: from nucleotides to networks"), the PRIME-XS project, grant agreement number 262067, and the 'ProteomeXchange' project, grant agreement number 260558, both funded by the European Union 7th Framework Program. H.B. is supported by the Research Council of Norway. The authors would like to thank Robert Behmüller and Raimund Tenhaken for providing the Arabidopsis thaliana plants, as well as An Staes, Evy Timmerman and Davy Maddelein for their support.

The authors declare no conflict of interest.

Supporting Information Available: This material is available free of charge via the Internet at http://pubs.acs.org.

References

1. Gevaert, K.; Van Damme, P.; Ghesquiere, B.; Impens, F.; Martens, L.; Helsens, K.; Vandekerckhove, J., A la carte proteomics with an emphasis on gel-free techniques. *Proteomics* **2007,** 7, (16), 2698-718.
2. Aebersold, R.; Mann, M., Mass spectrometry-based proteomics. *Nature* **2003,** 422, (6928), 198-207.
3. Nesvizhskii, A. I.; Roos, F. F.; Grossmann, J.; Vogelzang, M.; Eddes, J. S.; Gruissem, W.; Baginsky, S.; Aebersold, R., Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data: toward more efficient identification of post-translational modifications, sequence polymorphisms, and novel peptides. *Mol Cell Proteomics* **2006,** 5, (4), 652-70.
4. Flikka, K.; Martens, L.; Vandekerckhove, J.; Gevaert, K.; Eidhammer, I., Improving the reliability and throughput of mass spectrometry-based proteomics by spectrum quality filtering. *Proteomics* **2006,** 6, (7), 2086-2094.
5. Muth, T.; Benndorf, D.; Reichl, U.; Rapp, E.; Martens, L., Searching for a needle in a stack of needles: challenges in metaproteomics data analysis. *Mol Biosyst* **2012**.
6. Castellana, N. E.; Payne, S. H.; Shen, Z.; Stanke, M.; Bafna, V.; Briggs, S. P., Discovery and revision of Arabidopsis genes by proteogenomics. *Proceedings of the National Academy of Sciences of the United States of America* **2008,** 105, (52), 21034-8.
7. Veenstra, T. D.; Conrads, T. P.; Issaq, H. J., What to do with "one-hit wonders"? *Electrophoresis* **2004,** 25, (9), 1278-9.
8. Allmer, J., Algorithms for the de novo sequencing of peptides from tandem mass spectra. *Expert Rev Proteomics* **2011,** 8, (5), 645-57.
9. Pevtsov, S.; Fedulova, I.; Mirzaei, H.; Buck, C.; Zhang, X., Performance evaluation of existing de novo sequencing algorithms. *J Proteome Res* **2006,** 5, (11), 3018-28.

10. Ma, B.; Zhang, K.; Hendrie, C.; Liang, C.; Li, M.; Doherty-Kirby, A.; Lajoie, G., PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom* **2003,** 17, (20), 2337-2342.

11. Frank, A.; Pevzner, P., PepNovo: de novo peptide sequencing via probabilistic network modeling. *Anal Chem* **2005,** 77, (4), 964-973.

12. Montecchi-Palazzi, L.; Beavis, R.; Binz, P. A.; Chalkley, R. J.; Cottrell, J.; Creasy, D.; Shofstahl, J.; Seymour, S. L.; Garavelli, J. S., The PSI-MOD community standard for representation of protein modification data. *Nat Biotechnol* **2008,** 26, (8), 864-6.

13. Vaudel, M.; Barsnes, H.; Berven, F. S.; Sickmann, A.; Martens, L., SearchGUI: An open-source graphical user interface for simultaneous OMSSA and X!Tandem searches. *Proteomics* **2011,** 11, (5), 996-9.

14. Geer, L. Y.; Markey, S. P.; Kowalak, J. A.; Wagner, L.; Xu, M.; Maynard, D. M.; Yang, X.; Shi, W.; Bryant, S. H., Open mass spectrometry search algorithm. *J Proteome Res* **2004,** 3, (5), 958-964.

