References

Adams, M., *et al.* (1991), Complementary dna sequencing: expressed sequence tags and human genome project., *Science*, *252*(5013), 1651–1656. (page 24)

Aebersold, R., and M. Mann (2003), Mass spectrometry-based proteomics., *Nature*, *422*(6928), 198–207. (page 2, 3)

Alexandersson, M., S. Cawley, and L. Pachter (2003), Slam: cross-species gene finding and alignment with a generalized pair hidden markov model., *Genome Res*, *13*(3), 496–502. (page 25)

Altschul, S., W. Gish, W. Miller, E. Myers, and D. Lipman (1990), Basic local alignment search tool., *J Mol Biol*, *215*(3), 403–410. (page 30)

Anderson, D., W. Li, D. Payan, and W. Noble (2003), A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: support vector machine classification of peptide ms/ms spectra and sequest scores., *J Proteome Res*, *2*(2), 137–146. (page 34)

Anderson, N., N. Anderson, T. Pearson, C. Borchers, A. Paulovich, S. Patterson, M. Gillette, R. Aebersold, and S. Carr (2009), A human proteome detection and quantitation project., *Mol Cell Proteomics*, *8*(5), 883–886. (page 115)

Arnott, D., A. Kishiyama, E. Luis, S. Ludlum, J. J. Marsters, and J. Stults (2002), Selective detection of membrane proteins without antibodies: a mass spectrometric version of the western blot., *Mol Cell Proteomics*, *1*(2), 148–156. (page 115)

Ashurst, J., *et al.* (2005), The vertebrate genome annotation (vega) database., *Nucleic Acids Res*, *33*(Database issue), D459–65. (page 28, 84)

Bairoch, A., and R. Apweiler (1997), The swiss-prot protein sequence data bank and its supplement trembl, *Nucleic Acids Research*, *25*(1), 31. (page 28)

Beausoleil, S., J. Villen, S. Gerber, J. Rush, and S. Gygi (2006), A probability-based approach for high-throughput protein phosphorylation analysis and site localization., *Nat Biotechnol*, *24*(10), 1285–1292. (page 37)

Ben-Hur, A., C. Ong, S. Sonnenburg, B. Scholkopf, and G. Ratsch (2008), Support vector machines and kernels for computational biology., *PLoS Comput Biol*, *4*(10), e1000,173. (page 19)

Benjamini, Y., and Y. Hochberg (1995), Controlling the false discovery rate: a practical and powerful approach to multiple testing, *J. R. Stat. Soc., Ser B*, *57*(1), 289–300. (page 10)

Bern, M., Y. Cai, and D. Goldberg (2007), Lookup peaks: A hybrid of de novo sequencing and database search for protein identification by tandem mass spectrometry., *Anal Chem*. (page 5)

Bianco, L., J. Mead, and C. Bessant (2009), Comparison of novel decoy database designs for optimizing protein identification searches using abrf sprg2006 standard ms/ms datasets., *J Proteome Res*. (page 16)

Biemann, K. (1988), Contributions of mass spectrometry to peptide and protein structure., *Biomed Environ Mass Spectrom*, *16*(1-12), 99–111. (page 2, 37, 105)

Birney, E., and R. Durbin (1997), Dynamite: a flexible code generating language for dynamic programming methods used in sequence comparison., *Proc Int Conf Intell Syst Mol Biol*, *5*, 56–64. (page 24)

Birney, E., M. Clamp, and R. Durbin (2004), Genewise and genomewise., *Genome Res*, *14*(5), 988–995. (page 24)

Birney, E., *et al.* (2007), Identification and analysis of functional elements in 1% of the human genome by the encode pilot project., *Nature*, *447*(7146), 799–816. (page 23, 26, 29)

Bonferroni, C. (1935), Il calcolo delle assicurazioni su gruppi di teste, *Studi in Onore del Professore Salvatore Ortu Carboni*, *13*. (page 10)

Brent, M. (2005), Genome annotation past, present, and future: how to define an orf at each locus., *Genome Res*, *15*(12), 1777–1786. (page 23)

Brogna, S., and J. Wen (2009), Nonsense-mediated mrna decay (nmd) mechanisms., *Nat Struct Mol Biol*, *16*(2), 107–113. (page 117)

Brosch, M., S. Swamy, T. Hubbard, and J. Choudhary (2008), Comparison of mascot and x!tandem performance for low and high accuracy mass spectrometry and the development of an adjusted mascot threshold., *Mol Cell Proteomics*, *7*(5), 962–970. (page 39, 65, 76, 128)

