Bioinformatics, 31, 2015, i106–i115
doi: 10.1093/bioinformatics/btv236
ISMB/ECCB 2015

OXFORD

MSProGene: integrative proteogenomics beyond six-frames and single nucleotide polymorphisms

Franziska Zickmann and Bernhard Y. Renard*

Research Group Bioinformatics (NG4), Robert Koch Institute, 13353 Berlin, Germany

*To whom correspondence should be addressed.

Abstract

Summary: Ongoing advances in high-throughput technologies have facilitated accurate proteomic measurements and provide a wealth of information on genomic and transcript level. In proteogenomics, this multi-omics data is combined to analyze unannotated organisms and to allow more accurate sample-specific predictions. Existing analysis methods still mainly depend on six-frame translations or reference protein databases that are extended by transcriptomic information or known single nucleotide polymorphisms (SNPs). However, six-frames introduce an artificial sixfold increase of the target database and SNP integration requires a suitable database summarizing results from previous experiments. We overcome these limitations by introducing MSProGene, a new method for integrative proteogenomic analysis based on customized RNA-Seq driven transcript databases. MSProGene is independent from existing reference databases or annotated SNPs and avoids large six-frame translated databases by constructing sample-specific transcripts. In addition, it creates a network combining RNA-Seq and peptide information that is optimized by a maximum-flow algorithm. It thereby also allows resolving the ambiguity of shared peptides for protein inference. We applied MSProGene on three datasets and show that it facilitates a database-independent reliable yet accurate prediction on gene and protein level and additionally identifies novel genes.

Availability and implementation: MSProGene is written in Java and Python. It is open source and available at http://sourceforge.net/projects/msprogene/.

Contact: renardb@rki.de

1 Introduction

High-throughput technologies in both genomics and proteomics have driven the development of a variety of methods to analyze the large amounts of data generated. RNA-Seq techniques measure the transcriptome (Wang *et al.*, 2009), while mass spectrometry allows identification and quantification of proteins that were expressed (Nilsson *et al.*, 2010). The field of proteogenomics combines this multi-omics data for more accurate and sample-specific analyses (Castellana and Bafna, 2010; Nesvizhskii, 2014).

In recent years, proteogenomic studies have become more and more popular, focusing on deeper understanding of model organisms or exploring currently unannotated genomes (Ahn *et al.*, 2013; Castellana *et al.*, 2008; Fanayan *et al.*, 2013; Kelkar *et al.*, 2014). Despite this popularity, methods that are jointly focusing on genomics, transcriptomics and proteomics so far mainly rely on six-frame translations (Kelkar *et al.*, 2011; Krug *et al.*, 2013) or extensions of existing reference protein databases (Ahn *et al.*, 2013; Li *et al.*, 2010). Six-frame translation has the advantage of being independent from any a priori annotation of the nucleotide sequence. However, it introduces an artificial sixfold increase of the (unknown) target database, which can result in a bias in peptide identification (Blakeley *et al.*, 2012; Branca *et al.*, 2014; Jeong *et al.*, 2012; Reiter *et al.*, 2009).

In contrast, reference protein databases, for instance extended by known single nucleotide polymorphisms (SNPs) from databases such as dbSNP (Sherry *et al.*, 2001), are not as prone to this bias as six-frame translations. But these approaches depend on existing annotations and thus cannot be applied to unannotated organisms without reference proteomes. Further, they might not contain all information necessary to identify mutated or novel genes, and even error-tolerant search approaches (Renard *et al.*, 2012) may not be sufficient to recover these unannotated genes.

Thus, recent studies also rely on transcriptome information to provide better suited databases (Krug *et al.*, 2014; Ning and

Nesvizhskii, 2010; Safavi-Hemami *et al.*, 2014; Wang and Zhang, 2014). They focus on a more specific choice of six-frame translated open reading frames and on enhancing databases in a data-driven fashion, for instance by only including spliced parts or variations to the database (Wang and Zhang, 2013; Wang *et al.*, 2011; Woo *et al.*, 2013).

These approaches are either only suitable for eukaryotes (having splicing events) or are still only seen as an addition to or refinement of the standard approach using protein databases to identify peptides. Other approaches rely on the de novo assembly of transcript sequences, which are then six-frame translated to provide a sample-specific database (Evans *et al.*, 2012; Mohien *et al.*, 2013).

Further, all of these efforts are targeted on improving peptide identification, but rely on standard approaches to perform protein inference. Because of shared peptides that are present in more than one protein, often parsimonious approaches are employed that group proteins instead of selecting one specific match per peptide (Claassen, 2012; Huang *et al.*, 2012; Serang *et al.*, 2010). However, a possibility to select the most likely protein per peptide is desirable. Here, RNA-Seq is a valuable source to assist protein inference, as it provides an additional layer of confidence for a specific protein.

In this article, we present MSProGene (*M*ass *S*pectrometry and RNA-Seq based *Pro*tein and *Gene* Identification) as an integrative proteogenomic method that goes beyond the extension of existing reference databases by constructing customized transcript databases based on RNA-Seq. This sample-specific database avoids unnecessary enlargement by six-frame translations and increases the confidence in identified proteins. Further, RNA-Seq information is used to approach shared peptide protein inference without the need for protein grouping. To do so, MSProGene represents transcriptomic and peptide evidence in a network and performs a maximum-flow optimization formulated as an integer linear program.

We applied MSProGene on a *Bartonella henselae* and a *Litomosoides sigmodontis* dataset where it shows reliable and accurate identifications. Further, in a simulation based on *Escherichia coli* we demonstrate the suitability of the network optimization and RNA-Seq integration to resolve shared peptides for protein inference.

2 Methods

Figure 1 shows the overall workflow of MSProGene: First, an RNA-Seq read mapping is analyzed to infer transcript sequences, which are updated by including variations present in the RNA-Seq reads (Fig. 1.1.). These sequences are translated to amino acid sequences to serve as a database for a peptide search of tandem mass spectra (Fig. 1.2.). The resulting set of peptide spectrum matches (PSMs) is represented by a network. MSProGene then performs protein inference by re-assigning shared peptides using a linear program approach based on RNA-Seq information (Fig. 1.3.). Finally, peptide identifications are controlled with regard to their false discovery rate (FDR) and transcripts with a sufficient number of peptide hits are reported (Fig. 1.4.).

2.1 Transcript database

MSProGene uses evidence from RNA-Seq reads to derive a customized transcript database for the spectra search. This database reflects sample-specific mutations present in the reads and is independent from any a priori knowledge, in particular it is independent from known annotations or protein sequences. Per default, the gene finder GIIRA (Zickmann *et al.*, 2014) is used to extract transcripts based on a mapping of the RNA-Seq reads. However, also other methods for gene and transcript prediction can be used, for instance Cufflinks (Trapnell *et al.*, 2010).

