

Top-Down-Assisted Bottom-Up Method for Homologous Protein Sequencing: Hemoglobin from 33 Bird Species

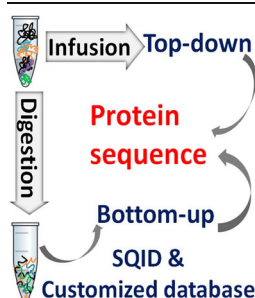
Yang Song,^{1,2} Ünige A. Laskay,² Inger-Marie E. Vilcins,³ Alan G. Barbour,⁴
Vicki H. Wysocki^{1,2}

¹Department of Chemistry and Biochemistry, The Ohio State University, Columbus, OH 43210, USA

²Department of Chemistry and Biochemistry, The University of Arizona, Tucson, AZ 85721, USA

³Emerging and Acute Infectious Diseases Branch, Department of State Health Services, Austin, TX 78756, USA

⁴Microbiology and Molecular Genetics, Medicine, and Ecology and Evolutionary Biology, University of California, Irvine, CA 92687, USA



Abstract. Ticks are vectors for disease transmission because they are indiscriminant in their feeding on multiple vertebrate hosts, transmitting pathogens between their hosts. Identifying the hosts on which ticks have fed is important for disease prevention and intervention. We have previously shown that hemoglobin (Hb) remnants from a host on which a tick fed can be used to reveal the host's identity. For the present research, blood was collected from 33 bird species that are common in the U.S. as hosts for ticks but that have unknown Hb sequences. A top-down-assisted bottom-up mass spectrometry approach with a customized searching database, based on variability in known bird hemoglobin sequences, has been devised to facilitate fast and complete sequencing of hemoglobin from birds with unknown sequences. These

hemoglobin sequences will be added to a hemoglobin database and used for tick host identification. The general approach has the potential to sequence any set of homologous proteins completely in a rapid manner.

Keywords: Bottom-up, Top-down, Customized database, Sequencing, Hemoglobin

Received: 23 February 2015/Revised: 8 May 2015/Accepted: 8 May 2015/Published Online: 26 June 2015

Introduction

Hemoglobin (Hb), a tetrameric protein with four polypeptide chains, typically two α and two β , is the most abundant protein in erythrocytes. It plays an important role in oxygen transportation [1]. Modifications and mutations on the amino acids of the hemoglobin sequences may reveal health states of individuals. For example, glycated hemoglobin is commonly used for diagnosis of diabetes [2, 3]. Severe hemoglobinopathies, including sickle cell disease, may involve only one amino acid substitution or deletion on hemoglobin chains [4]. Mutations, together with thalassemia, can lead to a more severe anemia [5]. Also, the fully sequenced hemoglobin can be utilized in sequence evolution analysis and ancestral sequence reconstruction [6]. The ability to identify small

variations in hemoglobin sequences unambiguously and quickly is therefore desirable in a broad clinical context.

Tick larvae often have their first blood meal on small mammals and birds and, hence, the possibility exists for the transmission of pathogenic microorganisms from those initial hosts to tick larvae. After feeding, the larvae leave their host and molt to nymphs. Later on, nymphs need to obtain a second blood meal from a larger host to molt to adult. The viruses and/or bacteria survive in the midgut of the tick for months without infecting the tick itself. As a consequence, disease acquired from the first host, if any exists, can be transmitted to a second host [7]. There are many bird species known to be host reservoirs for tick-borne diseases [8, 9]. The ability to identify major reservoirs in one geographical region is useful for disease prevention and intervention. It has been shown that host blood proteins still persist in ticks months after feeding and molting [10]. Because hemoglobin is the most abundant protein in erythrocytes and its sequence varies in different species [11], Hb remnants from host in ticks can be identified to reveal host identities [12]. We have obtained blood samples from 33 bird species (142 individuals) that are common in the U.S. as hosts

Electronic supplementary material The online version of this article (doi:10.1007/s13361-015-1185-z) contains supplementary material, which is available to authorized users.

Correspondence to: Vicki Wysocki; e-mail: wysocki.11@osu.edu

for ticks but that have unreported Hb sequences. Hb from most bird species contains one major component, HbA ($\alpha^A_2\beta_2$) and one minor component, HbD ($\alpha^D_2\beta_2$) [13–16]. Therefore, most bird species have three types of Hb chains: α^A , α^D , and β , with α^A and β chains being the most abundant Hb chains.

Bottom-up mass spectrometry, in which proteins are digested to form peptides that are fragmented in the mass spectrometer, is a very powerful tool in Hb identification. Gatlin and co-workers identified six variants of human hemoglobin by bottom-up proteomics and the SEQUEST-SNP algorithm, which dynamically generate all possible single-nucleotide polymorphisms (SNPs) [17]. Laskay and co-workers utilized a bottom-up mass spectrometry approach to sequence hemoglobin from nine mammalian species [12]. Guo and co-workers combined bottom-up mass spectrometry and X-ray diffraction to achieve complete sequence coverage of hemoglobin from two feline species [18]. Onder et al. employed spectral matching in a proteome profiling technique to identify vertebrate hosts of ticks without a priori knowledge of protein sequence information [19, 20]. The method which Onder et al. adopted is similar to a spectral library searching approach for peptide identification in bottom-up mass spectrometry, which involves matching query spectra to library spectra [21–23]. The most common method for identifying digested peptide ions from bottom-up experiments is database searching. During this process, the mass and fragmentation pattern of the ions are matched with in silico generated peptide ions from known protein sequences in a database [24]. Obviously, one drawback of this database searching strategy is that the existing databases may not necessarily cover the protein/peptide under study. To overcome this disadvantage, several de novo sequencing algorithms have been developed [25, 26]. However, de novo sequencing results are lower throughput and the output can be hard to organize into interconnected peptides because each tandem mass spectrum will return several candidate peptide sequences and a typical run usually result in tens of thousands tandem mass spectra. In this study, a different bottom-up data processing method is needed.

Top-down mass spectrometry, which accurately measures the intact molecular weight of a multiply-protonated protein and directly fragments the protein ions [27] without prior digestion to peptides is becoming increasingly popular because of advances in mass spectrometry instrumentation, such as higher resolution and accuracy [27–29], and improvements in top-down sequencing software [30, 31]. No digestion is performed in the top-down approach; thus, the connectivity of the peptides is not lost. The use of multiple fragmentation methods, such as collision-induced dissociation (CID) and electron transfer dissociation (ETD), facilitate better fragmentation coverage [32]. The advantages of top-down MS include precise determination of protein isoforms and localization of amino acid substitutions and post-translational modifications (PTMs). Edwards and coworkers have successfully applied a top-down approach to screen neonatal dried blood spot samples for hemoglobin variants [33, 34].