Brosch, M., L. Yu, T. Hubbard, and J. Choudhary (2009), Accurate and sensitive peptide identification with mascot percolator., *J Proteome Res*, *8*(6), 3176–3181. (page 64, 128)

Brunner, E., *et al.* (2007), A high-quality catalog of the drosophila melanogaster proteome., *Nat Biotechnol*, *25*(5), 576–583. (page 31)

Bunger, M., B. Cargile, J. Sevinsky, E. Deyanova, N. Yates, R. Hendrickson, and J. J. Stephenson (2007), Detection and validation of non-synonymous coding snps from orthogonal analysis of shotgun proteomics data., *J Proteome Res*, *6*(6), 2331–2340. (page 31)

Burge, C., and S. Karlin (1997), Prediction of complete gene structures in human genomic dna., *J Mol Biol*, *268*(1), 78–94. (page 25)

Carninci, P., *et al.* (2005), The transcriptional landscape of the mammalian genome., *Science*, *309*(5740), 1559–1563. (page 21)

Castellana, N., S. Payne, Z. Shen, M. Stanke, V. Bafna, and S. Briggs (2008), Discovery and revision of arabidopsis genes by proteogenomics., *Proc Natl Acad Sci U S A*, *105*(52), 21,034–21,038. (page 31, 115, 121)

Chepanoske, C., B. Richardson, M. von Rechenberg, and J. Peltier (2005), Average peptide score: a useful parameter for identification of proteins derived from database searches of liquid chromatography/tandem mass spectrometry data., *Rapid Commun Mass Spectrom*, *19*(1), 9–14. (page 30)

Chernushevich, I., A. Loboda, and B. Thomson (2001), An introduction to quadrupole-time-of-flight mass spectrometry., *J Mass Spectrom*, *36*(8), 849–865. (page 4)

Choi, H., and A. Nesvizhskii (2008), Semisupervised model-based validation of peptide identifications in mass spectrometry-based proteomics., *J Proteome Res*, *7*(1), 254–265. (page 18)

Choi, H., D. Ghosh, and A. Nesvizhskii (2008), Statistical validation of peptide identifications in large-scale proteomics using the target-decoy database search strategy and flexible mixture modeling., *J Proteome Res*, *7*(1), 286–292. (page 18)

Choudhary, J., W. Blackstock, D. Creasy, and J. Cottrell (2001a), Matching peptide mass spectra to est and genomic dna databases., *Trends Biotechnol*, *19*(10 Suppl), S17–22. (page 29)

Choudhary, J., W. Blackstock, D. Creasy, and J. Cottrell (2001b), Interrogating the

human genome using uninterpreted mass spectrometry data., *Proteomics*, *1*(5), 651–667. (page 29)

Clamp, M., B. Fry, M. Kamal, X. Xie, J. Cuff, M. Lin, M. Kellis, K. Lindblad-Toh, and E. Lander (2007), Distinguishing protein-coding and noncoding genes in the human genome., *Proc Natl Acad Sci U S A*. (page 21, 26)

Clauser, K., P. Baker, and A. Burlingame (1999), Role of accurate mass measurement (+/- 10 ppm) in protein identification strategies employing ms or ms/ms and database searching., *Anal Chem*, *71*(14), 2871–2882. (page 4)

Claverie, J. (2005), Fewer genes, more noncoding rna., *Science*, *309*(5740), 1529–1530. (page 21, 23, 25, 29)

Colinge, J., A. Masselot, M. Giron, T. Dessingy, and J. Magnin (2003), Olav: towards high-throughput tandem mass spectrometry data identification., *Proteomics*, *3*(8), 1454–1463. (page 14)

Craig, R., and R. Beavis (2004), Tandem: matching proteins with tandem mass spectra., *Bioinformatics*, *20*(9), 1466–1467. (page 7, 34, 36)

Crick, F. (1958), On protein synthesis., *Symp Soc Exp Biol*, *12*, 138–163. (page 22)

Crick, F. (1970), Central dogma of molecular biology., *Nature*, *227*(5258), 561–563. (page 22)

Curwen, V., E. Eyras, T. Andrews, L. Clarke, E. Mongin, S. Searle, and M. Clamp (2004), The ensembl automatic gene annotation system., *Genome Res*, *14*(5), 942–950. (page 25)

Dancik, V., T. Addona, K. Clauser, J. Vath, and P. Pevzner (1999), De novo peptide sequencing via tandem mass spectrometry., *J Comput Biol*, *6*(3-4), 327–342. (page 5)

de Godoy, L., J. Olsen, G. de Souza, G. Li, P. Mortensen, and M. Mann (2006), Status of complete proteome analysis by mass spectrometry: Silac labeled yeast as a model system., *Genome Biol*, *7*(6), R50. (page 2, 111)

