Leibniz-Gemeinschaft

Final report

Title of the project:

(Reverse) Proteomics as novel tool for biodiversity research

Leibniz-Institute: Leibniz-Institut für Analytische Wissenschaften – ISAS – e.V., Bunsen-Kirchhoff-Straße 11, 44139, Dortmund, Germany

Reference number: SAW-2014-ISAS-2 291

Project period: 01.07.2014 – 31.12.2017

Contact partner: Leibniz-Institut für Analytische Wissenschaften – ISAS – e.V.

Contents

Executive summary

Biodiversity on all levels of biological organisation from molecules to ecosystems is the basis of our daily life. However, the diversity of species is drastically declining and we are thus increasingly losing the foundation of our life. Moreover, a significant part of existing species cannot be distinguished by traditional taxonomic techniques. We are thus in need of fast and accurate methods to detect and re-identify biological diversity at the species level. The aim of this project was to probe the opportunities and limitations of peptide/protein based classification of organisms, “reverse” (proteome) annotation of raw genome data and detection of candidate sequences for taxonomy.
Model organisms chosen for this project are foraminifera species *Amphistegina* and the freshwater gastropod *Radix auricularia*. Foraminifera – single cell organisms with a calcareous skeleton – are of particular interest in environmental research and are known to react differently to external conditions based on difference in their genome and like corals are particularly affected by climate change. Molluscs are the second largest group of animals after gastropods, but the scientific knowledge is minor compared to many other less abundant and diverse organisms. Freshwater gastropods as well as foraminifers are difficult to assess by molecular methods due to their kind of tissue and quantity of DNA.

In the course of the project, a proteome-wide comparison of tandem mass spectra (MS/MS) similarity algorithm called DISMS2 was developed and foraminifera species were used as test organisms. Using this algorithm, the different holobiont species could be differentiated and clustered and further results of a blind-test of this approach, including four different foraminifera species, and the several sample groups of the same species collected at different times and in different habitats, show a remarkably high success rate.

By comparing the responses of different species and the same species originating from different environments, i.e. depths, when exposed to repetitive vs. chronic stress exposure (i.e. heat stress) we revealed insight into underlying molecular mechanisms on the protein level and detect the so far unknown bleaching mechanisms in photosymbiotic foraminifera. For the first time, we could differentiate between symbionts and host responses on a cellular level by separating the detected proteins using a homology-driven search approach on our two-compartment (foraminifera and diatom) database.

The genome sequencing of *R. auricularia* during this project resulted in an assembly comprised 4,823 scaffolds with a cumulative length of 910 Mb and an overall read coverage of 72. The assembly contained 94.6% of a metazoan core gene collection, indicating an almost complete coverage of the coding fraction. A comprehensive proteomics dataset was generated and a proteogenomics data analysis pipeline was established, combining information from a predicted protein database, the genome level as well as from independent *de novo* peptide sequencing. Although using the peptide sequences as sole source for informing gene annotation did not result in a reasonably complete annotation, the peptides improved in particular the exact exon/intron location annotation of many genes. About 769 genes were additionally detected, some additional exons were identified for previously found genes and others excluded as potential artefacts.

In summary, the methods and data developed and generated in this project could considerably increase the knowledge in particularly underrepresented areas of research, providing a firm basis for further investigations and already proofing its potential to be utilised in a wide-ranging field of applications.