Here, we developed a top-down-assisted bottom-up mass spectrometry approach with a customized search database to successfully sequence the α^A and β hemoglobin chains from the 33 bird species under investigation. The customized database was produced by aligning the known avian hemoglobin sequences from the National Center for Biotechnology Information (NCBI, 06/12/12), keeping sequences that are invariant in those constant and generating peptide sequences where variable amino acids were successively replaced with each canonical amino acid. The final resulting sequences produced by using this customized database to sequence the Hb of birds with unknown sequence and confirmed further with the top-down approach will be available to append to existing hemoglobin databases so that LC-MS/MS of protein digests or ESI or MALDI MS of intact proteins can be searched against this database to reveal host identities.

Experimental

Full Scan Mass Spectrometry Analysis of Bird Blood

Blood samples from 33 bird species were collected in purple Vacutainers containing anticoagulants and stored in a -80°C freezer. The birds were trapped as part of ongoing vector competence studies at Stone Lakes National Wildlife Refuge. The area is composed of seasonal and permanent wetlands, riparian forest, grasslands, and freshwater lakes, and is home to several colonial and migratory water bird species. Within those, five species were represented by only one bird per species. The other 28 species were represented by multiple birds per species serving as biological replicates. To determine the difference of the intact mass of Hb that may exist between individuals within the same species, the blood samples were lysed, diluted, and analyzed by a SYNAPT G2S Q-IM-TOF mass spectrometer (Waters Corporation, Manchester, UK). To prepare the spray solution, the blood aliquot was centrifuged at 14,000g for 10min to separate plasma (supernatant) and packed red blood cells. Two microliters of red blood cells were incubated with 25 μL of water (Optima LC/MS water, Thermo Fisher Scientific, Waltham, MA, USA) for 5min for cell lysis. Cell lysate was centrifuged at 14,000g for 5min. The supernatant from the red blood cell lysate was diluted 200-fold with 0.1% formic acid (Thermo Fisher Scientific) in water in preparation for analysis. Multiply charged hemoglobin subunits were observed and deconvoluted to zero charge mass with the SYNAPT's embedded MaxEnt 1 algorithm. Some settings for MaxEnt 1 were: mass range 5000–20,000, resolution 1Da/channel, damage model uniform Gaussian width at half height 0.5Da, minimum intensity ratios left 33% right 33%.

Bottom-Up Experiments

All the chemicals were purchased from Sigma-Aldrich (St. Louis, MO, USA) unless otherwise specified. For 16 bird species, the bottom-up sequencing approach was performed after in-gel tryptic digestion of bird hemoglobin, with LC-MS/MS

analysis on an LTQ mass spectrometer (Thermo Fisher Scientific) coupled to a capillary-flow LC (Surveyor, Thermo Finnigan, San Jose, CA, USA). Fifteen microliters of red blood cell lysate was mixed with 5 μ L 4 \times Laemmli protein sample buffer (250mM Tris-HCl, pH6.8, 8% SDS, 30% glycerol, 0.02% bromophenol blue, 1% β -mercaptoethanol) followed by heating at 95°C for 5min. The sample was loaded on a 4%–20% precast protein gel (Mini-PROTEAN TGX; Bio-Rad, Hercules, CA, USA) together with 10 μ L Precision Plus Protein standard (Bio-Rad) in the adjacent lane. The gel was run at 90V constantly for approximately 1.5h until the dye front reached the reference line. The gel was rinsed three times in pure water (Millipore, Billerica, MA, USA), and stained with Bio-Safe Coomassie (Bio-Rad) for 1h, then background destained with pure water overnight on a shaker (Boekel, Feasterville, PA, USA). The hemoglobin band (around 15kDa) was excised, diced into 1mm³ cubes, and transferred to a 1.5mL Eppendorf tube. The gel pieces were washed with 200 μ L pure water for 30s and destained twice with 200 μ L of methanol:50mM ammonium bicarbonate (1:1v/v) for 1min with vortex mixing. The gel pieces were then incubated for 5min in 200 μ L of acetonitrile:50mM ammonium bicarbonate (1:1v/v) with vortex mixing. The gel pieces were further dehydrated in 200 μ L acetonitrile then dried in a SpeedVac (Savant, Irvine, CA, USA) for 15min. The dried gel pieces were reduced with 200 μ L freshly prepared 25mM dithiothreitol for 20min at 56°C, then alkylated with 200 μ L 55mM iodoacetamide at room temperature for 20min in the dark. The gel pieces were washed twice with 500 μ L of pure water by vortex mixing briefly. Dehydration and drying as described before were repeated. In-gel digestion was performed by incubating in 50 μ L of 20ng/ μ L Trypsin Gold (Promega, Madison, WI, USA) in 0.01% ProteaseMAX (Promega) at 50°C for 1h. The digestion was stopped by adding in 0.5% (final v/v) trifluoroacetic acid to inactivate trypsin. The solution was transferred to a new 1.5mL Eppendorf tube and centrifuged at 14,000g for 10min. The supernatant was subjected to LC-MS/MS analysis. The LC system was modified from a microbore HPLC (Surveyor) to achieve capillary flow. The column (8cm \times 100 μ m i.d.) used was packed with 5 μ m Xorbax C18 resin (Agilent, Santa Clara, CA, USA). Peptides were separated with a gradient using mobile phase A (0.1% formic acid in water) and B (0.1% formic acid in acetonitrile). The initial flow was 5min at 1.25 μ L/min at 5% B, followed by 100min of 0.5 μ L/min with 5% B linearly ramping up to 45% B. Five min at 1.25 μ L/min at 95% B was used as a final wash step. An LTQ was used, with the full MS range set to 400 to 1600, and the three most intense ions automatically selected by Xcalibur software for fragmentation.

In addition to the above approach, which was used at the University of Arizona for approximately 16 bird species, the bottom-up sequencing approach for 17 bird species was performed at OSU by in-solution tryptic digestion of bird blood, with LC-MS/MS analysis on a dual linear ion trap Velos Pro mass spectrometer (Thermo Fisher Scientific) coupled to a nanoACQUITY UPLC (Waters). Fifteen microliters of blood cell lysate obtained as described above was mixed with 28.5 μ L 100mM ammonium bicarbonate, 1.5 μ L 1% ProteaseMAX

(Promega), and 5 μ g Trypsin Gold (Promega). Digestion was performed at 37°C for 3h and stopped by adding in 0.5% (final v/v) trifluoroacetic acid. After centrifuging the digestion solution at 14,000g for 10min, supernatant was subjected to LC-MS/MS analysis. A Thermo EASY-Spray Column (particle size: 3 μ m; 150 \times 0.075mm i.d.) was used. Flow rate was kept at 0.4 μ L/min and the gradient started at 1% B, ramping to 35% B in 84min. The full MS range was set to 400 to 2000, and the 10 most intense ions were automatically selected using Xcalibur software for fragmentation.