15. Craig, R.; Beavis, R. C., TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **2004,** 20, (9), 1466-1467.

16. Barsnes, H.; Vaudel, M.; Colaert, N.; Helsens, K.; Sickmann, A.; Berven, F. S.; Martens, L., compomics-utilities: an open-source Java library for computational proteomics. *BMC Bioinformatics* **2011,** 12, 70.

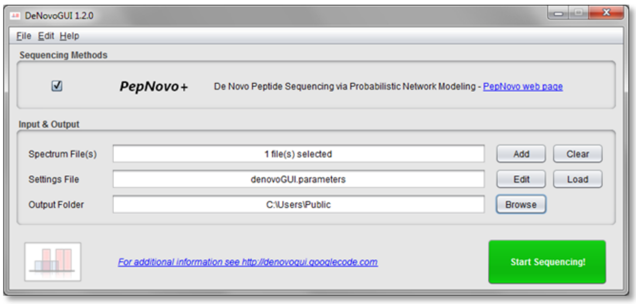
17. Walcher, W.; Oberacher, H.; Troiani, S.; Hölzl, G.; Oefner, P.; Zolla, L.; Huber, C. G., Monolithic Capillary Columns for Liquid Chromatography-Electrospray Ionization Mass Spectrometry in Proteomic and Genomic Research. *J. Chromatogr. B* **2002,** 782, 111-125.

Figure legends

Figure 1: (A) The main DeNovoGUI interface allows the user to input the spectrum files, the settings, and the output folder for the results. (B) The *de novo* sequencing settings dialog allows the user to specify the fragment ion and precursor mass tolerances, and the fixed and variable post-translational modifications. Additional settings to fine tune the PepNovo+ algorithm can also be configured.

Figure 2: The DeNovoGUI *de novo* results viewer shows the currently selected *de novo* peptide solution and its corresponding fragment ion annotations on the selected spectrum. The ‘Query Spectra’ section at the top allows the user to browse through the input spectra, while the ‘De Novo Peptides’ section below provides sequence and scoring information for all peptide solutions for the currently selected input spectrum.

Figure 1



A

DeNovoGUI 1.2.0

File Edit Help

Sequencing Methods

☑ *PepNovo+* De Novo Peptide Sequencing via Probabilistic Network Modeling - PepNovo web page

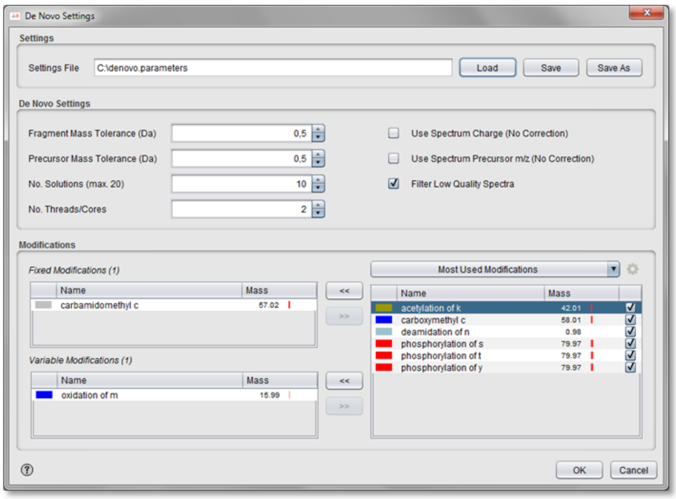
Input & Output

Spectrum File(s): 1 file(s) selected

Settings File: denovoGUI.parameters

Output Folder: C:\Users\Public

Add
Clear
Edit
Load
Browse
For additional information see http://denovogui.googlecode.com
Start Sequencing!