1. Opening questions and aims of the project

Under the current global and climate change, the loss of biodiversity is a major threat. However, only a small percentage of organisms has been appropriately described and identified yet, because methods to efficiently estimate biodiversity are lacking May (1988), (Bailie, Hilton-Taylor et al. 2004). This has the consequence that many species are going extinct before they were even recognized. Traditional taxonomic techniques are often not able to distinguish between highly similar appearing species(Pfenninger and Schwenk 2007), i.e. cryptic species, especially in small and highly diverse marine protists such as foraminifera. These unicellular organisms are mostly identified by the shape of their calcareous shells, called tests and form an inconspicuous element of global marine biodiversity. Some benthic foraminifers harbour photosynthetic symbionts, and like corals these species are particularly affected by climate change. Unfortunately, foraminifera are difficult to assess by molecular methods due to their kind of tissue or quantity of DNA (Pawlowksi and Holzmann 2002, Pereia and Chaves 2011). One of the most well recognized genera in terms of physiology, biogeography and ecology, *Amphistegina* spp., genuinely lacks genomic data and remains a molecular and evolutionary mystery as to how they adapt to environmental change. The genus harbours diatoms as symbionts and is widespread in the Indo-Pacific, Caribbean, Red Sea and West Atlantic, and is known for its morphological test modification (Hallock, Forward et al. 1986), but also sensitivity to environmental changes (Schmidt et al. 2011, Hallock et al. 2006, Talge & Hallock 1995 & 2003) . Their species concepts have a history of continuous revision, because based on test shape (exoskeleton) alone, foraminiferal identification largely disregards morphological plasticity. In most cases plasticity is a direct and systematic consequence of environmental conditions, as previously observed from genotyping planktonic foraminifera (e.g. Jan Pawlowski group, and(Darling and Wade 2008)), which reveal distinct ecologies, novel adaptations, and overall the true species diversity. Here, traditional species concepts have partitioned morphological types into distinct species based on test shape, but genetic studies show that individual morphospecies are actually complexes of several discrete genetic types (genotypes). Many of these genotypes have distinct ecologies and novel adaptations that are consistent with species-level classification, suggesting that also the true diversity of benthic foraminifers has been greatly underestimated.

At the intraspecies molecular level, which is the basis of evolution and thus essential for the resilience and the evolutionary potential of species (Pauls, Nowak et al. 2012), knowledge in warm-water shallow foraminifera species such as *Amphistegina* knowledge is even more scarce. Based on differences in their genome, individuals react in different ways to environmental conditions, and develop in different ways in their ontogeny. These differences can be due to structural differences in the proteins, but also in their temporal or spatial expression patterns. Moreover, larger benthic foraminifers react to stressful environmental conditions such as high temperature by bleaching, alike the phenomenon known from tropical shallow-water corals. Studying the various symbiont species would therefore also add to a better understanding of their resilience. It was expected that this group of foraminifers has one of the highest chances of resilience towards ocean acidification and global warming, because they can select resilient photosymbiont type, have a skeletal mineralogy (low-Mg calcite) of low solubility and their test plasticity allows to activity change the light conditions within the test.

Benthic foraminifers in modern and ancient sediment are widely used to reconstruct environmental conditions. The question whether they really represent the same taxon is of great importance. For photosymbiotic species such as *Amphistegin*a, this question is even more interesting because of the potential influence of different species of the symbiotic algae that may allow for adaptation to a wider range of environmental conditions than if only one algal species was available for symbiosis.

In order to improve our knowledge of the diversity of the genus and identify genetic adaptive traits to environmental conditions and ecological pressure, and to provide a complete, high quality and freely accessible resource of information about modern foraminiferal species, we aimed to present morphological description, photos/ scanning electron microscope (SEM) images, DNA sequences, and references to related publications for the “proteomics” barcode database, in accord with the PR2(Guillou and Bachar 2013) database catalogue requirements. In the frame of the network it was envisioned to focus on the latter part of the project and use foraminifera to probe the feasibility to do a proteomics based taxonomic analysis, and potentially define novel candidate genes / proteins for taxonomic classification, improve classification of foraminifer (determination of possible cryptospecies in foraminifers) as well as their symbionts (distribution and diversity of symbionts within and between regional habitats).

Proteomics technologies are now on the way to become the next “genomics” (Cox and Mann 2007). However, the main limitation of most proteomics strategies is the availability of complete sequenced and correctly annotated model organisms, because of the large scale (semi-)automated annotation peptide spectra with the correct peptide sequence. For this process annotated genome is translated to protein sequences and afterwards ‘digested’ in silico. It is vital for the establishment of the technique that the proteome is being sequenced from exactly the same individual or strain the genome sequence was derived. This restricts annotation problems arising from polymorphisms to a minimum. In order to maximise the scientific exploit of this venture, we proposed the use of *Radix sp*., a basommatophoran freshwater snail which is cultured in the Pfenninger laboratory (Senckenberg Biodiversity and Climate Research Centre (BiK-F) Frankfurt am Main, Germany). Despite Mollusca being the second largest animal phylum, only three genomes (two mussels, one sea slug) were published at the time of the project proposal. The annotation of the coding genes and their exon-intron structure is poor and difficult (Yoshida, Ishikura et al. 2011), mostly because of the large phylogenetic distance to the well-annotated and –characterised model organisms like mouse, fruit-fly and humans.