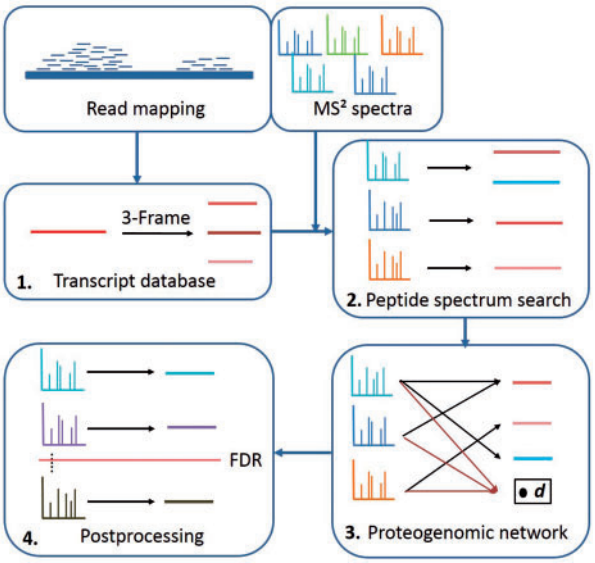

Fig. 1. The overall workflow of MSProGene. (1) An RNA-Seq read mapping is analyzed to infer transcript sequences, which (2) provide the database for spectra search. (3) The resulting PSMs are represented by a network, which is analyzed to resolve protein inference and to select the correct frame per transcript. (4) Finally, peptide identifications are controlled with regard to their FDR

MSProGene analyzes the read mapping and refines the transcript sequence according to mutations present in the RNA-Seq reads. A variation (SNP or insertion or deletion) is integrated if (i) it is present in more than one read (this ensures that regions with low coverage are not biased towards more mutations) and (ii) it is supported by the majority of the reads. Note that the first condition is only a default threshold specified to reduce bias introduced by low coverage. This threshold can be changed by the user. Further, also a vcf file with previously called mutations by external tools can be provided. These mutations are directly integrated to the reference sequence and are thus respected in the transcript reconstruction.

In case a specific database is intended for the peptide spectrum search, MSProGene can also be provided with custom sequences in fasta format, without the need for RNA-Seq evidence. Also gene models based on evidence different from RNA-Seq or the combined results of varying prediction methods [for instance combinations by IPred (Zickmann and Renard, 2015), or EVidenceModeler (Haas *et al.*, 2008)] are accepted as input for MSProGene. Note that in this case mutations already need to be included in the sequences, and the sequence header must contain information on the strand and start and stop position of the gene (an example file is provided with the MSProGene installation).

To be suitable for spectra search, nucleotide sequences need to be translated into amino acid sequences. Initially, we rely on a three-frame translation since in RNA-Seq experiments the ends of genes are often not recovered with high precision. Hence, the predicted start codon might not be the correct one and translating only one frame would potentially lead to a loss in peptide identifications. However, (i) increasing the transcript database with a six-frame translation is only necessary if no strand information is available (as it for instance is the case for unspliced Cufflinks predictions). Thus, bias resulting from unnecessary extension of the database can be avoided. Further, (ii) in order to create a tailored transcript database without artificial increase we perform a second MSProGene iteration based on the analysis of the first spectra search.

Note that only one out of the initial three frames is correct; hence, the translated protein sequence of the incorrect frames might contain stop codons. Since an early stop codon can also be due to an incorrectly inserted mutation, MSProGene does not stop the entire translation in case of a stop codon but can extract several amino acid subsequences per transcript frame. Since the user can specify a minimum peptide length for spectra search (per default 5 amino acids), subsequences with smaller length are removed.

Finally, each transcript t with sequence length l^t is initially scored based on the original GIIRA gene score (or score from other prediction methods) s^g and its read coverage c^t. The coverage is calculated by taking the number of reads n^t mapping to the transcript and their corresponding length l^r into account:

$$c^t = \frac{n^t \cdot l^r}{l^t}. \quad (1)$$

The initial transcript score s^t is normalized over minimum (m^i) and maximum (m^a) score of all original gene scores to indicate the relative evidence for a transcript in comparison to other transcripts:

$$s^t = s^g \cdot \frac{c^t}{m^a - m^i + 1}. \quad (2)$$

2.2 Peptide spectrum search

Once the transcript database has been created, the input tandem mass spectra are searched against the resulting set of amino acid sequences. Per default, MSProGene uses MSGF+ (Kim and Pevzner, 2014) as the search engine, but can easily be adapted to also work with other search methods. After the search, the resulting PSMs are extracted by MSProGene, independent of whether they are unique peptides or shared peptides (i.e. one peptide mapping to more than one transcript sequence). Further, the PSM score provided by the search engine is extracted, and normalized to the [0,1] interval.

2.3 Proteogenomic network

After the spectra search, each identified spectrum is assigned to one peptide sequence that can be found in one or more transcript sequences. Since each spectrum can only arise from one peptide and one transcript, we (i) need to assign shared peptides to their most likely origin. An additional challenge is the presence of potentially multiple supported reading frames per transcript. Since we initially provide at least three frames (*sister frames*) per transcript, a peptide can independently be mapped to each of the frames, although only one of the frames can be correct. Hence, (ii) we also have to identify the correct frame for each transcript and erase all incorrectly mapped peptides. Furthermore, not necessarily all PSMs are correct. Thus, (iii) we have to detect and remove incorrect identifications.

To meet these three objectives we first represent the inference problem as a network, which is then optimized in order to solve the inference. The network $G = \{N, E\}$ (depicted in Fig. 2) with edge set E and node set $N = P \cap F \cap T \cap so \cap si \cap d$ has nodes $p_i \in P$ representing the individual peptides and nodes $f_j \in F$ representing the sister frames of each transcript. Further, also the transcript itself is included as a node $t_k \in T$. For technical purposes, also a source node *so* and a sink node *si* are integrated to the network, as well as a *dummy* node *d*.

For each match between peptide p_i and frame f_j, a directed edge $e_{p_i,f_j} \in E$ is integrated to G connecting the two nodes. Further, all sister frames are connected to their corresponding transcript. Note that each peptide node is not only connected to its mapped frames but also to the dummy node. This ensures that whenever no target frame remains possible for a peptide, this peptide can be assigned to the dummy without creating inconsistency. The set of connections of a peptide p_i can become infeasible in case p_i only maps to frames that were marked as incorrect because their competing sister frames have more support. In this case, p_i is likely to be an incorrect identification, which is indicated by assigning p_i to d. For an example refer to Figure 2: here p_2 and p_3 match to different frames of the same transcript; hence, only one match can be correct, and the other peptide is assigned to d.