Dempster, A., N. Laird, and D. Rubin (1977), Maximum likelihood from incomplete data via the em algorithm, *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*(1), 1–38. (page 18)

Desiere, F., *et al.* (2005), Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry., *Genome Biol*, *6*(1), R9. (page 29, 83)

Desiere, F., *et al.* (2006), The peptideatlas project., *Nucleic Acids Res*, *34*(Database issue), D655–8. (page 4, 29, 98)

Douglas, D., A. Frank, and D. Mao (2005), Linear ion traps in mass spectrometry., *Mass Spectrom Rev*, *24*(1), 1–29. (page 4)

Dowell, R., R. Jokerst, A. Day, S. Eddy, and L. Stein (2001), The distributed annotation system., *BMC Bioinformatics*, *2*, 7. (page 30, 96, 104)

Elias, J., and S. Gygi (2007), Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry., *Nat Methods*, *4*(3), 207–214. (page 16, 37, 62)

Elias, J., W. Haas, B. Faherty, and S. Gygi (2005), Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations., *Nat Methods*, *2*(9), 667–675. (page 4)

Eng, J., A. McCormack, and J. Yates (1994), An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database, *J. Am. Soc. Mass Spectrom*, *5*(11), 976–989. (page 6, 7)

Everley, P., C. Bakalarski, J. Elias, C. Waghorne, S. Beausoleil, S. Gerber, B. Faherty, B. Zetter, and S. Gygi (2006), Enhanced analysis of metastatic prostate cancer using stable isotopes and high mass accuracy instrumentation., *J Proteome Res*, *5*(5), 1224–1231. (page 37)

Fenyo, D., and R. Beavis (2003), A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes., *Anal Chem*, *75*(4), 768–774. (page 35, 36, 50)

Finn, R., J. Stalker, D. Jackson, E. Kulesha, J. Clements, and R. Pettett (2007), Proserver: a simple, extensible perl das server., *Bioinformatics*, *23*(12), 1568–1570. (page 96)

Fitzgibbon, M., Q. Li, and M. McIntosh (2007), Modes of inference for evaluating the confidence of peptide identifications, *Journal of Proteome Research Journal of Proteome Research J. Proteome Res.* (page 16)

Foster, L., C. de Hoog, Y. Zhang, Y. Zhang, X. Xie, V. Mootha, and M. Mann (2006), A mammalian organelle map by protein correlation profiling., *Cell*, *125*(1), 187–199. (page 2)

Frank, A., and P. Pevzner (2005), Pepnovo: de novo peptide sequencing via probabilistic network modeling., *Anal Chem*, *77*(4), 964–973. (page 5)

Frank, A., M. Savitski, M. Nielsen, R. Zubarev, and P. Pevzner (2007), De novo peptide sequencing and identification with precision mass spectrometry., *J Proteome Res*, *6*(1), 114–123. (page 5)

Frottin, F., A. Martinez, P. Peynot, S. Mitra, R. Holz, C. Giglione, and T. Meinnel (2006), The proteomics of n-terminal methionine cleavage., *Mol Cell Proteomics*, *5*(12), 2336–2349. (page 87, 100)

Fullwood, M., C. Wei, E. Liu, and Y. Ruan (2009), Next-generation dna sequencing of paired-end tags (pet) for transcriptome and genome analyses., *Genome Res*, *19*(4), 521–532. (page 28)

Furuno, M., T. Kasukawa, R. Saito, J. Adachi, H. Suzuki, R. Baldarelli, Y. Hayashizaki, and Y. Okazaki (2003), Cds annotation in full-length cdna sequence., *Genome Res*, *13*(6B), 1478–1487. (page 24)

Fusaro, V., D. Mani, J. Mesirov, and S. Carr (2009), Prediction of high-responding peptides for targeted protein assays by mass spectrometry., *Nat Biotechnol*, *27*(2), 190–198. (page 109)

Guigo, R., and M. Reese (2005), Egasp: collaboration through competition to find human genes., *Nat Methods*, *2*(8), 575–577. (page 102)

Guigo, R., *et al.* (2006), Egasp: the human encode genome annotation assessment project., *Genome Biol*, *7 Suppl 1*, S2.1–31. (page 26)

Gupta, N., *et al.* (2008), Comparative proteogenomics: combining mass spectrometry and comparative genomics to analyze multiple genomes., *Genome Res*, *18*(7), 1133–1142. (page 32)