B

De Novo Settings

Settings

Settings File: C:\denovo.parameters

De Novo Settings

Fragment Mass Tolerance (Da): 0,5

Precursor Mass Tolerance (Da): 0,5

No. Solutions (max. 20): 10

No. Threads/Cores: 2

☐ Use Spectrum Charge (No Correction)

☐ Use Spectrum Precursor m/z (No Correction)

☑ Filter Low Quality Spectra

Modifications

Fixed Modifications (1)

Name	Mass
carbamidomethyl c	57.02

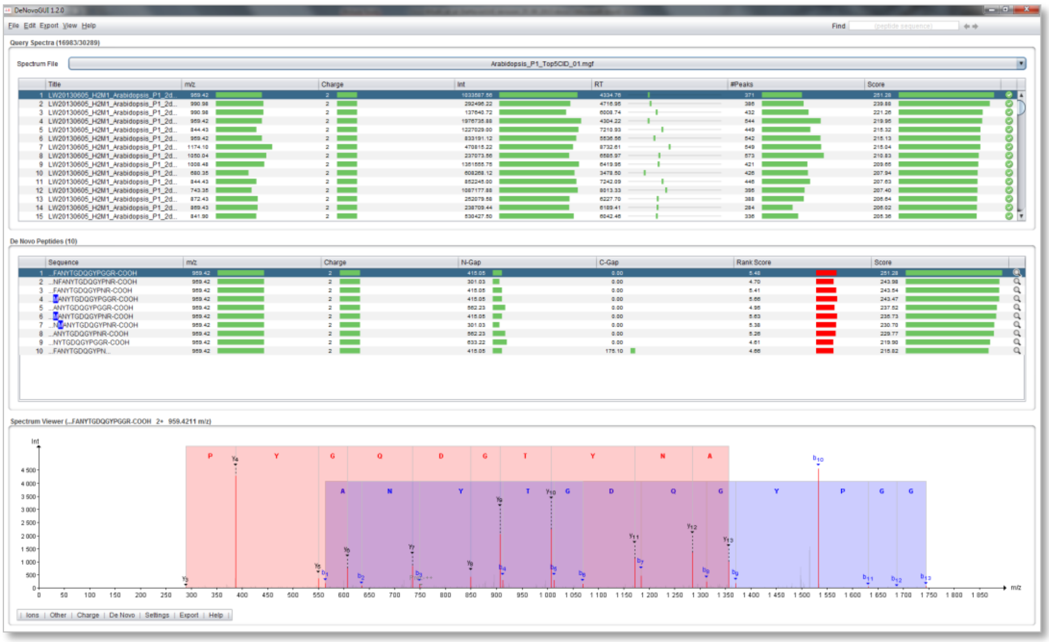
Variable Modifications (1)

Name	Mass
oxidation of m	15.99

Most Used Modifications

Name	Mass	
acetylation of k	42.01	☑
carboxymethyl c	58.01	☑
deamidation of n	0.98	☑
phosphorylation of s	79.97	☑
phosphorylation of t	79.97	☑
phosphorylation of y	79.97	☑

Figure 2



DeNovoGUI 1.2.0

File Edit Export View Help

Find (peptide sequence)

Query Spectra (16983/30289)