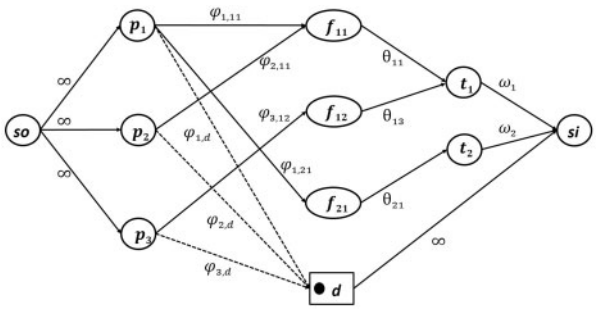

Fig. 2. Simplified example of a proteogenomic network: peptide nodes p_i are connected to the frames f_j they map to, and all sister frames are connected to their corresponding transcript node t_k. A so called *dummy* node *d* ensures that incorrect peptide identifications can be reassigned. All edges are labeled according to their capacity indicating the support from experimental data for a connection between the two neighboring nodes. The capacities define the overall throughput that can be passed through the network, starting from source node *so* towards the sink *si*

Since we aim at choosing connections between nodes that reflect the most likely correct identification, each edge is assigned a capacity representing the reliability of the associated match. Edges starting from the source are connected to peptide nodes and have an unlimited capacity, whereas edges e_{p_i,f_j} connecting peptides to frames have a capacity φ_{p_i,f_j} that is initially determined by the score calculated by the peptide search engine. Further, it is restricted by a binary variable $y_{p_i,f_j} \in \{0,1\}$ indicating whether this connection is chosen as the most likely connection ($y_{p_i,f_j} = 1$) or not ($y_{p_i,f_j} = 0$):

$$0 \leq \varphi_{p_i,f_j} \leq y_{p_i,f_j} \ \forall e_{p_i,f_j} \in E. \quad (3)$$

Further, edges $e_{t_k,si} \in E$ connecting transcript nodes $t_k \in T$ to *si* have a capacity ω_k that is determined by the initial transcript score calculated in step 1 (Equation 2) of the overall workflow. The capacity θ_{f_j,t_k} of connections of sister frames to their transcript is initially set to this transcript score, weighted by the number of peptides originally associated to the frame.

Since only one of the sister frames can be correct, θ_{f_j,t_k} is also restricted by a binary variable $m_{f_j,t_k} \in \{0,1\}$ that indicates whether a frame is chosen or not:

$$0 \leq \theta_{f_j,t_k} \leq m_{f_j,t_k} \forall e_{f_j,t_k} \in E. \quad (4)$$

Two additional constraints ensure that only one match per peptide (Equation 5) and only one frame per transcript (Equation 6) is selected, respectively:

$$\sum_j y_{p_i,f_j} = 1 \forall i \,|\, p_i \in P, \quad (5)$$

$$\sum_j m_{f_j,t_k} = 1 \forall k \,|\, t_k \in T. \quad (6)$$

The capacities define the maximal throughput that is allowed to be passed through an edge. Given these capacities, we can formulate a maximum-flow problem in order to optimize the throughput—in

this case the reliability of connections—that is passed from source towards sink node:

$$\max \sum_{e_{p_i,f_j} \in E} \varphi_{p_i,f_j} + \sum_{e_{f_j,t_k} \in E} \theta_{f_j,t_k} + \sum_{e_{t_k,si} \in E} \omega_k + \sum_{e_{p_i,d} \in E} \lambda_{p_i,d}\, y_{p_i,d}, \tag{7}$$

where $\lambda_{p_i,d}$ corresponds to a penalty term similar to a Lagrange multiplier for connections to the dummy node: In the maximum-flow description above, all capacities of chosen edges add to the overall maximal flow. However, an important difference holds for the dummy node d: since assignments to d are required for peptides that are likely incorrect identifications, a chosen connection to the dummy results in a penalty on the overall flow. This is realized by a form of Lagrangian relaxation on constraints describing edges to the dummy node. Whenever such a connection is chosen (i.e. $y_{p_i,d} = 1$), a penalty λ (i.e. the Lagrange multiplier) that equals the confidence score of the PSM is applied to the overall objective.

Although nodes have an unlimited throughput, a requirement of the maximum-flow is that for each node the input has to equal the output flow. Hence, the number of peptides that can be assigned to each frame and transcript is restricted by the overall evidence for this transcript because the higher ω_k, the more flow can be assigned to the transcript. Given the capacities $\theta_{f_j,t_k} \leq \omega_k$ of the connections of sister frames to their corresponding transcript, we derive the following constraint:

$$\sum_{i \mid e_{p_i,f_j} \in E} \varphi_{p_i,f_j} \leq \theta_{f_j,t_k} \forall e_{f_j,t_k} \in E. \tag{8}$$

Note that the dummy node has an unlimited outgoing capacity, such that in theory an unlimited number of peptides can be assigned to d. However, due to the penalty this connection is only chosen if the penalty is balanced by the benefit of supporting the competing frames.

Finally, the described maximum-flow problem is formulated as an integer linear program, which can be solved for instance using the CPLEX Optimizing studio (CPLEX, 2011). As a result, each peptide is either indicated as an incorrect match or associated to the most likely transcript frame. Note that thereby the graphical model ensures that the reassignment is performed in a non-greedy fashion that for instance distributes peptides between multiple observed alternative isoforms, rather than selecting only one isoform.

2.4 Postprocessing

After all PSMs have been reassigned to their most likely frame or are indicated as likely incorrect predictions, the confidence in each transcript sequence and corresponding frame has to be recalculated.

MSProGene proceeds through the original transcripts and assigns the frame chosen in the linear program. Note that at this point MSProGene uses the sequences supported by the spectra search for a second iteration: The supported frames are used to create a second and more specific amino acid database for a second run with a peptide spectrum search engine. The initial database was artificially increased by the three-frame translation, whereas the updated database is tailored to the (unknown) true database. Also the second PSM results are represented in a network to resolve shared peptides and identify incorrectly mapped peptides (refer to former section). Afterwards, the transcripts are finally analyzed for their peptide support and FDR controlled.

Since decoy protein sequences which are classically used for FDR computation in proteomics are artificial sequences without RNA-Seq evidence, the network representation and maximum-flow optimization is not applicable to decoy identifications. Hence, only target peptide hits are reassigned in the maximum-flow and can thus be used for FDR calculation. Therefore, the FDR cannot be calculated by a standard target-decoy approach, but is determined in a decoy-free approach based on the expectation-maximization (EM) algorithm (Dempster *et al.*, 1977). The aim is to fit two distributions on the frequencies of overall scores, one that explains the correct (i.e. target) and one the incorrect (i.e. decoy) identifications, similar to the approach in (Renard *et al.*, 2010). The observed frequencies of scores should be a mixture of these two distributions, where we assume an underlying normal distribution for both target and decoy identifications (assumption confirmed in independent experiments, data not shown).

Note that since the EM is not guaranteed to find the global maximum, the search is performed several times with differing initial values to identify the model best fitting the data. With the resulting target N_T and decoy N_D distribution we can compute a false discovery rate FDR_i at each PSM p_i with score s_i^p, using the cumulative density functions $F_T(s_x^p)$ for N_T and $F_D(s_x^p)$ for N_D:

$$FDR_i = \frac{w_D \cdot (1 - F_D(s_i^p))}{w_T \cdot (1 - F_T(s_i^p)) + w_D \cdot (1 - F_D(s_i^p))}, \tag{9}$$

where w_T and w_D are the weights of the target and decoy distribution, respectively.