Bottom-Up Data Analysis

Because most of the bird species examined here have unreported hemoglobin sequences, we decided that a statistical analysis of known avian Hb sequences should first be performed to identify conserved regions in avian hemoglobin. The conserved regions could then be used in a customized database. We chose to use our in-house protein identification algorithm SQID [35] for peptide identification because it showed an improved number of identified peptides, attributable to incorporating pairwise fragmentation intensity information. The work flow of data analysis is shown in Figure 1. The statistical analysis of avian Hb in the National Center for Biotechnology Information (NCBI, 06/12/12) bank indicates that basic residue (arginine and lysine) sites are 98.3% and 99.8% conserved among 1045 avian Hb α chains and 760 avian Hb β chains. This gave us confidence in segmenting the Hb chain into tryptic peptides ending with K or R. Peptides with enumerated mutations at variable sites were then generated to assemble hypothetical position-labeled peptide groups. The customization step is elaborated in the Results section. A general sequence of a Hb chain can be obtained by combining the SQID top hits from each hypothetical position-labeled peptide group in numerical order. The peptide group name (e.g., A1–A9; B1–B12) represents the position of the peptide in the Hb chain.

Top-Down Experiments and Data Analysis

Top-down data acquisition was performed on an Orbitrap Elite mass spectrometer (Thermo Fisher Scientific) with ESI infusion of bird blood lysate diluted in water/acetonitrile (1:1v/v) containing 0.1% formic acid via a Heated Electrospray Ionization (HESI) source. Typical parameters used were: flow rate 4 μ L/min, spray voltage 3.5kV, sheath gas flow rate 5, aux. gas flow rate 0, sweep gas flow rate 0, capillary temperature 275°C. Collision induced dissociation (CID) and electron transfer dissociation (ETD) were performed on the multiply charged (mostly, between +14 to +18) ions of the intact proteins. The isolation window was 2Da. The CID energy was 20%–30% and ETD reaction time was 10–15ms depending on precursor ion charge state and mass. Top-down data were charge-deconvolved and deisotoped by Thermo Xtract software. The Xtract settings were: generate MH⁺ masses mode, resolution 240,000 @400, S/N threshold 3, fit factor 44%, remainder 25%, AveragineLowSulfur model, max. charge 30. After the fragments with S/N over 3 are deconvoluted and deisotoped,

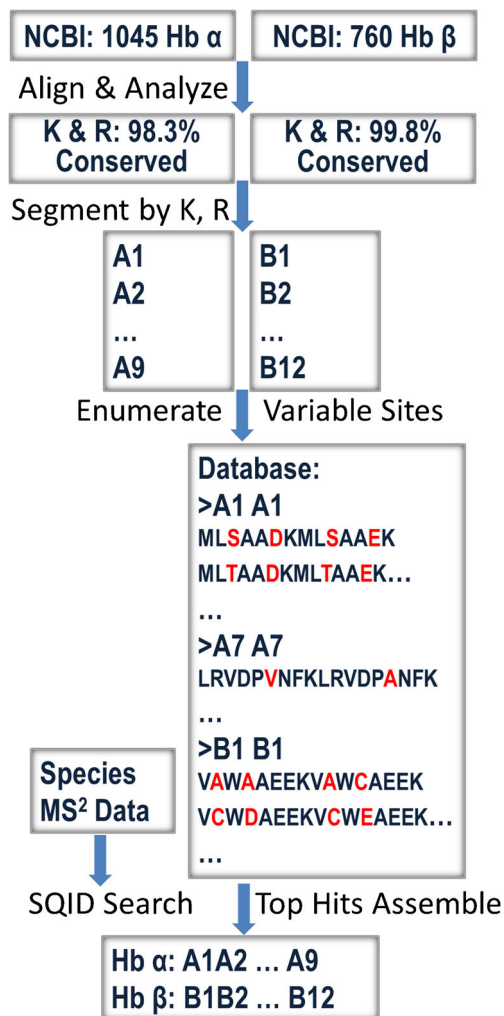


Figure 1. Flow chart of bottom-up data analysis

10ppm mass accuracy was used for manual matching based on the results from bottom-up. To show the abundance of products produced in top-down sequencing, product maps were generated by an in-house built Matlab script with input of the final sequences and the charge-deconvoluted MS² peak lists. The Matlab script is available in Online Resource 3.

Phylogenetic analysis of the Hb α^A chains was done via phylogeny.fr [36, 37]. Two sequenced bird species in the NCBI database were subjected to analysis together with 35 sequenced α^A chains (two out of 33 species have two α^A isoforms) in this work.

Results and Discussions

Intact Mass of Hemoglobin Chains

Blood samples from 28 out of the 33 bird species under study were collected from more than one individual bird. To check the heterogeneity of Hb within a species, the mass of Hb chains were measured on the SYNAPT G2S mass spectrometer. Figure 2 and Online Resource 1 “ESM_1. Intact average mass of hemoglobin

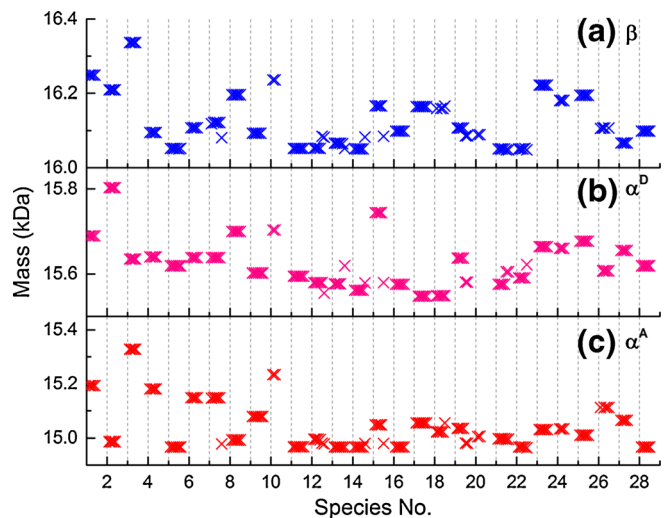


Figure 2. Intact mass of hemoglobin chains (a) β , (b) α^D , (c) α^A from 28 bird species containing multiple bird individuals. Individual bird masses within a species are represented by “X” symbols offset on the x-axis to improve readability. Most bird species have a consensus mass among different individuals. Several species, namely 7, 12, 13, 14, 15, 18, and 19, contain outliers

