



Data Article

“Controlled, cross-species dataset for exploring biases in genome annotation and modification profiles”

Alison McAfee, Sarah Michaud¹, Leonard J. Foster*

Department of Biochemistry and Molecular Biology, Centre for High-Throughput Biology and Centre for Sustainable Food Systems, University of British Columbia, 2125 East Mall, Vancouver, BC, Canada V6T 1Z4

ARTICLE INFO

Article history:

Received 8 August 2015

Received in revised form

22 October 2015

Accepted 29 October 2015

Available online 10 November 2015

Keywords:

Apis mellifera

Proteomics

Mass spectrometry

Nanoelectrospray ionization

Proteome coverage

ABSTRACT

Since the sequencing of the honey bee genome, proteomics by mass spectrometry has become increasingly popular for biological analyses of this insect; but we have observed that the number of honey bee protein identifications is consistently low compared to other organisms [1]. In this dataset, we use nanoelectrospray ionization-coupled liquid chromatography–tandem mass spectrometry (nLC–MS/MS) to systematically investigate the root cause of low honey bee proteome coverage. To this end, we present here data from three key experiments: a controlled, cross-species analyses of samples from *Apis mellifera*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, *Mus musculus* and *Homo sapiens*; a proteomic analysis of an individual honey bee whose genome was also sequenced; and a cross-tissue honey bee proteome comparison. The cross-species dataset was interrogated to determine relative proteome coverages between species, and the other two datasets were used to search for polymorphic sequences and to compare protein cleavage profiles, respectively.

© 2015 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

Specifications Table

Subject area	Biology
	Shot-gun proteomics

* Corresponding author.

E-mail addresses: alison.mcafee@alumni.ubc.ca (A. McAfee), sarah@proteincenter.com (S. Michaud), foster@chibi.ubc.ca (L.J. Foster).

¹ Present address: Genome BC Proteomics Center, University of Victoria, 4464 Markham St., Victoria, BC, Canada, V8Z 7X8I

<http://dx.doi.org/10.1016/j.dib.2015.10.042>

2352-3409/© 2015 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

More specific sub- ject area	
Type of data	Mass spectrometry
How data was acquired	Easy-nLC1000 coupled to a Q-Exactive orbitrap
Data format	Raw data (RAW files), search results (TXT files)
Experimental factors	Comparison of proteome coverage between species; comparison of honey bee proteome coverage with and without accounting for sequence poly-morphisms; comparison of protease activity across honey bee tissues
Experimental features	Protein samples were treated with dithiothreitol and iodoacetamide before trypsin digestion. Samples were desalted, then analyzed by nanoelectrospray ionization mass spectrometry (nESI-MS)
Data source location	Samples for the cross-species comparison were donated by researchers at the University of British Columbia, Vancouver, Canada. The bee for poly-morphism analysis came from York University, Toronto, Canada. All other bee tissues originated from the apiaries at the University of British Columbia Farm
Data accessibility	ProteomeXchange (PXD002275)

Value of the data

- The mass spectrometry dataset represents the highest honey bee proteome coverage to date and provides peptide evidence to help refine the honey bee genome annotation.
- We describe a detailed example of creating a personalized proteome database for a honey bee, and the code provided here can be used to construct a personalized protein database for any organism with known SNPs.
- The controlled cross-species proteomes dataset is suitable for evolutionary and bioinformatic hypothesis testing.
- These datasets, while focused on honey bees, should allow others to test hypotheses around the relative completeness of genome annotation or differential modification profiles among the main model organisms.

1. Data

We provide here the data used to investigate why honey bee proteomics experiments tend to result in fewer protein identifications compared to other commonly studied species. We include the raw mass spectrometry data files for the cross-species comparison, the honey bee whose genome was also sequenced, and the cross-tissue comparison. We also include the MaxQuant protein search file of the tissue comparison data and the perl script used to generate the customized polymorphic protein database as [Supplementary files 1](#) and [2](#), respectively. We have provided a navigation table ([Table 1](#))

Table 1
Description of the dataset.

Database(s)	Result file(s)	Data file(s)	Experiment
uniprot-(*).fasta; amel_OGSv3.2_pep.fa	[bee HeLa fly mouse worm yeast]_peptides.txt [bee HeLa dros mouse worm yeast]_summary.txt	species_ (*.).raw	"Cross-species comparison of proteome coverage"
amel_OGSv3.2_pep.fa	Supplementary material	tissue_ (*.).raw;	"Cross-tissue comparison of protease activity"
finalApisSNPPersonalizedDB.fasta	customOGS_peps.txt	customOGS_bee_head_3hr.raw	"Impact of accounting for genetic diversity"

to aid readers in identifying the relevant data files to download from our ProteomeXchange submission.

