Data in Brief 5 (2015) 829–833

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

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

Alison McAfee, Sarah Michaud [1], 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.

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 7X81

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

More specific subject 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 polymorphisms; 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 polymorphism 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|dro|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"