chains” show the intact average mass of α^A , α^D , and β chains for the 28 bird species that contain more than one individual bird. For most of the species, the same masses were present for the individual sampled birds in a given species, whereas some species, namely No.7 *Catharus ustulatus*, No.12 *Icterus bullockii*, No.13 *Junco hyemalis*, No.14 *Melospiza melodia*, No.15 *Myiarchus cinerascens*, No.18 *Pheucticus melanocephalus*, and No.19 *Picoides nuttallii* showed one or two outlier(s). We assume hemoglobin chains within a specific species have the same amino acid sequence if they have the same mass. The different masses observed in the outliers may be due to sampling of different phenotypes, misassignment of the species at sampling, or environmentally-induced hemoglobin mutations [38]. The sequences reported for a given species in this paper are from the individual bird samples that all provided the same masses. Outlier mass values are included in Online Resource 1.

Customizing a Database for Searching

Bottom-up proteomics typically utilizes sequence database searching [24], a popular approach to identify sequence of the digested peptides, although spectral library searching [21–23] is also used. In sequence database searching, the intact mass and fragmentation pattern of the experimental peptide ions are matched to fragmentation spectra or m/z lists of peptide fragment ions produced for peptide sequences generated by in silico digestion of protein sequences in a database. This approach is a reasonably fast and high-throughput method to identify proteins. Because the Hb chains from the bird species in this study have not been sequenced before, a customized database was created in order to adopt the database searching approach.

The concept of customizing a sequence database has been applied to many studies to increase peptide detection sensitivity

[39, 40]. Khatun et al. generated a database by 3-frame translation of a transcript database and 6-frame translation of genomic sequences [41]. Wang and coworkers created a protein sequence database from abundant RNA-seq data of a specific sample and further incorporated high-quality single nucleotide variations into the customized protein database [42, 43]. Park et al. considered single nucleotide variants, insertions, deletions, and stop-codon mutations derived from Exome-seq and RNA-seq to build a database for a specific sample [44].

In our study, the steps of generating a customized database are shown in Figure 1. First, a statistical analysis of amino acids in already sequenced avian Hb chains was performed. From the NCBI database, 760 entries of avian Hb β chains were available; these were aligned and all were 146 amino acids in length. Arginine and lysine sites were 99.8% conserved among the 760 entries, so the β chain can be segmented into 12 peptides ending with K or R. For example, for the first peptide VHWT AEEK, with variants noted at H2 and T4, and no R, K, or I included (because I and L were treated as one amino acid, labeled L), there are 17 (20 minus 3) potential amino acid substitutions on the second and the fourth positions. The second site mutation and the fourth site mutation were considered independently. Therefore, a total of 289 (17 multiplied by 17) VXWXAEEK peptide segments are generated as hypothetical position-labeled peptide group “B1,” in which X represents amino acids except for K, R, and I. In a similar way, 12 (B1 to B12) position-labeled peptide groups were constructed in the database. Each peptide group B1–B12 exists as 2 to 83,521 sequences in the database. Hb α chains are more varied because there are two copies of the α chain, one major component α^A , and one minor component α^D . Both are 141 amino acids long. In order to compromise between computational expense and the numerous variations of the peptides, some variable sites were limited to a certain number of substitutes that have over 1% occurrence in the sequenced Hb chains in the NCBI database. This compromise was justified by the results: more than 60% sequence coverage in bottom-up alone for α^A from any species was achieved. Top-down experiments performed later on enabled us to fully differentiate α^A from α^D . Nine (A1 to A9) position-labeled peptide groups for Hb α chain were generated for the database along with the 12 for the Hb β chain.

The constitution of each position-labeled peptide group is listed in Table 1. In this table, variable site X represents all amino acids except for K, R, and I. Variable site X^k represents all amino acids except for R and I (due to occurrence of K at these specific sites in some of the birds of known sequence). Sites with limited substitution have all the possible amino acids listed vertically. The limited substitution is determined by the fact that in 99% of known sequences of avian hemoglobin, one of the listed residues is occupied at those sites. Peptides with enumerated mutations at variable sites (highlighted in yellow) were generated to assemble hypothetical position-labeled peptide groups (A1–A9, B1–B12). The number of sequences in each peptide group is listed in the third column of Table 1. An experimental peptide mass and MS/MS data were searched by the SQID proteomics search engine against each peptide in the

hypothetical database. Because the peptide name represents the position of the peptide in the chain, a general sequence of a Hb chain can be obtained by combining the most confident peptide from each hypothetical position-labeled peptide group in the numerical order (A1–A9; B1–B12).

Bottom-Up Approach

The software SQID [35] first indexed the database into peptides ending with K or R. The allowed peptides had mass ranges from 700 to 4500Da and up to one missed cleavage. One missed cleavage is allowed during indexing of the database to accommodate the detection of regions containing internal R or K, such as A4, A7, and A8. Variable modifications included oxidation of methionine and carbamidomethylation on cysteine because reduction and alkylation were performed only on about half the hemoglobin samples (those subjected to in-gel digestion). Thermo Discoverer (V1.4) was used to convert the LC-MS/MS raw data to .dta files. Each .dta file contained the precursor mass (calculated based on measured m/z), charge, fragments' m/z values, and fragments' intensities. In order to determine the peptide sequence of each experimental tandem mass spectrum or .dta file, the precursor mass was compared with those of the in silico digested peptides and sequence candidates were chosen within 1.5Da mass error of the precursor mass. A series of b and y ions and their neutral loss ions were then calculated for each of the candidates. To evaluate the accuracy of the match between the candidates and the experimental tandem mass spectrum, a scoring function considering the number of matched ions (within 0.8Da mass error), the number of consecutively matched ions, and a bonus for strong pairwise fragmentation intensity was utilized. The sequence candidate with the highest score was assigned to the experimental tandem spectrum. The output of SQID for each sample was a spreadsheet that included information such as spectral name, corresponding sequence, the group number that contains the peptide, etc. For each peptide group, which is a collection of all the possible sequences at one tryptic cleavage region of the hemoglobin chains, the sequence with the highest SQID score is more likely to be the true sequence of the region. To be more specific, if more than one sequence was identified for a position-labeled peptide group, a minimum SQID score that corresponds to an FDR of 5% and a maximum delta SQID score of 0.3 were set to choose the sequence candidates. The FDR was established by searching against reverse decoys appended to our customized database, and the SQID cutoff was determined to be 16–21 for different species. The delta SQID score is represented by the normalized difference between the SQID score of the top ranked sequence for the position-labeled peptide group and current sequence as shown in the following equation:

$$\text{Delta SQID score} = (\text{SQID score}_{\text{rank1}} - \text{SQID score}) / \text{SQID score}_{\text{rank1}}$$