2. Experimental design, materials and methods

2.1. Sample sources

For the species comparison experiment, *Mus. musculus* liver tissue was provided by Nicholas E. Scott, *Saccharomyces cerevisiae* was provided by Patrick Chan, *Drosophila melanogaster* was provided by Carol Pollock, *Caenorhabditis elegans* was provided by George Chung. The *Apis mellifera*

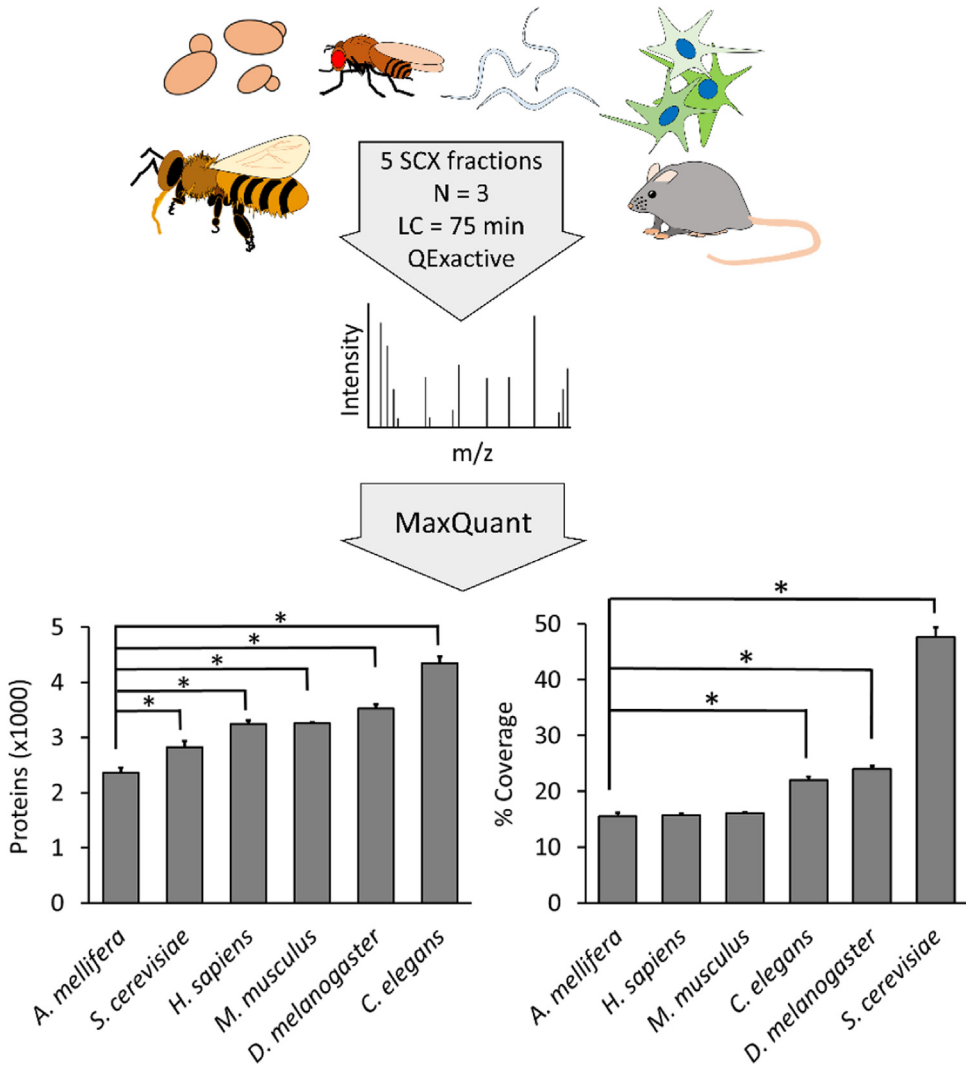


Fig. 1. Confirming low proteome coverage in honey bees. Proteome coverage across several well-studied species (*Apis mellifera*, *Saccharomyces cerevisiae*, *Homo sapiens*, *Drosophila melanogaster*, *Caenorhabditis elegans* and *Mus musculus*) was compared in a controlled experiment. Sample preparation techniques were identical wherever possible and all samples were run sequentially on the same Q-Exactive mass spectrometer and liquid chromatography system. Left graph: significantly fewer honey bee proteins were identified compared to all other species. Right graph: honey bees showed the lowest canonical proteome coverage overall. * $P < 0.05$; Student's *T*-test, $df = 4$.

sample for polymorphism analysis originated from York University in the summer of 2010, courtesy of Amro Zayed and Brock A. Harpur. All other *A. mellifera* samples were collected the University of British Columbia apiary in the summer of 2014.

3. Cross-species comparison of proteome coverage

3.1. Sample preparation

To provide experimental evidence for low protein identification rates in bees, we compared proteome coverage across six model systems in a controlled experiment (Fig. 1; Table 1; data files starting with “species_comparison”). All samples were processed and analyzed in biological triplicate. Protein was extracted in 50 mM ammonium bicarbonate (plus 1% sodium deoxycholate), with the aid of ceramic beads for mechanical disruption of coarse tissues. Clarified lysates were heated to 99 °C for 5 min, quantified using a bicinchoninic assay (Pierce), then were reduced (1 µg dithiothreitol per 50 µg protein), alkylated (5 µg iodoacetamide per 50 µg protein) and digested (1 µg porcine modified trypsin per 50 µg protein) overnight at 37 °C. Samples were acidified with 1% trifluoroacetic acid until pH < 2.0 and the precipitated deoxycholic acid was pelleted by centrifuging at 16,000g. Peptides in the supernatant were desalted using a high-capacity C18 STAGE-tip [2] and split into five fractions by strong cation exchange as described previously [3].

