Bioinformatics for Proteomics

Mass spectrometry based proteomics studies typically aim at measuring and comparing changes in biological samples. This however strongly relies on our ability to interpret the massive amounts of data produced by modern mass spectrometers. Therefore, methods and tools have been developed to tackle three major topics:

1. Identification
2. Functional Analysis
3. Data Sharing

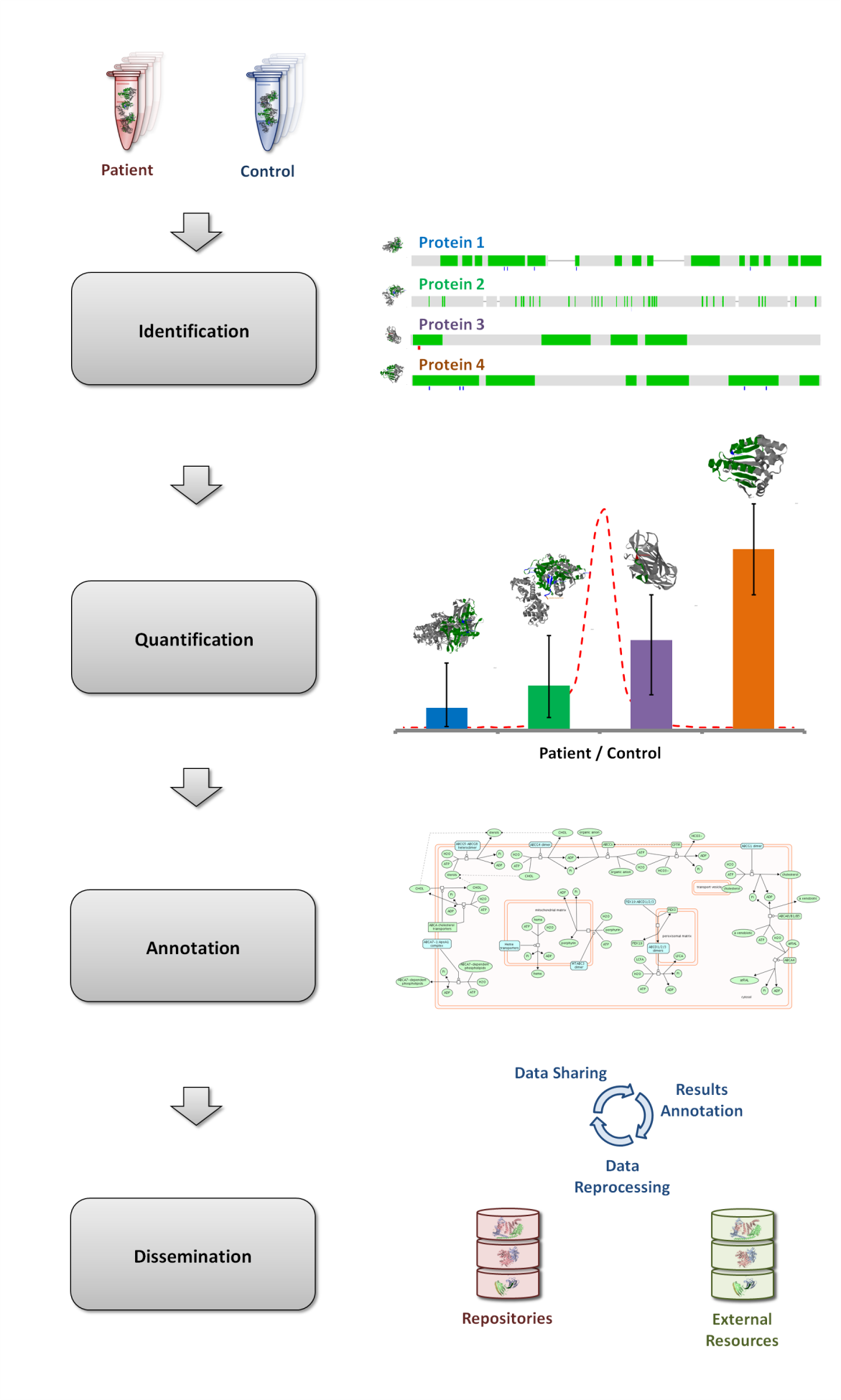
This tutorial aims at providing the basis for any user to go through the following workflow: (1) identify peptides, proteins and their modifications, (2) annotate the data with existing biological knowledge, and (3) share the data using online repositories.