Sequences with small, but not 0, delta SQID scores are potential sequences for a specific Hb region, and thus should

Table 1. Peptide Components in the Customized Database (Variable Sites are Highlighted)

Position	Sequence Variation	#
A1	V M L S T X X D E K	2,312
A2	I G G H A E E Y G A N S D E A L Q R A A S T H A E E Y G A N S D E A L Q R V S T Q Q D D F S E A L Q T	31,104
A3	M L F X T Y P X T K L F X T S Y P X T K	1734
A4	T Y F P H F D L V X X ^k G S A E D G I K T Y F P H F D L V X X ^k G S A E D G I K T Y F P H F D L V X X ^k G S A E D G I K	9792
A5	V V G A L V E N A V A N H S N I V D N I A S Q T L S K V V G A L V E N A V A N H S N I V D N I A S Q T L S K V V G A L V E N A V A N H S N I V D N I A S Q T L S K	15,552
A6	L S D L H A Q Y K L S D L H A Q Y K	4
A7	L R V D P V A N F K L R V D P V A N F K	2
A8	F L G S H C F L V V V A V H H L P G A A D L T S P E V M H A S A L F D K F L G S H C F L V V V A V H H L P G A A D L T S P E V M H A S A L F D K F L G S H C F L V V V A V H H L P G A A D L T S P E V M H A S A L F D K	2,654,208
A9	F M C A V G T A T A E K F M C A V G T A T A E K F M C A V G T A T A E K	192
B1	V X W X A E E K	289
B2	Q L I X X X W X K	83,521
B3	V N V X X C G X E A L A R	4,913
B4	L L X X Y P W T Q R	289
B5	F F A S T F G N L S S G P T A V X G N P K	544
B6	V L X S F G X A V K	289
B7	N X D X I K	289
B8	X T F X X L S E L H C D K	4,913
B9	L H V D P E N F R K	2
B10	L L G D V L V V V L A X H F X K	4,624
B11	D F S T P X X Q A A W Q K	578
B12	X V A H A L A R	17

X represents amino acids except for K, R, I.

X^k represents amino acids except for R, I.

be considered when constructing the general sequences of Hb from the bottom-up approach.

The Online Resource 2 “ESM_2. Peptide MS/MS Spectra Identified by SQID with Customized Database” contains all the successfully identified hemoglobin peptides for each of the bird species by this method. The confidence of assigning these sequences is strengthened by top-down analysis. Information on each page of the .PDF file includes species number, species’ common name, peptide sequence, hypothetical protein name (which represents the location of the peptide in the hemoglobin chain), raw data file identifier, SQID score, the peptide spectrum, and a three-row sequence cell. In each spectrum, b and y ion peaks are labeled and plotted in blue and red, respectively.

The green peaks represent neutral losses from the b/y ions. The three-row sequence cell facilitates counting the b and y ions.

Top-Down Completion of Partial Bottom-Up Sequences

Top-down data analysis was used to fill in missing peptides and to confirm the sequences identified in the bottom-up approach. Both collision-induced dissociation (CID) and electron transfer dissociation (ETD) methods were performed to improve the sequence coverage. All the precursor masses and fragment masses are within 10ppm error. Figure 3 shows the product ions of intact hemoglobin α^A and β chains from *Zonotrichia*

atricapilla species generated by CID (b, y ions; blue mark) and ETD (c, z ions; red arrow). The abundance of fragment ion information generated from the top-down approach provided sequence information that connected peptides identified in the bottom-up approach. For example, Figure 4 compares the result obtained from the bottom-up approach only to the result from the top-down-assisted bottom-up approach. The result from the top-down-assisted bottom-up approach is shown on the rows with amino acid number labels (1, 31, 61, 91, and 121) and the amino acids codes are in black or gray. A1 to A9 and B1 to B12 correspond to regions of hypothetical position-labeled peptide groups. A1, B7, B9, and B11 do not have confident hits returned from bottom-up database searching because all hits had SQID scores lower than 16 (FDR 5%). B2 has a confident hit “QLIPSGWGK” returned from the bottom-up approach (SQID score 29), but the sequence does not explain the fragment ions in the top-down experiments. A2 through A9, B1, B3, B4, B5, B6, B8, B10, and B12 are partially or completely sequenced correctly by the bottom-up approach. For the Hb α^A and β chains of *Zonotrichia atricapilla*, 67% and 66% sequence coverages, respectively, were achieved without the assistance of the top-down method. However, the top-down-assisted bottom-up method achieved 100% confidence sequence coverage with ambiguity only at I and L. Moreover, the top-down method can provide post-translational modification information. For example, top-down data showed that K16 of the *Corvus brachyrhynchos* Hb α^A chain was methylated. This modification was unexpected and, thus, was not included in the database searching of the bottom-up data. The bottom-up approach is also necessary because there are some regions in the middle of the hemoglobin sequences that do not have much fragmentation in the top-down approach, for example as shown in Figure 3, not much fragmentation has been observed in the β^{67-76} (VLTSFGAEAVK) region, but this region was successfully detected in the bottom-up approach as a top hit for the hypothetical position-labeled peptide group B6. All the product ion maps of α^A and β from the 33 bird species are in Online Resource 3. The resulting sequences and deconvoluted top-down mass spectra were further subjected to ProSight Lite [45] matching and the *P*-score and the protein characterization

score (PCS) were obtained. For all the Hb sequences, the median *P*-score is 2.4E-71 and the median PCS is 822.68. These scores add confidence to our sequencing results.