Haas, W., B. Faherty, S. Gerber, J. Elias, S. Beausoleil, C. Bakalarski, X. Li, J. Villen, and S. Gygi (2006), Optimization and use of peptide mass measurement accuracy in shotgun proteomics., *Mol Cell Proteomics*, *5*(7), 1326–1337. (page 4)

Han, X., M. Jin, K. Breuker, and F. McLafferty (2006), Extending top-down mass spectrometry to proteins with masses greater than 200 kilodaltons., *Science*, *314*(5796), 109–112. (page 3)

Harrow, J., A. Nagy, A. Reymond, T. Alioto, L. Patthy, S. Antonarakis, and R. Guigo (2009), Identifying protein-coding genes in genomic sequences., *Genome Biol*, *10*(1), 201. (page 24)

Harrow, J., *et al.* (2006), Gencode: producing a reference annotation for encode., *Genome Biol*, *7 Suppl 1*, S4.1–9. (page 26, 28)

Heeren, R., A. Kleinnijenhuis, L. McDonnell, and T. Mize (2004), A mini-review of mass spectrometry using high-performance fticr-ms methods., *Anal Bioanal Chem*, *378*(4), 1048–1058. (page 4)

Hsueh, H., J. Chen, and R. Kodell (2003), Comparison of methods for estimating the number of true null hypotheses in multiplicity testing., *Journal of biopharmaceutical statistics*, *13*(4), 675. (page 16)

Hu, Q., R. Noll, H. Li, A. Makarov, M. Hardman, and R. Graham Cooks (2005), The orbitrap: a new mass spectrometer., *J Mass Spectrom*, *40*(4), 430–443. (page 4)

Hubbard, T., *et al.* (2002), The ensembl genome database project., *Nucleic Acids Res*, *30*(1), 38–41. (page 25, 84)

Hubbard, T., *et al.* (2009), Ensembl 2009., *Nucleic Acids Res*, *37*(Database issue), D690–7. (page 25, 28)

Hunt, D., R. Henderson, J. Shabanowitz, K. Sakaguchi, H. Michel, N. Sevilir, A. Cox, E. Appella, and V. Engelhard (1992), Characterization of peptides bound to the class i mhc molecule hla-a2.1 by mass spectrometry., *Science*, *255*(5049), 1261–1263. (page 2)

Imanishi, T., *et al.* (2004), Integrative annotation of 21,037 human genes validated by full-length cdna clones., *PLoS Biol*, *2*(6), e162. (page 24)

International Human Genome Sequencing Consortium (2004), Finishing the euchromatic sequence of the human genome., *Nature*, *431*(7011), 931–945. (page 26)

Ishihama, Y., Y. Oda, T. Tabata, T. Sato, T. Nagasu, J. Rappsilber, and M. Mann (2005), Exponentially modified protein abundance index (empai) for estimation of

absolute protein amount in proteomics by the number of sequenced peptides per protein., *Mol Cell Proteomics*, *4*(9), 1265–1272. (page 107)

Jaffe, J., H. Berg, and G. Church (2004), Proteogenomic mapping as a complementary method to perform genome annotation., *Proteomics*, *4*(1), 59–77. (page 29)

Jin, J., and T. Cai (2006), Estimating the null and the proportion of non-null effects in large-scale multiple comparisons, *Arxiv preprint math.ST/0611108*. (page 16)

Jones, A., J. Siepen, S. Hubbard, and N. Paton (2009), Improving sensitivity in proteome studies by analysis of false discovery rates for multiple search engines., *Proteomics*, *9*(5), 1220–1229. (page 34)

Jones, P., R. Cote, S. Cho, S. Klie, L. Martens, A. Quinn, D. Thorneycroft, and H. Hermjakob (2008), Pride: new developments and new datasets., *Nucleic Acids Res*, *36*(Database issue), D878–83. (page 4)

Käll, L., J. Canterbury, J. Weston, W. Noble, and M. MacCoss (2007), Semi-supervised learning for peptide identification from shotgun proteomics datasets., *Nat Methods*, *4*(11), 923–925. (page 18, 19, 20, 63, 64, 65, 67, 73, 74, 128)

Käll, L., J. Storey, M. MacCoss, and W. Noble (2008a), Assigning significance to peptides identified by tandem mass spectrometry using decoy databases., *J Proteome Res*, *7*(1), 29–34. (page 11, 16)

Käll, L., J. Storey, M. MacCoss, and W. Noble (2008b), Posterior error probabilities and false discovery rates: two sides of the same coin., *J Proteome Res*, *7*(1), 40–44. (page 11, 13, 20)