Notably, we will only employ user friendly and open source software applicable to any kind of mass spectrometer. Despite our best efforts however, such software, or the resources they rely on, are not always available yet. For example, proteomics resources are heavily focused on a few model organisms, and working with data from other species is a lot more challenging.

Finally, if you encounter issues not tackled by the present tutorial, feel free to contact the authors.

For further reading about bioinformatics for proteomics we recommend:

* **Computational Methods for Mass Spectrometry Proteomics**, 2007, Wiley-VCH,   
  Ingvar Eidhammer, Kristian Flikka, Lennart Martens, Svein-Ole Mikalsen.
* **Computational and Statistical Methods for Protein Quantification by Mass Spectrometry**, 2012, Wiley-VCH, Ingvar Eidhammer, Harald Barsnes, Geir Egil Eide, Lennart Martens.



There are several tools that can be used in a proteomics workflow as outlined above. A detailed list (mainly focused on free software) is given in the appendix below. Among these, we recommend the use of the following tools:

1. To convert raw files we recommend **MSConvert** as part of the **Proteowizard[1](#_ENREF_1" \o "Kessner, 2008 #14)** package (<http://proteowizard.sourceforge.net>). We recommend **OpenMS[2](#_ENREF_2" \o "Sturm, 2008 #2)** for files manipulation and **TOPPAS**[**3**](#_ENREF_3) for workflow design (<http://openms.de>).
2. We strongly recommend the use of **UniProt[4](#_ENREF_4" \o "Apweiler, 2004 #34)** ([www.uniprot.org](file:///C:\Users\hba041\Teaching\WT%20Course%20EBI%202012\www.uniprot.org)) databases, and for their manipulation **dbtoolkit[5](#_ENREF_5" \o "Martens, 2005 #19)** (<http://dbtoolkit.googlecode.com>).
3. To match peptides to spectra, we will here use two distinct, freely available search engines: **OMSSA**[**6**](#_ENREF_6) and **X!Tandem[7](#_ENREF_7" \o "Craig, 2004 #46)**, both of which are made easily accessible *via* a free tool called **SearchGUI[8](#_ENREF_8" \o "Vaudel, 2011 #17)** ([http://searchgui.googlecode.com](http://dbtoolkit.googlecode.com)).
4. To analyze the search results, and to do the peptide and protein inference, we recommend the use of **PeptideShaker** (<http://peptide-shaker.googlecode.com>).
5. For the validation of the identifications we also recommend the use of **PeptideShaker** (<http://peptide-shaker.googlecode.com>).
6. Many external resources are available on the internet. Among these we will use:   
   **UniProt**[4](#_ENREF_4" \o "Apweiler, 2004 #34) (<http://www.uniprot.org>), **Reactome**[9](#_ENREF_9" \o "Haw, 2011 #139) (<http://www.reactome.org>),   
   **PICR**[10](#_ENREF_10) (<http://www.ebi.ac.uk/Tools/picr>) and **Dasty**[11](#_ENREF_11" \o "Jones, 2005 #81) (<http://www.ebi.ac.uk/dasty>).   
   Note that additional resources are listed in **PeptideShaker**, and will also be used to conduct the gene ontology analysis of the data.
7. In order to make your proteomics data publicly available, you can upload them in public repositories. We recommend **ProteomeXchange** (<http://proteomexchange.org>) and **PRIDE**[**12**](#_ENREF_12) (<http://www.ebi.ac.uk/pride>).

You will find a folder named software containing all the software needed for this tutorial as well as folders corresponding to the different chapters. Although it is recommended to follow the tutorial in its entirety, the chapters can also be followed independently. For every chapter, the resources folder contained in the chapter folder provides all the required files. If you experience any difficulty with a program, feel free to contact the developers. We will also welcome any feedback on these tutorials.