Phylogenetic analysis (Figure 5) on the 35 sequenced Hb α^A chains (two species have two α^A isoforms) together with two known sequences was performed by uploading the sequences to the server phylogeny.fr [36, 37] in “One click” phylogeny analysis mode, and the result is consistent with the conventional classification of the avian organisms. In Figure 5, the bird name is in the format of “Order_Family_Genus_Species”, which is the conventional classification. The phylogenetic tree is indicative of how similar the sequences are: dissimilar sequences diverge early on the tree while similar sequences stay in the same branch. For example, the three species *Colaptes auratus*, *Melanerpes formicivorus*, and *Picoides nuttallii* are known to come from the same order PICIFORMES. They are distant from all the other species that belong to the PASSERIFORMES order or the GALLIFORMES order. This is confirmed by the phylogenetic tree produced from the protein sequences of the 37 Hb α^A chains. Hb sequences from species *Vireo gilvus*, *Vireo cassinii*, and *Vireo huttoni* are from the same genus *Vireo*, and they are grouped together. Similarly, Hb sequences from the species *Catharus guttatus* and *Catharus ustulatus* are from the same family TURDIDEA and the same genus *Catharus*, and they grouped into the same branch. *Turdus migratorius* are from the same family (TURDIDAE) as *Catharus guttatus* and *Catharus ustulatus*, but are not the same genus (*Turdus* versus *Catharus*), and diverged a little earlier on the tree. *Thryomanes bewickii* and *Troglodytes aedon* from the TROGLODYTIDAE family, *Haemorrhous mexicanus* and *Haemorrhous purpureus* from the FRINGILLIDAE family, and *Haemorrhous* genus, *Euphagus cyanocephalus* and *Icterus bullockii* from the ICTERIDAE family, *Junco hyemalis*, *Melospiza melodia*, *Melospiza crissalis*, *Passerella iliaca*, *Pipilo maculatus*, *Zonotrichia atricapilla*, and *Zonotrichia leucophrys* from the EMBERIZIDAE family were all grouped separately based on their family. Aberrant from the classic classification are the species from the CARDINALIDAE family and the CORVIDAE family. They were grouped closely but not as close

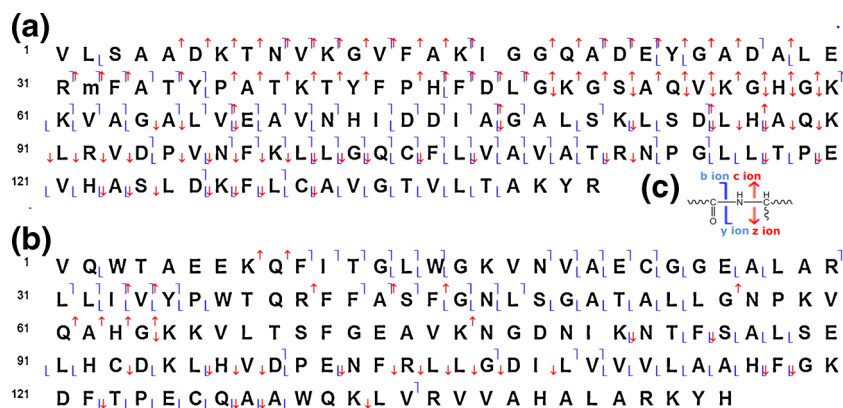


Figure 3. MS/MS product ion map of Hb α^A chain (a) and β chain (b) from *Zonotrichia atricapilla*. (c) Legend for fragmentation symbols

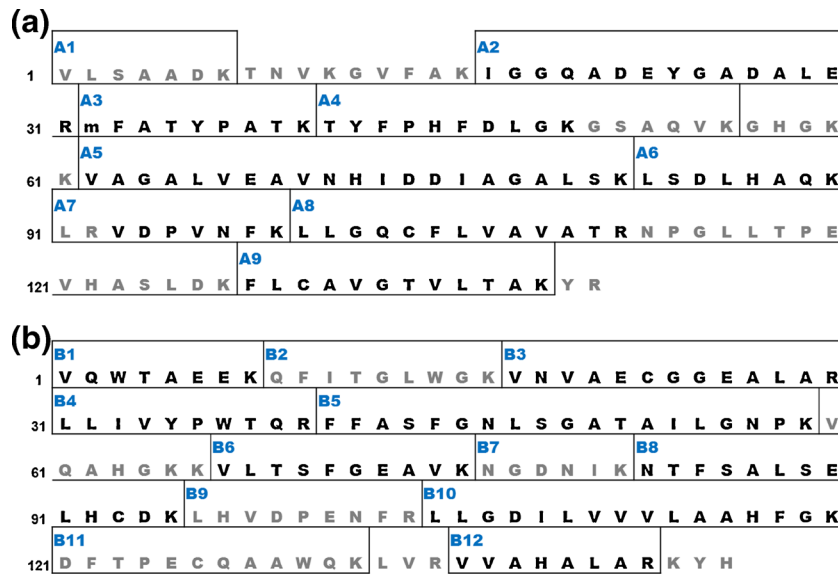


Figure 4. Peptides aligned from a SQUID search with a customized database in the bottom-up method for *Zonotrichia atricapilla* α^A chain (a) and β chain (b). The amino acid letters show the sequencing results from the top-down-assisted bottom-up method. The boxes with the blue A1-A9 or B1-B12 show the Hb sequence segmentation in the bottom-up method. Peptides in black are correctly identified by bottom-up. The gray regions are only identified in the top-down approach

as exactly the same branch. The conventional classification is manifested by phylogenetic tree construction using genetic data, which contains 19 loci, topological constraints, and fossil

constraints [46]. Our phylogenetic tree is based only on one parameter, the hemoglobin sequences, which may explain the aberrant results. The already sequenced *Gallus gallus*, which