2.5 Output

After the reassigned peptides are FDR controlled and hits below the threshold are removed, the set of transcripts with spectra support is reported. For postprocessing and visualization, the coordinates as well as confidence score and number of spectrum matches are presented in the well-established GTF format, accompanied by the actual sequences in fasta format.

The final confidence score s^c combines the original transcript score s^t with its coverage and quality of PSMs (set denoted as P^t):

$$s^c = s^t \cdot \frac{1}{l^t} \cdot \sum_{i \mid p_i \in P^t} s_i^p \cdot l_i^p, \tag{10}$$

where l^t is the length of the transcript sequence and l_i^p is the length of a peptide $p_i \in P^t$ with score s_i^p.

Since the combination of RNA-Seq read support and tandem mass spectra support does not only increase the confidence in protein identifications, but can also be used to verify variation observed in the read mapping, MSProGene additionally outputs a VCF file. This file contains all mutations present in the transcripts compared to the given reference sequence. Further, we indicate whether mutations are also supported by spectra (as an additional layer of confidence).

3 Experimental setup

3.1 Algorithm evaluation

As a proof-of-principle evaluation of the algorithm for peptide reassignment we conducted a simulation experiment. We used the NCBI reference annotation of *Escherichia coli* (NCBI accession: NC_000913.3) and integrated SNPs simulated with a mutation-rate of 1% to the gene sequences (to simulate deviances from the reference sequence as occurring in real datasets). Based on the mutated sequences, we simulated Illumina RNA-Seq reads with the read simulator Mason (Holtgrewe, 2010) in varying expression levels.

Tandem mass spectra were generated with the spectra simulator MSSimulator (Bielow *et al.*, 2011; OpenMS Release1.11) with a gradient of 3000s, an instrument resolution of 200 000, 10 tandem mass spectra per retention time bin, and default settings otherwise. Each of the resulting spectra is linked to its original peptide and protein, such that we can compare the peptide assignments of the network optimization integrated in MSProGene against the ground truth peptides.

3.2 *Bartonella henseale*

MSProGene was tested on data of *B.henselae*, a pathogenic bacterium that causes infections such as the cat scratch disease (Omasits *et al.*, 2013). Tandem mass spectra and RNA-Seq reads originate from a study by Omasits *et al.* (2013) (GEO Series accession number: GSE44564). We pooled data from the two conditions (induced and uninduced) of replicate 1 resulting in 1.16 million tandem mass spectra and 211 million AB-Solid RNA-Seq reads. Reads were mapped to the *B.henselae* reference genome (strain Houston-1, NCBI accession: NC_005956) using BFAST (Homer *et al.*, 2009; version: 0.7.0 a). For settings we followed the mapping pipeline and parameters recommended in the BFAST manual. As in the original study, the resulting mapping was filtered using samtools (Li *et al.*, 2009) to remove contamination with rRNA. Further, all raw spectra were converted to MGF format using the Trans-Proteomic Pipeline (Deutsch *et al.*, 2010). MSProGene was applied with default settings, using GIIRA in prokaryote mode for construction of the transcript database, also with default settings.

To analyze the performance of reference-independent methods, we compared MSProGene to the approach by (Evans *et al.*, 2012; in the following called *Assembly*) based on de novo assembly with Trinity (Grabherr *et al.*, 2011), as well as a standard six-frame translation of the *B.henselae* genome (in the following denoted as *six-frame*). Assembly was applied with default settings in its genome-guided mode (using the BFAST mapping as a guide). The resulting assembly contained 1907 transcripts, which were six-frame translated to identify open reading frames. These frames served as the database for MSGF+ search. Six-frame translation was performed using the program getorf from the EMBOSS package (Rice *et al.*, 2000; version EMBOSS:6.4.0.0), requiring a minimum length of 200bp. These three reference-independent methods were analyzed regarding the overall number of identified proteins and the spectra coverage of identifications.

For a general analysis of the robustness of our method we also randomly divided the original set of 1.16 million spectra into two smaller sets, each including half of the spectra. The compared methods were applied using the smaller samples of spectra separately and the resulting predicted protein sequences were compared between runs. The higher the overlap between two runs on differing input samples, the more robust the method. As a measure of overlap we counted the number of proteins coinciding in both runs and divided it by the highest number of proteins predicted in one run.

Further, we compared our method to a standard database search (in the following denoted as *Standard*) on the 1488 annotated *B.henselae* proteins available at NCBI (http://www.ncbi.nlm.nih.gov/). In addition, we performed a standard search on a database including SNPs indicated by a samtools (Li *et al.*, 2009) mpileup variant call on the RNA-Seq mapping (in the following denoted as *Mutated*).

For all evaluations we chose the set of annotated *B.henselae* proteins as a ground truth reference protein set (note that not necessarily all of these proteins are actually expressed simultaneously). The output of the Standard and Mutated approach was directly compared to the reference. In contrast, for the reference-free methods we first compared the coordinates of predicted proteins to the reference coordinates in order to map predictions to reference proteins.

For evaluation of method quality we employed the metrics of recall and precision. Recall is calculated as the number of identified annotated proteins, divided by the total number of annotations (1488). Precision is calculated as the number of predicted proteins matching the annotation, divided by the total number of proteins predicted by the method. Note that by nature of the analysis, the Standard and Mutated method always have a precision of 100% because they are exclusively searched against the reference annotation.

We also calculated an annotation-based FDR on the protein identifications of reference-free methods, sorted by identification score. We regard an identified protein as incorrect in case it did not match the reference annotation. We note that since not necessarily all unmatched predictions are false positives, this is a conservative estimate that likely overestimates the actual rate of incorrect identification.

3.3 *Litomosoides sigmodontis*

We also compared MSProGene to a six-frame based analysis on a *L.sigmodontis* dataset (assembly nLS.2.1 from www.nematodes.org). *Litomosoides sigmodontis* is a popular model organism for filarial nematodes, that amongst other diseases cause lymphatic filariasis ('elephantiasis') and are the human-parasitic species with the highest overall impact on public health (Armstrong *et al.*, 2014).

Tandem mass spectra originate from a study by Armstrong *et al.* (2014) (PRIDE Project PXD000756, in total 856 380 spectra).

For this organism only very few proteins are already annotated (a search at NCBI on January 9, 2015 resulted in 75 protein sequences). Hence, here we only compare methods in regard to their overall identification confidence, the number of predicted proteins and their spectra coverage.