***All chapters are also available online:***[***http://compomics.com/bioinformatics-for-proteomics***](http://compomics.com/bioinformatics-for-proteomics-tutorial)

Appendix: Proteomics Software

The concepts and methods introduced in the present tutorial can also be applied to other software and packages. The table below provides a (non-exhaustive) list of software dedicated to proteomics, with brief descriptions and corresponding references that will help you to get started.

|  |  |  |
| --- | --- | --- |
| Type | Software | Description |
| Converter | ProteoWizard[1](#_ENREF_1" \o "Kessner, 2008 #14) | Converter accepting most mass spectrometer proprietary formats and converting them into open formats |
| mzML parser | jmzML[13](#_ENREF_13" \o "Cote, 2010 #150) | Mass spectrometry mzML file parser |
| General proteomics package | OpenMS[14](#_ENREF_14" \o "Bertsch, 2011 #15) | Package of tools for proteomics allowing the design  of workflows with a graphical interface |
| TPP[15](#_ENREF_15) | Package of tools for proteomics mainly command line driven |
| MaxQuant[16](#_ENREF_16" \o "Cox, 2008 #103) | Package for identification and quantification  of entire proteomes |
| Identification post-processor | PeptideShaker | Interpretation of proteomics identifications  from multiple search engines |
| MassSieve[17](#_ENREF_17" \o "Slotta, 2010 #128) | Identification processing software |
| De novo sequencing | Sherenga[18](#_ENREF_18" \o "Dancik, 1999 #42) | De novo sequencing tool |
| PepNovo[19](#_ENREF_19" \o "Frank, 2005 #97) | De novo sequencing tool |
| UniNovo[20](#_ENREF_20" \o "Jeong, 2013 #38) | De novo sequencing tool |
| Antilope[21](#_ENREF_21" \o "Andreotti, 2012 #39) | De novo sequencing tool |
| NovoHMM[22](#_ENREF_22" \o "Fischer, 2005 #40) | De novo sequencing tool |
| LutefiskXP[23](#_ENREF_23" \o "Taylor, 1997 #41) | De novo sequencing tool |
| EigenMS[24](#_ENREF_24" \o "Bern, 2006 #43) | De novo sequencing tool |
| PILOT[25](#_ENREF_25) | De novo sequencing tool |
| PEAKS[26](#_ENREF_26) | De novo sequencing tool (commercial) |
| Tag sequencing | GutenTag[27](#_ENREF_27" \o "Tabb, 2003 #191) | Finds amino acid patterns in spectra |
| DirecTag[28](#_ENREF_28" \o "Tabb, 2008 #190) | Finds amino acid patterns in spectra |
| TagRecon[29](#_ENREF_29" \o "Dasari, 2010 #45) | Matches amino acid patterns in protein databases |
| Database search engine | Sequest[30](#_ENREF_30" \o "Yates, 1995 #99) | Database search engine (commercial) |
| Mascot[31](#_ENREF_31) | Database search engine (commercial) |
| OMSSA[6](#_ENREF_6) | Database search engine |
| X!Tandem[7](#_ENREF_7" \o "Craig, 2004 #46) | Database search engine |
| Morpheus[32](#_ENREF_32) | Database search engine |
| Inspect[33](#_ENREF_33) | Database search engine |
| MyriMatch[34](#_ENREF_34) | Database search engine |
| MassWiz[35](#_ENREF_35" \o "Yadav, 2011 #101) | Database search engine |
| MS-GF+ | Database search engine |
| MS Amanda | Database search engine (Proteome Discoverer only) |
| Andromeda[36](#_ENREF_36) | Database search engine (MaxQuant only) |
| PRIDE Inspector[37](#_ENREF_37) | Graphical interface for the inspection of PRIDE XML files |
| TOPPAS[38](#_ENREF_38) | Graphical interface for the design of OpenMS workflows |
| User friendly interfaces | SearchGUI[8](#_ENREF_8" \o "Vaudel, 2011 #17) | Graphical interface for search engines |
| DenovoGUI[39](#_ENREF_39" \o "Muth, 2013 #46) | Graphical interface for de novo sequencing |
| BumberDash | Graphical interface for tag based processing |
| Spectral library searching | X!Hunter[40](#_ENREF_40" \o "Craig, 2006 #106) | Spectral libraries search engine |
| NIST MS search[41](#_ENREF_41) | Spectral libraries search engine |
| Pepitome[42](#_ENREF_42" \o "Dasari, 2012 #47) | Spectral libraries search engine |
| SpectraST[43](#_ENREF_43" \o "Lam, 2007 #108) | Spectral libraries search engine |
| Identification file parsers | MascotDatFile[44](#_ENREF_44" \o "Helsens, 2007 #59) | Java parser for Mascot .dat files |
| OMSSA parser[45](#_ENREF_45) | Java parser for OMSSA .omx files |
| X!Tandem parser[46](#_ENREF_46) | Java parser for X!Tandem XML files |
| Data structure | compomics-utilities[47](#_ENREF_47) | Java object structure for the handling and visualization  of identifications from different search engines |
| PSM rescoring | Percolator[48](#_ENREF_48) | Machine learning algorithm rescoring  PSMs and attaching them a p-value |
| PeptideProphet[49](#_ENREF_49" \o "Keller, 2002 #27) | Machine learning algorithm attaching  PSMs a PEP (integrated in TPP) |
| PepArML[50](#_ENREF_50" \o "Edwards, 2009 #179) | Machine learning algorithm merging results from  different search engines with web interface: https://edwardslab.bmcb.georgetown.edu/pymsio |
| Database manipulation | dbtoolkit[5](#_ENREF_5" \o "Martens, 2005 #19) | Tool allowing the manipulation of databases  and creation of custom ones |
| Peptide inference | iProphet[51](#_ENREF_51" \o "Shteynberg, 2011 #126) | Tool for statistical post-processing of PSMs  (integrated in TPP) |
| Protein inference | ProteinProphet[52](#_ENREF_52" \o "Nesvizhskii, 2003 #10) | Tool for protein inference (integrated in TPP) |
| IDPicker[53](#_ENREF_53" \o "Ma, 2009 #127) | Tool for protein inference |
| MassSieve[17](#_ENREF_17" \o "Slotta, 2010 #128) | Identification processing software |
| Protein annotation | UniProtKB[4](#_ENREF_4" \o "Apweiler, 2004 #34) | Protein knowledge database |
| Dasty[11](#_ENREF_11" \o "Jones, 2005 #81) | Cross reference tool for protein databases |
| GO enrichment | GOTree[54](#_ENREF_54" \o "Zhang, 2004 #134) | GO enrichment tool |
| Onotologizer[55](#_ENREF_55" \o "Bauer, 2008 #135) | GO enrichment tool |
| DAVID[56](#_ENREF_56) | Interface for enrichment of identification results |
| 3D structures | jmol[57](#_ENREF_57" \o "Hanson, 2010 #138) | Tool for the display of 3D structures |
| Pathways | Reactome[9](#_ENREF_9" \o "Haw, 2011 #139) | Pathway investigation interface allowing the mapping  of one’s results and pathway coverage estimation |
| Interactions | STRING[58](#_ENREF_58) | Protein interaction investigation interface |
| Repository | PRIDE[12](#_ENREF_12) | Protein identification repository |
| PeptideAtlas[59](#_ENREF_59" \o "Deutsch, 2008 #142) | Peptide identification repository |
| GPMDB[60](#_ENREF_60) | Peptide and protein identification repository |
| Quality control | SimpatiQCo[61](#_ENREF_61" \o "Pichler, 2012 #1) | Quality Control for proteomics |
| Local data management | MASPECTRAS[62](#_ENREF_62) | LIMS system |
| Proteios[63](#_ENREF_63" \o "Hakkinen, 2009 #160) | LIMS system |
| ms\_lims[64](#_ENREF_64) | LIMS system |