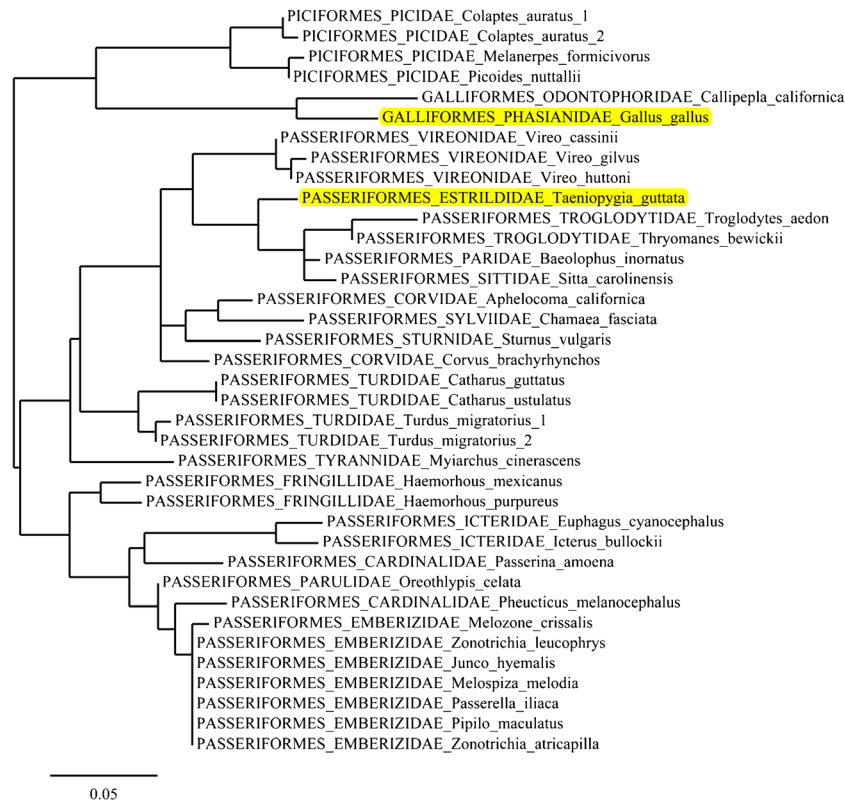


Figure 5. Phylogenetic tree of birds based on sequences of their Hb α^A chains and generated by using phylogeny.fr. The tree matches very well with the conventional classification (reflected by the name “Order_Family_Genus_Species”). The two known sequences from NCBI, and which were inserted as controls, are highlighted

belongs to GALLIFORMES order, was branched together with *Callipepla Californica* from our study, adding confidence to our sequencing results.

Conclusions

We have presented a top-down-assisted bottom-up sequencing approach. Building a customized database based on homologous and variable regions of already sequenced avian hemoglobin for the bottom-up approach expedited the alignment of the peptide data, thus providing a reliable model for top-down data analysis. The top-down approach played a role of quality control in confirming the peptides alignment in the customized database bottom-up approach but also provided sequence in regions where bottom-up failed. The combined approach has been successfully applied to sequence α^A and β hemoglobin chains from 33 avian species. The sequences obtained in this study will be appended to current hemoglobin databases for tick/host studies. In the future, the top-down-assisted bottom-up sequencing approach can be applied to facilitate fast and complete sequencing of other homologous proteins.

Acknowledgments

The authors acknowledge support for this research by the National Institutes of Health (NIH) grant U54 AI065359 and The Ohio State University internal funds. Bottom-up experiments were partially performed at the University of Arizona. The authors thank the Arizona Proteomics Consortium for instrument access.

References

- Lehninger, A.L., Nelson, D.L., Cox, M.M.: Lehninger principles of biochemistry, 4th edn, pp. 157–188. W. H. Freeman, New York (2005)
- Inzucchi, S.E.: Diagnosis of diabetes. *New. Engl. J. Med.* **367**, 542–550 (2012)
- Alqahtani, N., Khan, W.A.G., Alhumaidi, M.H., Ahmed, Y.A.A.R.: Use of glycated hemoglobin in the diagnosis of diabetes mellitus and pre-diabetes and role of fasting plasma glucose, oral glucose tolerance test. *Int. J. Preventive Med.* **4**, 1025–1029 (2013)
- Stuart, M.J., Nagel, R.L.: Sick-cell disease. *Lancet* **364**, 1343–1360 (2004)
- Williams, J.P., Scrivens, J.H., Green, B.N., Farrar, L.M., Sutcliffe, M.: Hb Leeds [β 56(D7)Gly→Cys]: a new hemoglobin that aggravates anemia in a child with β 0-thalassemia trait. *Hemoglobin* **31**, 367–373 (2007)
- Grispo, M.T., Natarajan, C., Projecto-Garcia, J., Moriyama, H., Weber, R.E., Storz, J.F.: Gene duplication and the evolution of hemoglobin isoform differentiation in birds. *J. Biol. Chem.* **287**, 37647–37658 (2012)
- Goodman, J.L., Dennis, D.T., Sonenshine, D.E.: Tick-borne diseases of humans, pp. 12–36. ASM Press, Washington, D.C. (2005)
- Björnsdóttir, A., Bergström, S., Massung, R.F., Haemig, P.D., Olsén, B.: Ehrlichia-infected ticks on migrating birds. *Emerg. Infect. Dis.* **7**, 877–879 (2001)
- Olsén, B., Jaenson, T.G.T., Bergström, S.: Prevalence of *Borrelia burgdorferi* sensu lato-infected ticks on migrating birds. *Appl. Environ. Microbiol.* **61**, 3082–3087 (1995)
- Wickramasekara, S., Bunikis, J., Wysocki, V.H., Barbour, A.G.: Identification of residual blood proteins in ticks by mass spectrometry proteomics. *Emerg. Infect. Dis.* **14**, 1273–1275 (2008)
- Hall, B.K.: Homology: the hierarchical basis of comparative biology. Academic Press, Waltham (1994)
- Laskay Ünige, A., Burg, J., Kaleta Erin, J., Vilcins Inger-Marie, E., Telford, I.I.S.R., Barbour Alan, G., Wysocki Vicki, H.: Development of a host blood meal database: de novo sequencing of hemoglobin from nine small mammals using mass spectrometry. *Biol. Chem.* **393**, 195 (2012)
- Vandecasserie, C., Paul, C., Schnek, A.G., Léonis, J.: Oxygen affinity of avian hemoglobins. *Comparative biochemistry and physiology. Part A: Physiol.* **44**, 711–718 (1973)
- Nothum, R., Weber, R.E., Kusters, J., Schneegeans, D., Braunitzer, G.: Amino-acid sequences and functional differentiation of hemoglobins A and D from swift (*Apus apus*, Apodiformes). *Biol. Chem. Hoppe-Seyler* **370**, 1197–1207 (1989)
- Tamburrini, M., Riccio, A., Romano, M., Giardina, B., di Prisco, G.: Structural and functional analysis of the two haemoglobins of the Antarctic seabird *Catharacta maccormicki*. *Eur. J. Biochem.* **267**, 6089–6098 (2000)
- Rana, M.S., Knapp, J.E., Holland, R.A.B., Riggs, A.F.: Component D of chicken hemoglobin and the hemoglobin of the embryonic Tammar wallaby (*Macropus eugenii*) self-associate upon deoxygenation: effect on oxygen binding. *Proteins: Struct. Funct. Bioinformatics* **70**, 553–561 (2008)
- Gatlin, C.L., Eng, J.K., Cross, S.T., Detter, J.C., Yates, J.R.: Automated identification of amino acid sequence variations in proteins by HPLC/microspray tandem mass spectrometry. *Anal. Chem.* **72**, 757–763 (2000)
- Guo, J., Uppal, S., Easton, L.M., Mueser, T.C., Griffith, W.P.: Complete sequence determination of hemoglobin from endangered feline species using a combined ESI-MS and X-ray crystallography approach. *Int. J. Mass Spectrom.* **312**, 70–77 (2012)
- Onder, O., Shao, W., Lam, H., Brisson, D.: Tracking the sources of blood meals of parasitic arthropods using shotgun proteomics and unidentified tandem mass spectral libraries. *Nat. Protoc.* **9**, 842–850 (2014)
- Onder, O., Shao, W., Kempes, B.D., Lam, H., Brisson, D.: Identifying sources of tick blood meals using unidentified tandem mass spectral libraries. *Nat. Commun.* **4**, 2730 (2013)
- 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)
- Frewen, B.E., Merrihew, G.E., Wu, C.C., Noble, W.S., MacCoss, M.J.: Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries. *Anal. Chem.* **78**, 5678–5684 (2006)
- Lam, H., Deutsch, E.W., Eddes, J.S., Eng, J.K., Stein, S.E., Aebersold, R.: Building consensus spectral libraries for peptide identification in proteomics. *Nat. Methods.* **5**, 873–875 (2008)
- Eng, J.K., McCormack, A.L., Yates III, J.R.: An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5**, 976–989 (1994)
- Frank, A., Pevzner, P.: PepNovo: De novo peptide sequencing via probabilistic network modeling. *Anal. Chem.* **77**, 964–973 (2005)
- 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.* **17**, 2337–2342 (2003)
- Kelleher, N.L.: Peer reviewed: top-down proteomics. *Anal. Chem.* **76**, 196A–203A (2004)
- McLafferty, F.W., Fridriksson, E.K., Horn, D.M., Lewis, M.A., Zubarev, R.A.: Biomolecule mass spectrometry. *Science* **284**, 1289–1290 (1999)
- Kelleher, N.L., Lin, H.Y., Valaskovic, G.A., Aaserud, D.J., Fridriksson, E.K., McLafferty, F.W.: Top down versus bottom up protein characterization by tandem high-resolution mass spectrometry. *J. Am. Chem. Soc.* **121**, 806–812 (1999)
- LeDuc, R.D., Taylor, G.K., Kim, Y.-B., Janusz, T.E., Bynum, L.H., Sola, J.V., Garavelli, J.S., Kelleher, N.L.: ProSight PTM: an integrated environment for protein identification and characterization by top-down mass spectrometry. *Nucleic Acids Res.* **32**, W340–W345 (2004)
- Zamdborg, L., LeDuc, R.D., Glowacz, K.J., Kim, Y.-B., Viswanathan, V., Spaulding, I.T., Early, B.P., Bluhm, E.J., Babai, S., Kelleher, N.L.: ProSight PTM 2.0: improved protein identification and characterization for top down mass spectrometry. *Nucleic Acids Res.* **35**, W701–W706 (2007)
- Ahlf, D.R., Compton, P.D., Tran, J.C., Early, B.P., Thomas, P.M., Kelleher, N.L.: Evaluation of the compact high-field orbitrap for top-down proteomics of human cells. *J. Proteome. Res.* **11**, 4308–4314 (2012)
- Edwards, R.L., Griffiths, P., Bunch, J., Cooper, H.J.: Top-down proteomics and direct surface sampling of neonatal dried blood spots: diagnosis of