Käll, L., J. Storey, and W. Noble (2008c), Non-parametric estimation of posterior error probabilities associated with peptides identified by tandem mass spectrometry., *Bioinformatics*, *24*(16), i42–8. (page 13, 16, 20)

Kapp, E., *et al.* (2005), An evaluation, comparison, and accurate benchmarking of several publicly available ms/ms search algorithms: sensitivity and specificity analysis., *Proteomics*, *5*(13), 3475–3490. (page 6)

Keller, A., A. Nesvizhskii, E. Kolker, and R. Aebersold (2002), Empirical statistical model to estimate the accuracy of peptide identifications made by ms/ms and database search., *Anal Chem*, *74*(20), 5383–5392. (page 13, 17)

Kersey, P., J. Duarte, A. Williams, Y. Karavidopoulou, E. Birney, and R. Apweiler (2004), The international protein index: an integrated database for proteomics experiments., *Proteomics*, *4*(7), 1985–1988. (page 29, 98)

Kersey, P., *et al.* (2009), Ensembl genomes: Extending ensembl across the taxonomic space., *Nucleic Acids Res.* (page 26)

Kim, S., N. Gupta, N. Bandeira, and P. Pevzner (2009), Spectral dictionaries: Integrating de novo peptide sequencing with database search of tandem mass spectra., *Mol Cell Proteomics*, *8*(1), 53–69. (page 5)

Klammer, A., and M. MacCoss (2006), Effects of modified digestion schemes on the identification of proteins from complex mixtures., *J Proteome Res*, *5*(3), 695–700. (page 14)

Klimek, J., *et al.* (2007), The standard protein mix database: A diverse data set to assist in the production of improved peptide and protein identification software tools., *J Proteome Res.* (page 43)

Knowles, D., and A. McLysaght (2009), Recent de novo origin of human protein-coding genes., *Genome Res*, *19*(10), 1752–1759. (page 25)

Korf, I., P. Flicek, D. Duan, and M. Brent (2001), Integrating genomic homology into gene structure prediction., *Bioinformatics*, *17 Suppl 1*, S140–8. (page 25)

Kuhn, R., *et al.* (2009), The ucsc genome browser database: update 2009, *Nucleic acids research*, *37*(Database issue), D755. (page 28)

Kuster, B., P. Mortensen, J. Andersen, and M. Mann (2001), Mass spectrometry allows direct identification of proteins in large genomes., *Proteomics*, *1*(5), 641–650. (page 29)

Lander, E., *et al.* (2001), Initial sequencing and analysis of the human genome., *Nature*, *409*(6822), 860–921. (page 25, 26)

Liolios, K., I. Chen, K. Mavromatis, N. Tavernarakis, P. Hugenholtz, V. Markowitz, and N. Kyrpides (2009), The genomes on line database (gold) in 2009: status of genomic and metagenomic projects and their associated metadata., *Nucleic Acids Res.* (page 22)

Lu, P., C. Vogel, R. Wang, X. Yao, and E. Marcotte (2007), Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation., *Nat Biotechnol*, *25*(1), 117–124. (page 107)

Mallick, P., *et al.* (2007), Computational prediction of proteotypic peptides for quantitative proteomics., *Nat Biotechnol*, *25*(1), 125–131. (page 109)

Mann, M., and M. Wilm (1994), Error-tolerant identification of peptides in sequence databases by peptide sequence tags., *Anal Chem*, *66*(24), 4390–4399. (page 5, 31)

Maquat, L. (2005), Nonsense-mediated mrna decay in mammals, *Journal of cell science*, *118(9)*(9), 1773–1776. (page 99)

Martens, L., H. Hermjakob, P. Jones, M. Adamski, C. Taylor, D. States, K. Gevaert, J. Vandekerckhove, and R. Apweiler (2005a), Pride: the proteomics identifications database., *Proteomics*, *5*(13), 3537–3545. (page 4)

Martens, L., J. Vandekerckhove, and K. Gevaert (2005b), Dbtoolkit: processing protein databases for peptide-centric proteomics., *Bioinformatics*, *21*(17), 3584–3585. (page 91)

McCormack, A., D. Schieltz, B. Goode, S. Yang, G. Barnes, D. Drubin, and J. r. Yates (1997), Direct analysis and identification of proteins in mixtures by lc/ms/ms and database searching at the low-femtomole level., *Anal Chem*, *69*(4), 767–776. (page 2)

McGlincy, N., and C. Smith (2008), Alternative splicing resulting in nonsense-mediated mrna decay: what is the meaning of nonsense?, *Trends Biochem Sci*, *33*(8), 385–393. (page 117)