References

1. Kessner, D., Chambers, M., Burke, R., Agus, D. & Mallick, P. ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* **24**, 2534-2536 (2008).

2. Sturm, M. et al. OpenMS - an open-source software framework for mass spectrometry. *BMC bioinformatics* **9**, 163 (2008).

3. Junker, J. et al. TOPPAS: a graphical workflow editor for the analysis of high-throughput proteomics data. *Journal of proteome research* **11**, 3914-3920 (2012).

4. Apweiler, R. et al. UniProt: the Universal Protein knowledgebase. *Nucleic acids research* **32**, D115-119 (2004).

5. Martens, L., Vandekerckhove, J. & Gevaert, K. DBToolkit: processing protein databases for peptide-centric proteomics. *Bioinformatics* **21**, 3584-3585 (2005).

6. Geer, L.Y. et al. Open mass spectrometry search algorithm. *J Proteome Res* **3**, 958-964 (2004).

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

8. 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* **11**, 996-999 (2011).

9. Haw, R., Hermjakob, H., D'Eustachio, P. & Stein, L. Reactome pathway analysis to enrich biological discovery in proteomics data sets. *Proteomics* **11**, 3598-3613 (2011).

10. Cote, R.G. et al. The Protein Identifier Cross-Referencing (PICR) service: reconciling protein identifiers across multiple source databases. *BMC Bioinformatics* **8**, 401 (2007).

11. Jones, P. et al. Dasty and UniProt DAS: a perfect pair for protein feature visualization. *Bioinformatics* **21**, 3198-3199 (2005).