Transcript prediction methods such as Cufflinks (Trapnell *et al.*, 2010) and GIIRA work best on high coverage RNA-Seq datasets. Hence, since at the time of this study only low coverage 454 transcriptome data was available for *L.sigmodontis*, we chose RNA-Seq data from *Brugia malayi*, a close relative of *L.sigmodontis*. We pooled 14 samples from different life cycle stages of *B.malayi* (BioProject-accession: PRJEB2709) and mapped the reads to the *L.sigmodontis* draft genome using TopHat2 (Kim *et al.*, 2013; version 2.0.11) with error tolerant parameter setting (N 5, read-gap-length 5, read-edit-dist 5). Transcript coordinates were obtained using Cufflinks (version 2.2.0) on the resulting mapping. The resulting GTF file was converted using in-house scripts to generate a fasta file with transcript sequences for MSProGene analysis. For the six-frame analysis the *L.sigmodontis* draft genome was translated using the program getorf from the EMBOSS package, requiring a minimum length of 200 bp.

In addition to the transcripts predicted by either Cufflinks or getorf, we included protein sequences from the *Wolbachia* symbiont of *L.sigmodontis*, obtained from www.nematodes.org (release wLs 2.0, 1042 sequences) for spectra search.

For further evaluation, we used BLAST (Altschul *et al.*, 1997) to compare the identified sequences to *B.malayi* proteins. Similar to (Armstrong *et al.*, 2014), we specified a bit score cutoff of 50. We did not use the BLAST E value for threshold definition to allow a fair comparison since an E value threshold may have favored the

evaluation towards MSProGene because it has a smaller query database size than the six-frame translation.

3.4 Peptide search parameters

All spectra searches were conducted using MSGF+ (Kim and Pevzner, 2014; version v9881) with a precursor mass tolerance of 5 ppm, a minimum peptide length of five amino acids, specifying a high-resolution LTQ, and using default settings otherwise. All analyses were performed in regard to a 1% FDR cutoff and excluding proteins with fewer than two spectra hits.

4 Results and discussion

4.1 Algorithm evaluation

We analyzed the PSMs before and after the network optimization of MSProGene. Details are shown in Figure 3. Of 21 715 spectra that MSGF+ matched to the original protein (sometimes among multiple proteins), 21 617 were assigned correctly (99.5%) by MSProGene. Overall, the algorithm correctly reassigned over 90% of the spectra that had multiple protein hits (933 of 1031).

This demonstrates that the network representation with integrated RNA-Seq information and its optimization is suitable to successfully resolve shared peptide protein inference, without the need for protein grouping.

4.2 *Bartonella henselae* data

4.2.1 Verification of transcripts with spectra support

First, we investigated the effect of integrating RNA-Seq evidence and spectra on the actual identification accuracy. As shown in Table 1, the transcript database constructed for spectra search contains 1568 sequences. This number is reduced to 1397 when taking spectra support into account. This leads to a decrease in recall from 78.2 to 76.5%. This shows that although generally transcriptome and proteome correlate well, we have to be aware of potential losses in protein identifications. For comprehensive studies on mRNA and protein level correlation we refer to Vogel and Marcotte (2012) and Nagaraj *et al.* (2011). Here, it is shown that the mRNA undergoes several modification steps that can reduce the correlation to the protein level. However, overall the transcriptome is regarded as a valuable source and verification technique for protein level analysis and evidence on the transcriptome level is a good indication for protein measurements. However, if possible, combinations with other searches should be considered in order to detect additional protein candidates.

In contrast to sensitivity, the precision strongly increases from 79.0% to 85.1% when spectra support is taken into account. This shows that the combination of RNA-Seq data and tandem mass spectra is a suitable verification method for accurate protein identification.

4.2.2 Comparison to reference-free methods

For the three methods compared we counted the number of annotations that were identified and the number of predictions that actually match the annotation. Both numbers can differ since a single annotated protein might be covered by several smaller predictions. The results of the analysis are summarized in Table 2. The transcript database constructed for spectra search by MSProGene contains 1568 sequences. This is significantly smaller than the number of sequences searched in the six-frame analysis and Assembly and shows the suitability of RNA-Seq data to provide smaller and more tailored search databases.

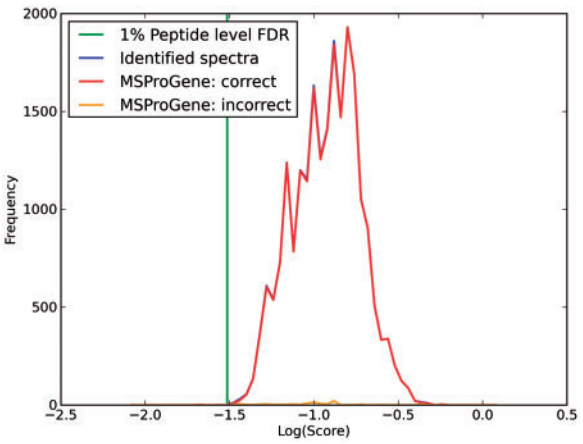

Fig. 3. Figure illustrating the distribution of peptides correctly and incorrectly reassigned by MSProGene. 99.5% of the peptides were assigned to their original ground truth protein

Table 1. Prediction results of MSProGene, exclusively based on RNA-Seq, verified by spectra support, and in addition excluding proteins with only one spectrum hit

	Without spectra	With spectra	With spectra Without single hits
Predicted	1568	1397	1286
#matches to annotation	1238	1189	1143
#identified annotations	**1164**	1139	1109
Recall (%)	**78.2**	76.5	74.5
Precision (%)	79.0	85.1	**88.9**

Evaluation on the *B.henselae* dataset, compared to the reference annotation comprising 1488 genes. Best values for each category are marked in bold.

Table 2. Prediction of reference-free methods on the *B.henselae* dataset, compared to the reference annotation comprising 1488 genes

	MSProGene	Six-frame	Assembly
Database size	1568	6091	5894
Predicted	1286	1502	1276
# matches to annotation	1143	1207	447
# identified annotations	1109	**1163**	372
Recall (%)	74.5	**78.2**	25.0
Precision (%)	**88.9**	80.4	35.0
Recall 1%-AnnotationFDR (%)	**51.5**	1.1	0.0
Median # spectra per protein	**90**	77	50

The row indicated as '1%-AnnotationFDR' shows results for an additional 1% annotation-based FDR on the protein level. The best value for each category is marked in bold.