Meinshausen, N., and J. Rice (2006), Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses, *Annals of statistics*, *34*(1), 373. (page 16)

Merrihew, G., *et al.* (2008), Use of shotgun proteomics for the identification, confirmation, and correction of c. elegans gene annotations., *Genome Res*, *18*(10), 1660–1669. (page 121)

Moore, R., M. Young, and T. Lee (2002), Qscore: an algorithm for evaluating sequest database search results., *J Am Soc Mass Spectrom*, *13*(4), 378–386. (page 14)

Nagaraj, S., R. Gasser, and S. Ranganathan (2007), A hitchhiker's guide to expressed sequence tag (est) analysis., *Brief Bioinform*, *8*(1), 6–21. (page 24)

Nesvizhskii, A., and R. Aebersold (2004), Analysis, statistical validation and dissemination of large-scale proteomics datasets generated by tandem ms., *Drug Discov Today*, *9*(4), 173–181. (page 2, 30, 45)

Nesvizhskii, A., A. Keller, E. Kolker, and R. Aebersold (2003), A statistical model for

identifying proteins by tandem mass spectrometry., *Anal Chem*, *75*(17), 4646–4658. (page 2, 30)

Nesvizhskii, A. I., O. Vitek, and R. Aebersold (2007), Analysis and validation of proteomic data generated by tandem mass spectrometry, *Nat Meth*, *4*(10), 787–797. (page 4)

Olsen, J., S. Ong, and M. Mann (2004), Trypsin cleaves exclusively c-terminal to arginine and lysine residues., *Mol Cell Proteomics*, *3*(6), 608–614. (page 5)

Parkinson, J., and M. Blaxter (2009), Expressed sequence tags: an overview., *Methods Mol Biol*, *533*, 1–12. (page 24)

Parks, B., L. Jiang, P. Thomas, C. Wenger, M. Roth, M. n. Boyne, P. Burke, K. Kwast, and N. Kelleher (2007), Top-down proteomics on a chromatographic time scale using linear ion trap fourier transform hybrid mass spectrometers., *Anal Chem*, *79*(21), 7984–7991. (page 3)

Parra, G., E. Blanco, and R. Guigo (2000), Geneid in drosophila., *Genome Res*, *10*(4), 511–515. (page 25)

Parra, G., P. Agarwal, J. Abril, T. Wiehe, J. Fickett, and R. Guigo (2003), Comparative gene prediction in human and mouse., *Genome Res*, *13*(1), 108–117. (page 25)

Pasa-Tolic, L., C. Masselon, R. Barry, Y. Shen, and R. Smith (2004), Proteomic analyses using an accurate mass and time tag strategy., *Biotechniques*, *37*(4), 621–4, 626–33, 636 passim. (page 39)

Patterson, S., and R. Aebersold (2003), Proteomics: the first decade and beyond., *Nat Genet*, *33 Suppl*, 311–323. (page 2)

Peng, J., J. Elias, C. Thoreen, L. Licklider, and S. Gygi (2003), Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (lc/lc-

ms/ms) for large-scale protein analysis: the yeast proteome., *J Proteome Res*, *2*(1), 43–50. (page 4)

Pennisi, E. (2009), Dna sequencing. no genome left behind., *Science*, *326*(5954), 794–795. (page 22)

Perkins, D., D. Pappin, D. Creasy, and J. Cottrell (1999), Probability-based protein identification by searching sequence databases using mass spectrometry data., *Electrophoresis*, *20*(18), 3551–3567. (page 6, 7, 34)

Pitzer, E., A. Masselot, and J. Colinge (2007), Assessing peptide de novo sequencing algorithms performance on large and diverse data sets., *Proteomics*. (page 5)

Pruitt, K., T. Tatusova, and D. Maglott (2006), Ncbi reference sequences (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins, *Nucleic acids research*. (page 28)

Pruitt, K., *et al.* (2009), The consensus coding sequence (ccds) project: Identifying a common protein-coding gene set for the human and mouse genomes., *Genome Res*, *19*(7), 1316–1323. (page 28)

Resing, K., *et al.* (2004), Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics., *Anal Chem*, *76*(13), 3556–3568. (page 34)

Rodriguez, J., N. Gupta, R. D. Smith, and P. A. Pevzner (2007), Does trypsin cut before proline?, *Journal of Proteome Research Journal of Proteome Research J. Proteome Res.* (page 5, 65)

Roepstorff, P., and J. Fohlman (1984), Proposal for a common nomenclature for sequence ions in mass spectra of peptides., *Biomed Mass Spectrom*, *11(11)*(11), 601. (page 2, 37, 105)