The taxon *Radix sp*. is an emerging model organism for climate response (Pfenninger Group, BiK-F), evolutionary development, host-parasite coevolution and mating-system evolution. Additionally it is used in ecotoxicological studies. Its ecology, taxonomy, population history, mating system variation and range wide population structure are well characterised (Pfenninger, Cordellier et al. 2006, Haun, Salinger et al. 2012, Teixeira, Rainha et al. 2012). As vector for several parasites, the species has also a not negligible socio-economic impact. The species easily breeds in the laboratory and can be manipulated in experiments. With the characterization of the transcriptome just prior to this project proposal (Feldmeyer, Wheat et al. 2011), the first step towards establishing the full set of genomics resources has been made. The aim was to test whether the method has the potential to supersede current transcriptome analyses, and thus opening the potential of the proteome sequencing method for (i) modern biodiversity research, (ii) to develop new methods to devise Gene Prediction models, (iii) to obtain a new, well annotated genome of a currently understudied taxon that will trigger new, additional research with this taxon and (iv) to finally inspire and facilitate the sequencing of other ‘exotic’ genomes.

2. Progression of the realized work including deviations from the initial concept, scientific failures, problems in the organisation of the project or the technical execution

1. Foraminifera

It was planned to collect the studied organisms from the biogeographic areas of interest (e.g. Aqaba, Bahamas and Zanzibar during field campaigns combined with ongoing research activities in these areas, and additionally culture them at Leibniz Center for Tropical Marine Ecology (ZMT), Bremen and MARUM - Center for Marine Environmental Sciences, University of Bremen, Germany) to ensure that the environmental parameters of the biogeographical study areas are well-constrained and sufficient material for the proteomics research is available. In the first work-package the genome of *Amphistegina* spp. was supposed to be sequenced, along with large-scale proteome analysis (including optimization of sample preparation and data analysis and interpretation), collection of public available genome and transcriptome data, and. After application and programming of genome data annotation algorithms, application to the generated large scale datasets, statistical evaluation of similarities between spectra and the usage of peptides features for multi parameter classification of organisms by the partners in Dortmund (Faculty of Statistics, Technical University Dortmund and ISAS), the analysis of intra species variations was supposed to follow, i.e. mass spectrometry (MS)-based analysis of the same species from different habitats to screen for intra species variations (development of species) and intra species protein expression profiles (adaptation to stress factors). Lastly, data integration and mining through analysis of curated and generated datasets for “taxon” and “stress/environment” markers, validation of derived markers and development of fast, reliable and matrix independent MS-based detection methods as well as the application of methods to species from different habitats was envisioned as final application of the novel approaches.

To cooperate with the most knowledgeable researcher worldwide on the genus *Amphistegina*, Pamela Hallock, specimens from the Florida Keys instead of the Bahamas were sampled, covering the same biogeographic area. Moreover, samples from Zanzibar were collected and cultures were set up successfully at ZMT, Bremen. During optimization of the proteomic workflow (at ISAS, Dortmund), the photosymbiotic character of the foraminifera and other contaminations on and (due to feeding on bacteria etc.) probably also inside the test showed to be problematic, as not only the foraminiferal proteins/genes were present, but also those of the symbiotic diatoms and other microorganisms living in the same habitats. In contrast to larger organisms where single organs can be extracted, the foraminifera had to be analysed completely and several specimens (e.g. 5-10) were pooled to generate sufficient amount of starting material for proteomics studies. To sequence the genome and establish a baseline proteomics dataset, uncontaminated material is necessary. Therefore, it was attempted to isolate host and symbionts separately. To achieve photosymbiont-free foraminifera with minimum microbial contamination, different approaches to entirely bleach, but not to kill them were tested, including exposure to stressful conditions such as high light intensities and total darkness combined with herbicide/photosynthesis-inhibitor DCMU (e.g. after Koestler et al. 1985, Lee et al. 1986, Lee et al. 1991, van Dam et al. 2012). Furthermore, the foraminifera were cleaned to maximum by brushing and rinsing with sterile seawater and incubation in sterile artificial seawater containing an antibiotic-antimycotic mixture. Unfortunately, the complete removal of symbionts seemed impossible, as even totally white specimens regained (symbiont) colour after rinsing and transfer into sterile seawater without DCMU. The symbionts were extracted and cultured under aseptic conditions (after mixed protocols by Barnes & Hallock unpublished, Schmidt 2015, Lee 1992, Koestler et al. 1985). As a result of several months of cultivation efforts, some diatoms were successfully found from *A. lessonii* and imaged by SEM scans, but not *A. gibbosa*, and the quantity achieved was very low and considered insufficient for proteomic analysis. Moreover, it was found that culturing is strongly biasing the resulting symbiont community depending on culturing conditions, especially due to the presence of more than one symbiont species in *Amphistegina*. Since neither the isolation of the host nor the symbionts was successful, and discussions with experts in foraminiferal genetics revealed that sequencing their genomes is nearly impossible under the current conditions; this part of the project had to be replaced by another usage of the proteomic tools available.