12. Martens, L. et al. PRIDE: the proteomics identifications database. *Proteomics* **5**, 3537-3545 (2005).

13. Cote, R.G., Reisinger, F. & Martens, L. jmzML, an open-source Java API for mzML, the PSI standard for MS data. *Proteomics* **10**, 1332-1335 (2010).

14. Bertsch, A., Gropl, C., Reinert, K. & Kohlbacher, O. OpenMS and TOPP: open source software for LC-MS data analysis. *Methods Mol Biol* **696**, 353-367 (2011).

15. Deutsch, E.W. et al. A guided tour of the Trans-Proteomic Pipeline. *Proteomics* **10**, 1150-1159 (2010).

16. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* **26**, 1367-1372 (2008).

17. Slotta, D.J., McFarland, M.A. & Markey, S.P. MassSieve: panning MS/MS peptide data for proteins. *Proteomics* **10**, 3035-3039 (2010).

18. Dancik, V., Addona, T.A., Clauser, K.R., Vath, J.E. & Pevzner, P.A. De novo peptide sequencing via tandem mass spectrometry. *Journal of computational biology : a journal of computational molecular cell biology* **6**, 327-342 (1999).

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

20. Jeong, K., Kim, S. & Pevzner, P.A. UniNovo: a universal tool for de novo peptide sequencing. *Bioinformatics* **29**, 1953-1962 (2013).

21. Andreotti, S., Klau, G.W. & Reinert, K. Antilope--a Lagrangian relaxation approach to the de novo peptide sequencing problem. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM* **9**, 385-394 (2012).

22. Fischer, B. et al. NovoHMM: a hidden Markov model for de novo peptide sequencing. *Analytical chemistry* **77**, 7265-7273 (2005).

23. Taylor, J.A. & Johnson, R.S. Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. *Rapid communications in mass spectrometry : RCM* **11**, 1067-1075 (1997).

24. Bern, M. & Goldberg, D. De novo analysis of peptide tandem mass spectra by spectral graph partitioning. *Journal of computational biology : a journal of computational molecular cell biology* **13**, 364-378 (2006).

25. DiMaggio, P.A., Jr. & Floudas, C.A. De novo peptide identification via tandem mass spectrometry and integer linear optimization. *Analytical chemistry* **79**, 1433-1446 (2007).

26. Ma, B. et al. PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom* **17**, 2337-2342 (2003).

27. Tabb, D.L., Saraf, A. & Yates, J.R., 3rd GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. *Anal Chem* **75**, 6415-6421 (2003).

28. Tabb, D.L., Ma, Z.Q., Martin, D.B., Ham, A.J. & Chambers, M.C. DirecTag: accurate sequence tags from peptide MS/MS through statistical scoring. *J Proteome Res* **7**, 3838-3846 (2008).

29. Dasari, S. et al. TagRecon: high-throughput mutation identification through sequence tagging. *Journal of proteome research* **9**, 1716-1726 (2010).

30. Yates, J.R., 3rd, Eng, J.K., McCormack, A.L. & Schieltz, D. Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal Chem* **67**, 1426-1436 (1995).

31. Perkins, D.N., Pappin, D.J., Creasy, D.M. & Cottrell, J.S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551-3567 (1999).

32. Wenger, C.D. & Coon, J.J. A proteomics search algorithm specifically designed for high-resolution tandem mass spectra. *Journal of proteome research* **12**, 1377-1386 (2013).

33. Tanner, S. et al. InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Anal Chem* **77**, 4626-4639 (2005).

34. Tabb, D.L., Fernando, C.G. & Chambers, M.C. MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *J Proteome Res* **6**, 654-661 (2007).

35. Yadav, A.K., Kumar, D. & Dash, D. MassWiz: a novel scoring algorithm with target-decoy based analysis pipeline for tandem mass spectrometry. *J Proteome Res* **10**, 2154-2160 (2011).

36. Cox, J. et al. Andromeda: a peptide search engine integrated into the MaxQuant environment. *J Proteome Res* **10**, 1794-1805 (2011).

37. Wang, R. et al. PRIDE Inspector: a tool to visualize and validate MS proteomics data. *Nat Biotechnol* **30**, 135-137 (2012).

38. Junker, J. et al. TOPPAS: A Graphical Workflow Editor for the Analysis of High-Throughput Proteomics Data. *J Proteome Res* (2012).