3.2. Data acquisition

The cross-species analysis was done in biological triplicate and equal amounts of protein were analyzed from each, all with equal amounts of instrument time. Fractionated peptides were analyzed by nLC–MS/MS using a nanoflow HPLC (Easy-nLC1000) coupled to a Q-Exactive mass spectrometer (Thermo). For each fraction, peptides were injected into the LC and loaded with Buffer A (0.5% acetic acid) onto an in-house packed fused-silica (5 µm Aqua C18 particles; Phenomenex) fritted trap column (2 cm long, 100 µm I.D., 360 µm O.D., 5 µL/min flow rate), then resolved on a reverse phase 75 µm I.D. fused silica, in-house packed 30 cm analytical column (ReproSil C18, 3 µm particle size; Dr. Maisch) using a 75 min linear gradient (90 min total run time) at a flow rate of 250 µL/min from 5% to 35% Buffer B (acetonitrile, 0.5% acetic acid), followed by a 15 min wash at 95% Buffer B. Instrument acquisition parameters included a 1% underfill ratio, intensity dependent MS/MS at 1.7e5 intensity threshold, and the instrument was set to scan from 300 to 2000 m/z with a 30 s dynamic exclusion time.

3.3. Search parameters

For each species, raw files were analyzed in a single MaxQuant search (version 1.5.2.8) [4] with all parameters left as default, except “match between runs” was enabled and data files were assigned their respective groups and fractions. The most recent honey bee Official Gene Set (OGS) protein database was downloaded from www.beebase.org (OGSv3.2). All other databases were downloaded from Uniprot proteomes (www.uniprot.org) on Feb. 5th, 2014.

4. Cross-tissue comparison of protease activity

We extracted protein from tissues with varying degrees of endogenous proteases (antennae, legs and digestive tract). The protein was processed and analyzed as described above, except it was not fractionated and the deoxycholic acid was removed by centrifuging through a 0.6 µm filter (Sartorius Stedim Biotech). 5 µg of peptides was injected (based on protein-level quantitation) per sample. The data was searched as above, but with both trypsin specificity and no enzyme specificity. In the absence of substantial protease activity, searches with no enzyme specificity typically decrease the total number of matched peptides owing to the increase in search space, except when endogenous

protease activity is abundant. This dataset represents the highest coverage of the honey bee proteome (protein identifications are provided as [Supplementary file 1](#)).

5. Impact of accounting for genetic diversity

Proteins from a single adult worker bee were extracted from the head essentially as outlined above but in 6 M urea lysis buffer [5] rather than ammonium bicarbonate and deoxycholate. Samples were quantified using a Bradford assay, then digested and analyzed on a Q-Exactive as described above, but with a three hour total analytical gradient. The data was searched using Mascot v2.5 with TrypsinMSIPI specificity and the decoy search option was used to filter for 1% false discovery rate. The protein database used in this search included this bee's specific SNPs, which was produced following the methods outlined in [6]. We have included our perl script used to generate the protein database as [Supplementary file 2](#).

Acknowledgments

This work was funded by a Discovery Grant to LJF from the Natural Sciences and Engineering Research Council of Canada (NSERC)(#311654-11). AM was supported by an NSERC Canada Graduate Scholarship. LJF is the Canada Research Chair in Quantitative Proteomics. The mass spectrometry infrastructure used here was supported by the Canada Foundation for Innovation, British Columbia (BC) Knowledge Development Fund and BC Proteomics Network.

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at <http://dx.doi.org/10.1016/j.dib.2015.10.042>.

References

- [1] Q.W.T. Chan, M.Y. Chan, M. Logan, M. Fang, H. Higo, L.J. Foster, Honey bee protein atlas at organ-level resolution, *Genome Res.* 23 (2013) 1951–1960. <http://dx.doi.org/10.1139/cjz-2014-0181>.
- [2] J. Rappsilber, Y. Ishihama, M. Mann, Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics, *Anal. Chem.* 75 (3) (2003) 663–670.
- [3] S. Michaud, H. Boncristiani, J.W. Gouw, et al., Response of the honey bee (*Apis mellifera* L.) proteome to Israeli acute paralysis virus infection, *Can. J. Zool.* (2015). <http://dx.doi.org/10.1139/cjz-2014-0181>.
- [4] J. Cox, M. Mann, MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification, *Nat. Biotechnol.* 26 (12) (2008) 1367–1372.
- [5] L.J. Foster, C.L. De Hoog, M. Mann, Unbiased quantitative proteomics of lipid rafts reveals high specificity for signaling factors, *Proc. Natl. Acad. Sci. USA* 100 (10) (2003) 5813–5818.
- [6] S. Schandorff, J.V. Olsen, J. Bunkenborg, et al., A mass spectrometry-friendly database for cSNP identification, *Nat. Methods* 4 (6) (2007) 465–466.