Our preliminary experiments showed that it was indeed possible to analyse the holobiont proteome (i.e. the host including the various symbionts) and identify proteins by homology-based searches against our successfully collected database of publically available gene and protein data from distantly related non-symbiotic foraminifera and free-living diatoms. Therefore, we decided to use a label-free bottom-up proteomic approach to focus more on the intra- and inter-specific variations as well as the adaptive potential of *Amphistegina*, potentially resulting from different environmental conditions or variations in the symbiont assemblages. By comparing the responses of different species and the same species originating from different environments, i.e. depths, when exposed to repetitive vs. chronic stress exposure (i.e. heat stress) we aimed to reveal insight into underlying molecular mechanisms on the protein level and detect the so far unknown bleaching mechanisms in photosymbiotic foraminifera. For the first time, we tried to differentiate between symbionts and host responses on a cellular level by separating the detected proteins using a homology-driven search approach on our two-compartment (foraminifera and diatom) database. To evaluate this approach, the proteomic analysis had to be combined with well-established physiological methods. These showed a highly similar response and thus validated the protein identification and related compartment assignment, and hence represented a great success of our approach provides basis for future research on large benthic foraminifera and other photosymbiotic organisms such as corals and has great potential to reveal key proteins/genes for resilience and the fate of coral reefs under climate change etc.

Besides, variations of the photosynthetic endosymbiont community in the respective *Amphistegina spp*. and its role for adaptation to environmental conditions were tested. For this, genetic fingerprinting (sequencing of SSU region) was conducted on the symbionts, using marker genes only present in photosymbiotic organisms such that only the algal symbionts are detected. This revealed large differenced in the specificity between host species and diatom assemblages, indicating that more resilient foraminifera species are more flexible in their symbiosis, i.e. harbour a wide range of symbiont species and are potentially able to adjust this to changing conditions

Although the usage of foraminiferal samples for the development of the novel taxonomic approach was partly obstructed by the above-mentioned problems, they were still used as test organisms for the proteome-wide comparison of tandem mass spectra (MS/MS) similarity algorithm called DISMS2 (*see section 3*) developed by the Statistics department, TU Dortmund (Rieder et al. 2017a). Using this algorithm, the different holobiont species could be differentiated and clustered and further results of a blind-test of this approach, including four different foraminifera species, and the several sample groups of the same species collected at different times and in different habitats, show a remarkably high success rate. This was also confirmed by comparison to a de novo annotation approach conducted on the peptide data by the ISAS. At the moment it still has to be evaluated how far this may relate to the differences in photosymbiont assemblages (i.e. genetic fingerprinting of the diatoms in these samples is under way) and whether the resulting *Amphistegina* phylogeny corresponds to the results of genetic markers (potentially to be done with colleagues at Marum soon).

In order to reach the goal of applying proteomics data in the proposed way, all major steps of analysis were evaluated and optimized for the given purpose. Foraminifera in general and the studied genus *Amphistegina* in particular are comprised predominantly of a calcareous shell and only contain minute amounts of protein per specimen, usually in the sub-µg range. Therefore, an adapted protocol of the filter-aided sample preparation protocol (FASP)(Wisniewski, Zougman et al. 2009) was established to allow for reproducible protein extraction, clean-up and digestion. The number of required specimens could be reduced from 20 to a single exemplar, with best results being achieved from 6 to 8 pooled foraminifera. This facilitated both efforts in environmental sampling and in-house culturing.