39. Muth, T. et al. DeNovoGUI: an open source graphical user interface for de novo sequencing of tandem mass spectra. *Journal of proteome research* (2013).

40. Craig, R., Cortens, J.C., Fenyo, D. & Beavis, R.C. Using annotated peptide mass spectrum libraries for protein identification. *J Proteome Res* **5**, 1843-1849 (2006).

41. Stein, S.E. & Scott, D.R. Optimization and testing of mass spectral library search algorithms for compound identification. *Journal of the American Society for Mass Spectrometry* **5**, 859-866 (1994).

42. Dasari, S. et al. Pepitome: evaluating improved spectral library search for identification complementarity and quality assessment. *Journal of proteome research* **11**, 1686-1695 (2012).

43. Lam, H. et al. Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* **7**, 655-667 (2007).

44. Helsens, K., Martens, L., Vandekerckhove, J. & Gevaert, K. MascotDatfile: an open-source library to fully parse and analyse MASCOT MS/MS search results. *Proteomics* **7**, 364-366 (2007).

45. Barsnes, H., Huber, S., Sickmann, A., Eidhammer, I. & Martens, L. OMSSA Parser: an open-source library to parse and extract data from OMSSA MS/MS search results. *Proteomics* **9**, 3772-3774 (2009).

46. Muth, T., Vaudel, M., Barsnes, H., Martens, L. & Sickmann, A. XTandem Parser: an open-source library to parse and analyse X!Tandem MS/MS search results. *Proteomics* **10**, 1522-1524 (2010).

47. Barsnes, H. et al. compomics-utilities: an open-source Java library for computational proteomics. *BMC Bioinformatics* **12**, 70 (2011).

48. Kall, L., Canterbury, J.D., Weston, J., Noble, W.S. & MacCoss, M.J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat Methods* **4**, 923-925 (2007).

49. Keller, A., Nesvizhskii, A.I., Kolker, E. & Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem* **74**, 5383-5392 (2002).

50. Edwards, N., Wu, X. & Tseng, C.-W. An Unsupervised, Model-Free, Machine-Learning Combiner for Peptide Identifications from Tandem Mass Spectra. *Clinical Proteomics* **5**, 23-36 (2009).

51. Shteynberg, D. et al. iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. *Mol Cell Proteomics* **10**, M111 007690 (2011).

52. Nesvizhskii, A.I., Keller, A., Kolker, E. & Aebersold, R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem* **75**, 4646-4658 (2003).

53. Ma, Z.Q. et al. IDPicker 2.0: Improved protein assembly with high discrimination peptide identification filtering. *J Proteome Res* **8**, 3872-3881 (2009).

54. Zhang, B., Schmoyer, D., Kirov, S. & Snoddy, J. GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. *BMC Bioinformatics* **5**, 16 (2004).

55. Bauer, S., Grossmann, S., Vingron, M. & Robinson, P.N. Ontologizer 2.0--a multifunctional tool for GO term enrichment analysis and data exploration. *Bioinformatics* **24**, 1650-1651 (2008).

56. Huang da, W., Sherman, B.T. & Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44-57 (2009).

57. Hanson, R. Jmol - a paradigm shift in crystallographic visualization. *Journal of Applied Crystallography* **43**, 1250-1260 (2010).

58. Szklarczyk, D. et al. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res* **39**, D561-568 (2011).

59. Deutsch, E.W., Lam, H. & Aebersold, R. PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. *EMBO Rep* **9**, 429-434 (2008).

60. Craig, R., Cortens, J.P. & Beavis, R.C. Open source system for analyzing, validating, and storing protein identification data. *J Proteome Res* **3**, 1234-1242 (2004).

61. Pichler, P. et al. SIMPATIQCO: a server-based software suite which facilitates monitoring the time course of LC-MS performance metrics on Orbitrap instruments. *Journal of proteome research* **11**, 5540-5547 (2012).

62. Hartler, J. et al. MASPECTRAS: a platform for management and analysis of proteomics LC-MS/MS data. *BMC Bioinformatics* **8**, 197 (2007).

63. Hakkinen, J., Vincic, G., Mansson, O., Warell, K. & Levander, F. The proteios software environment: an extensible multiuser platform for management and analysis of proteomics data. *J Proteome Res* **8**, 3037-3043 (2009).

64. Helsens, K. et al. ms\_lims, a simple yet powerful open source laboratory information management system for MS-driven proteomics. *Proteomics* **10**, 1261-1264 (2010).