- unknown hemoglobin variants. *J. Am. Soc. Mass Spectrom.* **23**, 1921–1930 (2012)
34. Edwards, R.L., Creese, A.J., Baumert, M., Griffiths, P., Bunch, J., Cooper, H.J.: Hemoglobin variant analysis via direct surface sampling of dried blood spots coupled with high-resolution mass spectrometry. *Anal. Chem.* **83**, 2265–2270 (2011)
35. Li, W., Ji, L., Goya, J., Tan, G., Wysocki, V.H.: SQID: an intensity-incorporated protein identification algorithm for tandem mass spectrometry. *J. Proteome. Res.* **10**, 1593–1602 (2011)
36. Dereeper, A., Guignon, V., Blanc, G., Audic, S., Buffet, S., Chevenet, F., Dufayard, J.-F., Guindon, S., Lefort, V., Lescot, M., Claverie, J.-M., Gascuel, O.: Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res.* **36**, W465–W469 (2008)
37. Dereeper, A., Audic, S., Claverie, J.-M., Blanc, G.: BLAST-EXPLORER helps you building datasets for phylogenetic analysis. *BioMed. Central* **10** (2010)
38. Zhang, H., Wu, C.X., Chamba, Y., Ling, Y.: Blood characteristics for high altitude adaptation in tibetan chickens. *Poultry Sci.* **86**, 1384–1389 (2007)
39. Li, J., Su, Z., Ma, Z.-Q., Slebos, R.J.C., Halvey, P., Tabb, D.L., Liebler, D.C., Pao, W., Zhang, B.: A bioinformatics workflow for variant peptide detection in shotgun proteomics. *Mol. Cell. Proteom.* **10**, M110.006536.1-11 (2011)
40. Nesvizhskii, A.I.: Proteogenomics: concepts, applications, and computational strategies. *Nat. Methods* **11**, 1114–1125 (2014)
41. Khatun, J., Yu, Y., Wrobel, J., Risk, B., Gunawardena, H., Secret, A., Spitzer, W., Xie, L., Wang, L., Chen, X., Giddings, M.: Whole human genome proteogenomic mapping for ENCODE cell line data: identifying protein-coding regions. *BMC Genomics* **14**, 141 (2013)
42. Wang, X., Zhang, B.: customProDB: an R package to generate customized protein databases from RNA-Seq data for proteomics search. *Bioinformatics* **29**, 3235–3237 (2013)
43. Wang, X., Slebos, R.J.C., Wang, D., Halvey, P.J., Tabb, D.L., Liebler, D.C., Zhang, B.: Protein identification using customized protein sequence databases derived from RNA-Seq Data. *J. Proteome. Res.* **11**, 1009–1017 (2011)
44. Park, H., Bae, J., Kim, H., Kim, S., Kim, H., Mun, D.-G., Joh, Y., Lee, W., Chae, S., Lee, S., Kim, H.K., Hwang, D., Lee, S.-W., Paek, E.: Compact variant-rich customized sequence database and a fast and sensitive database search for efficient proteogenomic analyses. *Proteomics* **14**, 2742–2749 (2014)
45. Fellers, R.T., Greer, J.B., Early, B.P., Yu, X., LeDuc, R.D., Kelleher, N.L., Thomas, P.M.: ProSight Lite: Graphical software to analyze top-down mass spectrometry data. *Proteomics* **15**, 1235–1238 (2015)
46. Jetz, W., Thomas, G.H., Joy, J.B., Hartmann, K., Mooers, A.O.: The global diversity of birds in space and time. *Nature* **491**, 444–448 (2012)