Overall the six-frame approach predicts the highest number of spectra-supported genes and also achieves the highest recall given the peptide level FDR. At first this is surprising given the supposedly high number of spurious sequences in six-frame translated databases (which should lead to reduced sensitivity). We suspect that the overall high coverage of this dataset (1.16 million spectra) prevents the originally expected loss in protein identifications. A loss in sensitivity is rather reflected in the spectra coverage of protein identifications, where six-frame shows 77 median spectra hits per protein, compared to 90 for MSProGene. The drawback of six-frame

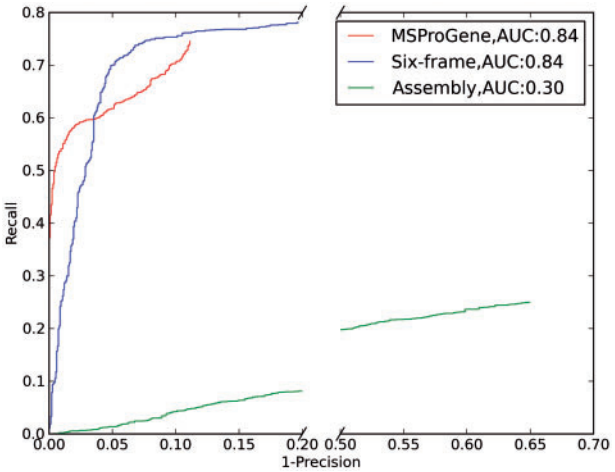

Fig. 4. Receiver operating curve illustrating recall and precision of MSProGene, six-frame and Assembly for the *B.henselae* dataset. MSProGene shows the highest precision of all three methods. In particular, for highly scored predictions it achieves better sensitivity at the same precision level

translations is also reflected on the precision level, where the high recall comes at the cost of specificity: six-frame has 3.7% higher recall but 8.5% less precision than MSProGene (also refer to Fig. 4). Hence, although MSProGene identifies slightly fewer proteins, it provides higher confidence in the resulting predictions. Further, if in addition to the peptide level FDR also an annotation-based FDR is applied on the protein level, the recall of six-frame decreases to 1% because of early false positive identifications. In contrast, MSProGene still achieves a recall of 51%.

The Assembly approach shows low agreement between predicted transcripts and the annotation, resulting in reduced precision and recall. This indicates that the two-step integration of RNA-Seq data (first de novo assembly followed by six-frame translation and later the independent spectra search) is not as suited for proteogenomic analysis as the integrative approach employed by MSProGene.

As illustrated in Figure 5, MSProGene and the six-frame approach coincide in 941 of the 1488 annotations. In contrast, Assembly only shared 304 and 317 annotations with MSProGene and six-frame, respectively.

Taken together, the three methods identified 1340 of the 1488 annotated *B.henselae* proteins. However, all methods identified proteins that were not predicted by the other methods, such that no approach shows a complete prediction by itself. Six-frame is sensitive, but lacks confidence and precision. MSProGene is specific but dependent on the quality of predicted transcript sequences. Here, gene identifications exclusively based on RNA-Seq (as performed by GIIRA for this dataset) might not identify all possible transcripts and a more comprehensive RNA-Seq based prediction might be more sensitive. Some of the missing transcripts can be recovered by the de novo assembly used in Assembly; however, this approach overall has the least accuracy. Hence, in regard to precision, customized transcript databases as employed by MSProGene should be preferred.

All three methods performed well in the robustness analysis. The overlap of six-frame (97.5%) is slightly higher than for MSProGene (96.0%) and Assembly (95.5%). However, all three approaches only vary little, indicating that they are robust to differing input data.

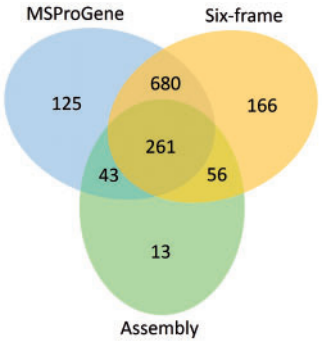

Fig. 5. Venn diagram illustrating the number of identified annotated proteins of the *B. henselae* dataset for MSProGene, six-frame, and Assembly. Together, 1340 of the annotated proteins were identified, although no method shows a complete prediction by itself

4.2.3 Comparison to reference-based methods

To generate the mutated database 2592 variants were called with samtools on the RNA-Seq read mapping and included in the reference protein sequences. Both Standard and Mutated method identified 1274 of the annotated proteins (recall: 85.6%). Interestingly, including mutations observed in the RNA-Seq mapping did not improve the overall recall, but instead even decreases the median spectra support for identified proteins from 106 (Standard) to 95 (Mutated) spectra. This indicates that some of the included SNPs are incorrect. Since thresholds for the filtering of incorrect mutations are hard to define (Giese *et al.*, 2014), this is a likely bias when including sample-specific mutations to reference proteins.

With 1109 identified proteins, MSProGene has a lower recall than both Standard and Mutated method. However, as shown in Figure 6, it identifies 84 proteins not detected by the standard searches.

When comparing MSProGene and the Mutated approach, 92 proteins are unique to MSProGene, and 257 proteins are unique to Mutated. The latter are not identified due to missing or incorrectly constructed transcript sequences. MSProGene not only needs to correctly identify the correct PSMs for a protein sequence, but also the correct coordinates of a transcript. Hence, the sensitivity of MSProGene strongly depends on the quality of constructed transcript sequences. Since RNA-Seq is challenging as the exclusive source for gene prediction (performed by GIIRA for this dataset), integrating additional evidence or other methods for prediction might lead to a more comprehensive set of transcripts and hence improved recall. We believe that the extensive studies dedicated to RNA-Seq analysis (a search of the term 'RNA-Seq' on google scholar resulted in more than 17 300 entries in year 2014) will also benefit MSProGene. Since our method is independent of the method used for transcript construction (except scores and mutations that need to be provided), better methods for RNA-Seq based gene and transcript prediction will lead to improved recall by MSProGene.

The proteins exclusively detected by MSProGene often have shared peptide support and in addition are supported by peptides that have scores below the FDR threshold in the Mutated approach. For instance, 51 of the missing 92 proteins of Mutated can be identified with an FDR threshold of 5%. This indicates the precision of MSProGene peptide assignments since it identifies these proteins under a more conservative FDR.

In general, the comparison against the complete reference can only be regarded as a relative rather than an absolute comparison between methods (since not all genes are necessarily expressed at the same time). Further, transcripts that do not match the reference are

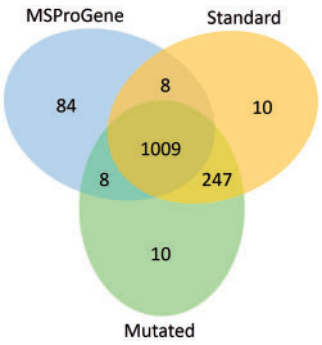

Fig. 6. Venn diagram illustrating the number of identified annotated proteins of the *B.henselae* dataset for MSProGene, Standard, and Mutated. Together, 1376 of the annotated proteins were identified, although no method shows a complete prediction by itself

not necessarily false positives but might be unannotated genes. However, for the evaluation of sensitivity and specificity all transcripts not matching the annotation are regardless counted as false positives. Hence, the evaluation is slightly biased against MSProGene.

Reference-dependent approaches fail to detect novel genes (examples detailed below) and in addition, even databases adapted or extended with SNPs are not always suited to identify mutated proteins. Hence, even for annotated organisms or fast evolving organisms such as viruses it is worth to employ alternative search strategies.