Roth, M., B. Parks, J. Ferguson, M. n. Boyne, and N. Kelleher (2008), "proteotyping": population proteomics of human leukocytes using top down mass spectrometry., *Anal Chem*, *80*(8), 2857–2866. (page 3)

Rudnick, P., Y. Wang, E. Evans, C. Lee, and B. Balgley (2005), Large scale analysis of mascot results using a mass accuracy-based threshold (math) effectively improves data interpretation., *J Proteome Res*, *4*(4), 1353–1360. (page 37, 53)

Savitski, M., M. Nielsen, and R. Zubarev (2005), New data base-independent, sequence tag-based scoring of peptide ms/ms data validates mowse scores, recovers below threshold data, singles out modified peptides, and assesses the quality of ms/ms techniques., *Mol Cell Proteomics*, *4*(8), 1180–1188. (page 37)

Schandorff, S., J. Olsen, J. Bunkenborg, B. Blagoev, Y. Zhang, J. Andersen, and M. Mann (2007), A mass spectrometry-friendly database for csnp identification., *Nat Methods*, *4*(6), 465–466. (page 31, 87, 100)

Sevinsky, J., B. Cargile, M. Bunger, F. Meng, N. Yates, R. Hendrickson, and J. J. Stephenson (2008), Whole genome searching with shotgun proteomic data: applications for genome annotation., *J Proteome Res*, *7*(1), 80–88. (page 31)

Shadforth, I., T. Dunkley, K. Lilley, D. Crowther, and C. Bessant (2005), Confident protein identification using the average peptide score method coupled with search-specific, ab initio thresholds., *Rapid Commun Mass Spectrom*, *19*(22), 3363–3368. (page 30)

Shadforth, I., W. Xu, D. Crowther, and C. Bessant (2006), Gapp: A fully automated software for the confident identification of human peptides from tandem mass spectra, *Journal of proteome research*. (page 30, 83)

Shaffer, J. (1995), Multiple hypothesis testing, *Annual Review of Psychology*, *46*(1), 561–584. (page 10)

Shevchenko, A., M. Wilm, O. Vorm, and M. Mann (1996), Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels., *Anal Chem*, *68*(5), 850–858. (page 2)

Shin, S., and G. Sanders (2006), Denormalization strategies for data retrieval from data warehouses, *Decision Support Systems*, *42*(1), 267–282. (page 87)

Simpson, R., L. Connolly, J. Eddes, J. Pereira, R. Moritz, and G. Reid (2000), Proteomic analysis of the human colon carcinoma cell line (lim 1215): development of a membrane protein database., *Electrophoresis*, *21*(9), 1707–1732. (page 2)

Slater, G., and E. Birney (2005), Automated generation of heuristics for biological sequence comparison., *BMC Bioinformatics*, *6*, 31. (page 24)

Sonnhammer, E., S. Eddy, and R. Durbin (1997), Pfam: a comprehensive database of protein domain families based on seed alignments., *Proteins*, *28*(3), 405–420. (page 125)

Stabenau, A., G. McVicker, C. Melsopp, G. Proctor, M. Clamp, and E. Birney (2004), The ensembl core software libraries, *Genome research*, *14*(5), 929. (page 86, 87)

Stanke, M., and S. Waack (2003), Gene prediction with a hidden markov model and a new intron submodel, *Bioinformatics-Oxford*, *19*(2), 215–225. (page 25, 86)

Stanke, M., A. Tzvetkova, and B. Morgenstern (2006), Augustus at egasp: using est, protein and genomic alignments for improved gene prediction in the human genome., *Genome Biol*, *7 Suppl 1*, S11.1–8. (page 102)

States, D., G. Omenn, T. Blackwell, D. Fermin, J. Eng, D. Speicher, and S. Hanash (2006), Challenges in deriving high-confidence protein identifications from data gathered by a hupo plasma proteome collaborative study., *Nat Biotechnol*, *24*(3), 333–338. (page 8)

Stein, L. (2001), Genome annotation: from sequence to biology., *Nat Rev Genet*, *2*(7), 493–503. (page 23)

Storey, J. (2002), A direct approach to false discovery rates, *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *64*(3), 479–498. (page 16)

Storey, J., and R. Tibshirani (2003), Statistical significance for genomewide studies., *Proc Natl Acad Sci U S A*, *100*(16), 9440–9445. (page 10, 16)

Syka, J., *et al.* (2004), Novel linear quadrupole ion trap/ft mass spectrometer: performance characterization and use in the comparative analysis of histone h3 post-translational modifications., *J Proteome Res*, *3*(3), 621–626. (page 4)