Spectrum File: Arabidopsis_P1_Top5CID_01.mgf

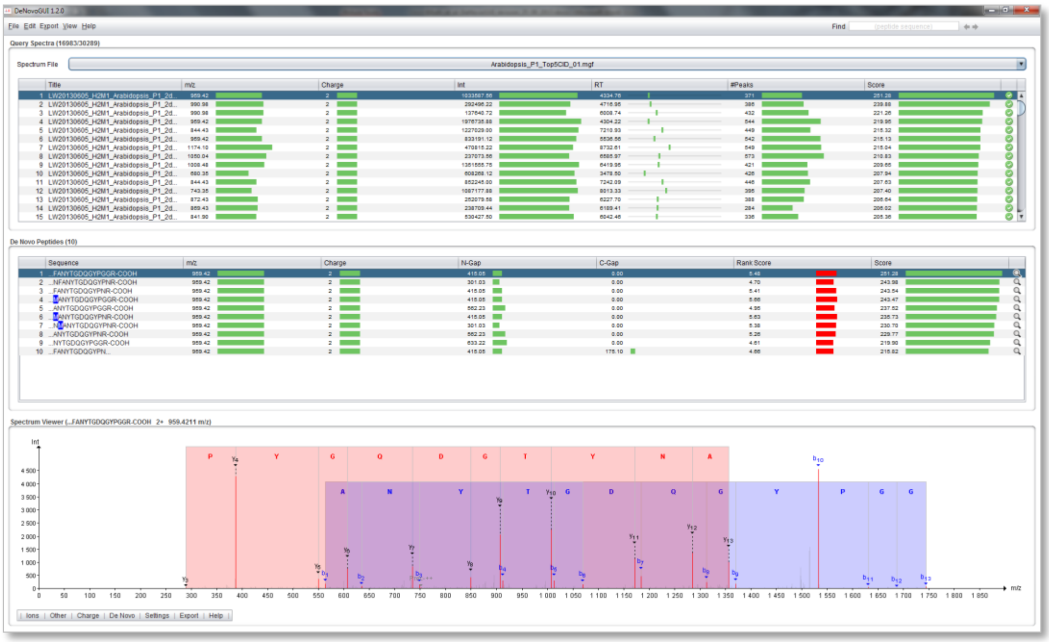
	Title	m/z	Charge	Int	RT	#Peaks	Score
1	LW20130605_H2M1_Arabidopsis_P1_2d...	959.42	2	1033587.56	4334.76	371	251.28
2	LW20130605_H2M1_Arabidopsis_P1_2d...	990.98	2	292496.22	4716.96	386	239.88
3	LW20130605_H2M1_Arabidopsis_P1_2d...	990.98	2	137640.72	6008.74	432	221.26
4	LW20130605_H2M1_Arabidopsis_P1_2d...	959.42	2	1976735.88	4304.22	544	219.96
5	LW20130605_H2M1_Arabidopsis_P1_2d...	844.43	2	1227029.00	7210.93	449	215.32
6	LW20130605_H2M1_Arabidopsis_P1_2d...	959.42	2	833191.12	5536.56	542	215.13
7	LW20130605_H2M1_Arabidopsis_P1_2d...	1174.10	2	470815.22	8732.61	549	215.04
8	LW20130605_H2M1_Arabidopsis_P1_2d...	1050.04	2	237073.56	6585.97	573	210.83
9	LW20130605_H2M1_Arabidopsis_P1_2d...	1008.48	2	1351555.76	6419.96	421	209.66
10	LW20130605_H2M1_Arabidopsis_P1_2d...	680.35	2	608268.12	3478.50	426	207.94
11	LW20130605_H2M1_Arabidopsis_P1_2d...	844.43	2	852245.00	7242.09	446	207.63
12	LW20130605_H2M1_Arabidopsis_P1_2d...	743.35	2	1087177.88	8013.33	395	207.40
13	LW20130605_H2M1_Arabidopsis_P1_2d...	872.43	2	252079.58	6227.70	388	206.64
14	LW20130605_H2M1_Arabidopsis_P1_2d...	869.43	2	238709.44	6189.41	284	206.02
15	LW20130605_H2M1_Arabidopsis_P1_2d...	841.90	2	530427.50	6042.46	336	205.36

De Novo Peptides (10)

	Sequence	m/z	Charge	N-Gap	C-Gap	Rank Score	Score
1	...FANYTGDQGYPGGR-COOH	959.42	2	415.05	0.00	5.48	251.28
2	...NFANYTGDQGYPNR-COOH	959.42	2	301.03	0.00	4.70	243.98
3	...FANYTGDQGYPNR-COOH	959.42	2	415.05	0.00	5.41	243.54
4	...MANYTGDQGYPGGR-COOH	959.42	2	415.05	0.00	5.66	243.47
5	...ANYTGDQGYPGGR-COOH	959.42	2	562.23	0.00	4.96	237.52
6	...MANYTGDQGYPNR-COOH	959.42	2	415.05	0.00	5.63	236.73
7	...NMANYTGDQGYPNR-COOH	959.42	2	301.03	0.00	5.38	230.70
8	...ANYTGDQGYPNR-COOH	959.42	2	562.23	0.00	5.26	229.77
9	...NYTGDQGYPGGR-COOH	959.42	2	633.22	0.00	4.61	219.90
10	...FANYTGDQGYPN...	959.42	2	415.05	175.10	4.66	215.82