When compared to the annotated reference database comprising 1488 genes, MSProGene predicted 76 genes with RNA-Seq and spectra support that do not match the annotation. Two of these genes (located at position 1 357 979 to 1 358 722 and 1 180 052 to 1180672, respectively) were chosen for further verification with BLAST (Altschul *et al.*, 1997). The first protein with length 248 was supported by 94 spectra, the second one of length 207 received 36 spectra.

A protein BLAST search of the two sequences (predicted by MSProGene on the Houston-1 reference strain) revealed that both proteins are annotated in other *B.henselae* strains. The first sequence shows high similarity to a peptide ABC transporter substrate-binding protein (BLAST E value: 1 e−178, identity: 99%), for instance present in strain BM1374165. The second one shows high similarity to a hemin binding protein E, for instance present in strain BM1374163 (BLAST E value: 5 e−145, identity: 100%).

Thus, both genes are likely candidates for novel genes in the Houston-1 reference strain of the *B.henselae* taxonomy. This highlights the relevance of reference database independent approaches because standard database searches cannot identify genes that are not already annotated.

4.3 *Litomosoides sigmodontis* data

The results of the evaluation on the *L.sigmodontis* dataset are shown in Table 3. Also for this dataset the RNA-Seq based transcript database used by MSProGene is significantly smaller than the six-frame translation.

Although the overall number of predicted sequences is higher for the six-frame approach, MSProGene sequences receive higher spectra support. The greater confidence of MSProGene transcripts is also significantly shown in the BLAST search: As illustrated in Figure 7, the confidence of BLAST hits is considerably improved for MSProGene sequences. Further, only 42% of the six-frame

Table 3. Evaluation for *L.sigmodontis* dataset, with best values for each category marked in bold

	MSProGene	Six-frame
Database size	28 009	189 512
Predicted	2146	**4297**
Median spectra count	**8**	6
BLAST hits all	1462	**1804**
Median bit score all	**54.5**	25.8
BLAST hits above threshold	**779**	42
Median bit score	**89.7**	70.1

BLAST hits were reported with a bit score threshold of 50. Although at first glance the six-frame approach predicted more proteins than MSProGene, less than half of them can be mapped by BLAST, with less confidence than MSProGene hits. Further, only a small fraction of six-frame predicted proteins passes the confidence score threshold.

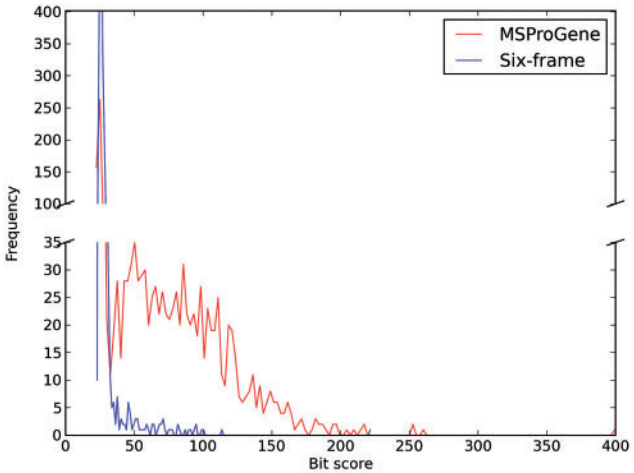

Fig. 7. The frequency of bit scores for MSProGene and the six-frame approach for the BLAST search of predicted sequences against a *B.malayi* reference. The confidence of MSProGene sequence alignments significantly exceeds the confidence of six-frame sequence alignments

sequences receive a BLAST hit at all, while in contrast 68% of MSProGene predictions can be mapped. With 1804 hits, the overall number of hits for the six-frame approach is still higher, but only in case no score cutoff for confidence control for the BLAST search is applied. When using a bit score cutoff of 50 as in (Armstrong *et al.*, 2014), the number of remaining BLAST hits of MSProGene is an order of magnitude higher than for the six-frame analysis. Hence, MSProGene identifies fewer transcripts with more confidence.

We are aware that the comparison against a *B.malayi* database can only identify proteins that are *L.sigmodontis* orthologs and does not determine proteins specific to *L.sigmodontis*. However, *L.sigmodontis* and *B.malayi* are close relatives. Hence, the BLAST search against *B.malayi* is a good indicator of the quality of *L.sigmodontis* protein identifications.

4.4 System requirements

The computational performance of MSProGene is evaluated using the transcripts predicted by GIIRA (for *B.henselae*) or Cufflinks (for *L.sigmodontis*). The main contributors to run time are the two spectra searches performed by MSGF+: The search of 1.16 million spectra on the *B.henselae* dataset required 35.7h. The search of 856 380 spectra on the *L.sigmodontis* dataset required 40.8h. Overall,

MSProGene used 30 GB RAM and 36.5 h to analyze the *B.henselae* dataset, and 30 GB RAM and 41.6 h to analyze the *L.sigmodontis* dataset.

5 Conclusion and outlook

We present MSProGene as a novel proteogenomic method for integration of proteomic, genomic and transcriptomic data beyond six-frame translation and the dependency on reference databases. We demonstrate the benefits of the new method in a comparison on three datasets and show that MSProGene provides an automated integrative framework for robust and precise proteogenomic analysis. We show that MSProGene performs peptide and protein identification with higher specificity than existing methods and constructs smaller customized spectra search databases. It is independent of *a priori* annotations and allows the identification of mutated and novel genes. Further, the network optimization employed by MSProGene successfully resolves shared peptides for protein inference without the need for protein grouping. This way, MSProGene distinguishes alternative isoforms and genes sharing homologous regions. Since the algorithm for peptide reassignment is independent of the constructed gene model, MSProGene can be combined with any prediction method or previously defined gene sequences of choice. Thus, given a suitable gene prediction, our method is also applicable to higher eukaryotes and polyploid organisms and can respect polyploid SNPs. Future applications of the software include a more thorough analysis of the simultaneous verification of SNPs on the transcriptome and proteome level and the analysis of variant peptides.

Acknowledgements

We thank all members of the NG4 bioinformatics group (Robert Koch Institute) and Wojtek Dabrowski (ZBS1, Robert Koch Institute) for inspirational discussions.

Funding

We gratefully acknowledge financial support by Deutsche Forschungsgemeinschaft (DFG), grant number RE3474/2-1 to B.Y.R.

Conflict of Interest: none declared

References

Ahn,J.-M. *et al.* (2013) Proteogenomic analysis of human chromosome 9-encoded genes from human samples and lung cancer tissues. *J. Proteome Res.,* **13**, 137–146.

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.,* **25**, 3389–3402.

Armstrong,S.D. *et al.* (2014) Comparative analysis of the secretome from a model filarial nematode (*Litomosoides sigmodontis*) reveals maximal diversity in gravid female parasites. *Mol. Cell Proteomics,* **13**, 2527–2544.

Bielow,C. *et al.* (2011) MSSimulator: simulation of mass spectrometry data. *J. Proteome Res.,* **10**, 2922–2929.