Tabb, D., A. Saraf, and J. r. Yates (2003), Gutentag: high-throughput sequence tagging via an empirically derived fragmentation model., *Anal Chem*, *75*(23), 6415–6421. (page 5)

Tanner, S., H. Shu, A. Frank, L. Wang, E. Zandi, M. Mumby, P. Pevzner, and V. Bafna (2005), Inspect: identification of posttranslationally modified peptides from tandem mass spectra., *Anal Chem*, *77*(14), 4626–4639. (page 5, 31)

Tanner, S., Z. Shen, J. Ng, L. Florea, R. Guigo, S. Briggs, and V. Bafna (2007), Improving gene annotation using peptide mass spectrometry., *Genome Res.* (page 31, 115, 123)

Taylor, J., and R. Johnson (1997), Sequence database searches via de novo peptide sequencing by tandem mass spectrometry., *Rapid Commun Mass Spectrom*, *11*(9), 1067–1075. (page 5)

The ENCODE Project Consortium (2004), The encode (encyclopedia of dna elements) project, *Science*, *306*, 636–640. (page 26)

Tress, M., B. Bodenmiller, R. Aebersold, and A. Valencia (2008), Proteomics studies confirm the presence of alternative protein isoforms on a large scale., *Genome Biol*, *9*(11), R162. (page 31, 114, 115)

Ulintz, P., J. Zhu, Z. Qin, and P. Andrews (2006), Improved classification of mass spectrometry database search results using newer machine learning approaches., *Mol Cell Proteomics*, *5*(3), 497–509. (page 34)

Venter, J., *et al.* (2001), The sequence of the human genome., *Science*, *291*(5507), 1304–1351. (page 25, 26)

Wang, E., R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. Kingsmore, G. Schroth, and C. Burge (2008), Alternative isoform regulation in human tissue transcriptomes., *Nature*, *456*(7221), 470–476. (page 29, 91)

Wang, Z., M. Gerstein, and M. Snyder (2009), Rna-seq: a revolutionary tool for transcriptomics., *Nat Rev Genet*, *10*(1), 57–63. (page 28)

Washburn, M., D. Wolters, and J. r. Yates (2001), Large-scale analysis of the yeast proteome by multidimensional protein identification technology., *Nat Biotechnol*, *19*(3), 242–247. (page 2)

Washietl, S., *et al.* (2007), Structured rnas in the encode selected regions of the human genome., *Genome Res*, *17*(6), 852–864. (page 21)

Waterston, R., *et al.* (2002), Initial sequencing and comparative analysis of the mouse genome., *Nature*, *420*(6915), 520–562. (page 98, 107)

Wilming, L., J. Gilbert, K. Howe, S. Trevanion, T. Hubbard, and J. Harrow (2008), The vertebrate genome annotation (vega) database., *Nucleic Acids Res*, *36*(Database issue), D753–60. (page 28)

Wolters, D., M. Washburn, and J. r. Yates (2001), An automated multidimensional protein identification technology for shotgun proteomics., *Anal Chem*, *73*(23), 5683–5690. (page 2)

Wright, J., D. Sugden, S. Francis-McIntyre, I. Riba-Garcia, S. Gaskell, I. Grigoriev, S. Baker, R. Beynon, and S. Hubbard (2009), Exploiting proteomic data for genome annotation and gene model validation in aspergillus niger., *BMC Genomics*, *10*, 61. (page 31)

Wu, C., *et al.* (2006), The universal protein resource (uniprot): an expanding universe of protein information, *Nucleic acids research*, *34*(Database Issue), D187. (page 28)

Yates, J. r., J. Eng, and A. McCormack (1995), Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases., *Anal Chem*, *67*(18), 3202–3210. (page 29)

Zhang, T., P. Haws, and Q. Wu (2004), Multiple variable first exons: a mechanism for cell- and tissue-specific gene regulation., *Genome Res*, *14*(1), 79–89. (page 115)

Zubarev, R. (2006), Protein primary structure using orthogonal fragmentation techniques in fourier transform mass spectrometry., *Expert Rev Proteomics*, *3*(2), 251–261. (page 4)

Zubarev, R., and M. Mann (2007), On the proper use of mass accuracy in proteomics., *Mol Cell Proteomics*, *6*(3), 377–381. (page 37, 45, 62)

Zubarev, R., P. Hakansson, and B. Sundqvist (1996), Accuracy requirements for peptide characterization by monoisotopic molecular mass measurements, *Anal. Chem*, *68*(22), 4060–4063. (page 62)