Spectrum Viewer (...FANYTGDQGYPGGR-COOH 2+ 959.4211 m/z)

Table of Contents/Abstract Graphic



Query Spectra (16983/30289)

Spectrum File: Arabidopsis_P1_Top5CID_01.mgf

	Title	m/z	Charge	Int	RT	#Peaks	Score	
1	LW20130605_H2M1_Arabidopsis_P1_2d...	959.42	2	1033587.56	4334.76	371	251.28	✓
2	LW20130605_H2M1_Arabidopsis_P1_2d...	990.98	2	292496.22	4716.96	386	239.88	✓
3	LW20130605_H2M1_Arabidopsis_P1_2d...	990.98	2	137640.72	6008.74	432	221.26	✓
4	LW20130605_H2M1_Arabidopsis_P1_2d...	959.42	2	1976735.88	4304.22	544	219.96	✓
5	LW20130605_H2M1_Arabidopsis_P1_2d...	844.43	2	1227029.00	7210.93	449	215.32	✓
6	LW20130605_H2M1_Arabidopsis_P1_2d...	959.42	2	833191.12	5536.56	542	215.13	✓
7	LW20130605_H2M1_Arabidopsis_P1_2d...	1174.10	2	470815.22	8732.61	549	215.04	✓
8	LW20130605_H2M1_Arabidopsis_P1_2d...	1050.04	2	237073.56	6585.97	573	210.83	✓
9	LW20130605_H2M1_Arabidopsis_P1_2d...	1008.48	2	1361565.75	6419.96	421	209.66	✓
10	LW20130605_H2M1_Arabidopsis_P1_2d...	680.35	2	608268.12	3478.50	426	207.94	✓
11	LW20130605_H2M1_Arabidopsis_P1_2d...	844.43	2	852245.00	7242.09	446	207.63	✓
12	LW20130605_H2M1_Arabidopsis_P1_2d...	743.35	2	1087177.88	8013.33	396	207.40	✓
13	LW20130605_H2M1_Arabidopsis_P1_2d...	872.43	2	252079.58	6227.70	388	206.64	✓
14	LW20130605_H2M1_Arabidopsis_P1_2d...	869.43	2	238709.44	6189.41	284	206.02	✓
15	LW20130605_H2M1_Arabidopsis_P1_2d...	841.90	2	530427.50	6042.46	336	205.36	✓

De Novo Peptides (10)

	Sequence	m/z	Charge	N-Gap	C-Gap	Rank Score	Score	
1	...FANYTGDQGYPGGR-COOH	959.42	2	415.05	0.00	5.48	251.28	🔍
2	...NFANYTGDQGYPNR-COOH	959.42	2	301.03	0.00	4.70	243.98	🔍
3	...FANYTGDQGYPNR-COOH	959.42	2	415.05	0.00	5.41	243.64	🔍
4	...MANYTGDQGYPGGR-COOH	959.42	2	415.05	0.00	5.66	243.47	🔍
5	...ANYTGDQGYPGGR-COOH	959.42	2	562.23	0.00	4.96	237.62	🔍
6	...MANYTGDQGYPNR-COOH	959.42	2	415.05	0.00	5.63	236.73	🔍
7	...NMANYTGDQGYPNR-COOH	959.42	2	301.03	0.00	5.38	230.70	🔍
8	...ANYTGDQGYPNR-COOH	959.42	2	562.23	0.00	5.26	229.77	🔍
9	...NYTGDQGYPGGR-COOH	959.42	2	633.22	0.00	4.61	219.90	🔍
10	...FANYTGDQGYPN...	959.42	2	415.05	175.10	4.66	215.82	🔍

Spectrum Viewer (...FANYTGDQGYPGGR-COOH 2+ 959.4211 m/z)