Blakeley,P. *et al.* (2012) Addressing statistical biases in nucleotide-derived protein databases for proteogenomic search strategies. *J. Proteome Res.,* **11**, 5221–5234.

Branca,R.M. *et al.* (2014) HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics. *Nat. Methods,* **11**, 59–62.

Castellana,N. and Bafna,V. (2010) Proteogenomics to discover the full coding content of genomes: a computational perspective. *J. Proteomics,* **73**, 2124–2135.

Castellana,N.E. *et al.* (2008) Discovery and revision of arabidopsis genes by proteogenomics. *PNAS,* **105**, 21034–21038.

Claassen,M. (2012) Inference and validation of protein identifications. *Mol. Cell Proteomics,* **11**, 1097–1104.

CPLEX. (2011) International Business Machines Corporation. v12.4: User's manual for CPLEX. *IBM ILOG CPLEX*.

Dempster,A.P. *et al.* (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Series. B,* **39**, 1–38.

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

Evans,V.C. *et al.* (2012) De novo derivation of proteomes from transcriptomes for transcript and protein identification. *Nat. Methods,* **9**, 1207–1211.

Fanayan,S. *et al.* (2013) Proteogenomic analysis of human colon carcinoma cell lines lim1215, lim1899, and lim2405. *J. Proteome Res.,* **12**, 1732–1742.

Giese,S.H. *et al.* (2014) Specificity control for read alignments using an artificial reference genome-guided false discovery rate. *Bioinformatics,* **30**, 9–16.

Grabherr,M.G. *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.,* **29**, 644–652.

Haas,B.J. *et al.* (2008) Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.,* **9**, R7.

Holtgrewe,M. (2010) Mason - a read simulator for second generation sequencing data. Technical Report TR-B-10-06, Fachbereich für Mathematik und Informatik, Freie Universität Berlin.

Homer,N. *et al.* (2009) BFAST: an alignment tool for large scale genome resequencing. *PloS One,* **4**, e7767.

Huang,T. *et al.* (2012) Protein inference: a review. *Briefings Bioinf*, **13**, 586–614.

Jeong,K. *et al.* (2012) False discovery rates in spectral identification. *BMC Bioinformatics,* **13**, S2.

Kelkar,D.S. *et al.* (2011) Proteogenomic analysis of mycobacterium tuberculosis by high resolution mass spectrometry. *Mol. Cell Proteomics,* **10**, M111–011627.

Kelkar,D.S. *et al.* (2014) Annotation of the zebrafish genome through an integrated transcriptomic and proteomic analysis. *Mol. Cell Proteomics,* **13**, 3184–3198.

Kim,D. *et al.* (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.,* **14**, R36.

Kim,S. and Pevzner,P.A. (2014) MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat. Commun.,* **5**, 5277.

Krug,K. *et al.* (2013) Deep coverage of the *Escherichia coli* proteome enables the assessment of false discovery rates in simple proteogenomic experiments. *Mol. Cell Proteomics,* **12**, 3420–3430.

Krug,K. *et al.* (2014) Construction and assessment of individualized proteogenomic databases for large-scale analysis of nonsynonymous single nucleotide variants. *Proteomics,* **14**, 2699–2708.

Li,H. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li,J. *et al.* (2010) Canprovar: a human cancer proteome variation database. *Hum. Mutat.,* **31**, 219–228.

Mohien,C.U. *et al.* (2013) A bioinformatics approach for integrated transcriptomic and proteomic comparative analyses of model and non-sequenced anopheline vectors of human malaria parasites. *Mol. Cell Proteomics,* **12**, 120–131.

Nagaraj,N. *et al.* (2011) Deep proteome and transcriptome mapping of a human cancer cell line. *Mol. Syst. Biol.,* **7**, 548.

Nesvizhskii,A.I. (2014) Proteogenomics: concepts, applications and computational strategies. *Nat. Methods,* **11**, 1114–1125.

Nilsson,T. *et al.* (2010) Mass spectrometry in high-throughput proteomics: ready for the big time. *Nat. Methods,* **7**, 681.

Ning,K. and Nesvizhskii,A.I. (2010) The utility of mass spectrometry-based proteomic data for validation of novel alternative splice forms reconstructed from RNA-seq data: a preliminary assessment. *BMC Bioinformatics,* **11**, S14.

Omasits,U. *et al.* (2013) Directed shotgun proteomics guided by saturated rna-seq identifies a complete expressed prokaryotic proteome. *Genome Res.,* **23**, 1916–1927.

Reiter,L. *et al.* (2009) Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. *Mol. Cell Proteomics,* **8**, 2405–2417.

Renard,B.Y. *et al.* (2010) Estimating the confidence of peptide identifications without decoy databases. *Anal. Chem.,* **82**, 4314–4318.

Renard,B.Y. *et al.* (2012) Overcoming species boundaries in peptide identification with Bayesian information criterion-driven error-tolerant peptide search (BICEPS). *Mol. Cell Proteomics,* **11**, M111–014167.

Rice,P. *et al.* (2000) EMBOSS: the European molecular biology open software suite. *Trends Genet.,* **16**, 276–277.

Safavi-Hemami,H. *et al.* (2014) Combined proteomic and transcriptomic interrogation of the venom gland of *Conus geographus* uncovers novel components and functional compartmentalization. *Mol. Cell Proteomics,* **13**, 938–953.

Serang,O. *et al.* (2010) Efficient marginalization to compute protein posterior probabilities from shotgun mass spectrometry data. *J. Proteome Res.,* **9**, 5346–5357.

Sherry,S.T. *et al.* (2001) dbsnp: the ncbi database of genetic variation. *Nucleic Acids Res.,* **29**, 308–311.

Trapnell,C. *et al.* (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.,* **28**, 511–515.

Vogel,C. and Marcotte,E.M. (2012) Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat. Rev. Genet.,* **13**, 227–232.

Wang,X. and Zhang,B. (2013) customprodb: an r package to generate customized protein databases from RNA-seq data for proteomics search. *Bioinformatics*, **29**, 3235–3237.

Wang,X. and Zhang,B. (2014) Integrating genomic, transcriptomic and interactome data to improve peptide and protein identification in shotgun proteomics. *J. Proteome Res.,* **13**, 2715–2723.

Wang,X. *et al.* (2011) Protein identification using customized protein sequence databases derived from RNA-seq data. *J. Proteome Res.,* **11**, 1009–1017.

Wang,Z. *et al.* (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.,* **10**, 57–63.

Woo,S. *et al.* (2013) Proteogenomic database construction driven from large scale RNA-seq data. *J. Proteome Res.,* **13**, 21–28.

Zickmann,F. *et al.* (2014) GIIRA – RNA-Seq driven gene finding incorporating ambiguous reads. *Bioinformatics,* **30**, 606–613.

Zickmann,F. and Renard,B.Y. (2015) IPred-integrating ab initio and evidence based gene predictions to improve prediction accuracy. *BMC Genomics,* **16**, 134.