To provide a complementary approach to DNA-barcoding based on proteomics data in taxonomic classification studies, it was decided to implement direct analysis of MS/MS spectra by *de novo* peptide sequencing, as this approach is entirely independent of the availability of genomic or transcriptomic data. However, while the appropriate control of false discovery rate (FDR) is well established and implemented in common database-search based proteomics analyses, this was not the case for *de novo* peptide sequencing. Therefore datasets of well characterized samples (HeLa, human cervical cancer cell line; C2C12, mouse muscle cell line; W303, *S. cerevisiae*) were generated as ground truth in order to optimize both parameters for data acquisition and interpretation. As *de novo* peptide sequencing benefits greatly from high-resolution MS/MS data, the Q Exactive HF MS (Thermo Scientific) was the instrument of choice for these analyses. For data analysis and interpretation, commercially available software PEAKS Studio 7.5 (Jing Zhang, Lei Xin et al. 2012) (which is by far the best tool for analysing *de novo* peptide sequencing datasets) was compared to the freely available *de novo* peptide sequencing algorithms pNovo+, PepNovo and Novor (Frank and Pevzner 2005, Chi, Chen et al. 2013, Ma 2015). While no single algorithm performed considerably well in comparison to a reference database search (Mascot 2.4, Sequest, and MS Amanda (John R. Yates, Jimmy K. Eng et al. 1996, Dorfer, Pichler et al. 2014), using Proteome Discoverer 1.4 software (Thermo Fisher Scientific) and FDR-controlled by Percolator), it was established that a combinatory approach of two or three *de novo* peptide sequencing algorithms with the subsequent replacement of individual scoring schemes with the combined sequence agreement as discriminatory measure was particularly advantageous. Chimeric spectra stemming from co-isolation of simultaneously eluting peptides was found to be a major driver of sequence miss-annotations. Usage of a sufficiently narrow isolation window could be shown to reduce the false positive identifications by 50% while maintaining the same sensitivity. Altogether, the number of PSMs meeting a 5% FDR value could be increased more than threefold compared to the single best *de novo* sequencing algorithm alone. (Blank-Landeshammer et al., 2017).

Peptide sequences assigned by the optimized *de novo* peptide sequencing workflow could be used in conjunction with the DISMS2 algorithm to perform a proteome-wide comparison and subsequent differentiation of several foraminifera holobiont species (Stuhr et al., in preparation).

As proposed in the project application, the use of an alternative enzyme LysN was evaluated to aid in *de novo* peptide sequencing, as spectra from LysN-generated peptides should be dominated by b-ions after collision-induced dissociation, which would facilitate correct peptide annotation. However, while this proposition generally holds true, we found that the generated peptides are considerably longer than tryptic peptides, thus hampering the *de novo* peptide sequencing accuracy so that the overall identification rate could not be improved compared to tryptic digests.

In the course of the project the focus of interest broadened from taxonomic comparisons to functional analysis and the adaptive potential of *Amphistegina*. Hence, a label-free proteomics approach was chosen and an adapted homology-based database search strategy had to be established, as adequate protein inference is necessary achieve meaningful functional understanding. This was done by collecting publicly available protein and genome sequence information from both distantly related non-symbiotic foraminifera and free-living diatoms which were combined to a custom database. The acquired MS/MS spectra were subsequently searched with PEAKS Spider, allowing for single amino acid variants likely to occur due to phylogenic distance. Finally the confidently identified proteins were clustered by the use of a similarity-driven algorithm (cd-hit). (Stuhr et al. 2017)

2. *Radix auricularia*

Genome sequencing and –annotation for protein database enlargement

As model organism, an inbred strain of the freshwater snail *Radix auricularia* was chosen. Six whole genome shotgun libraries with different layouts were sequenced. The resulting assembly comprised 4,823 scaffolds with a cumulative length of 910 Mb and an overall read coverage of 72. The assembly contained 94.6% of a metazoan core gene collection, indicating an almost complete coverage of the coding fraction. The discrepancy of ~ 690 Mb compared with the estimated genome size of 1.6 Gb results from a high repeat content of 70%, mainly comprising DNA transposons. The annotation of 17,338 protein coding genes was supported by the use of publicly available transcriptome data. Because of the respectively unique and specific competence profile of the network partners, the genome sequencing, -assembly and structural annotation was performed by the SGN partner alone and published by Schell et al. (2017).

The high quality genome allowed the ISAS partners to supplement their protein databases successfully. To make full use of the data, regular project meetings (quarterly through the project period) were indispensable to communicate and discuss their specifications and peculiarities.

In parallel, at ISAS a comprehensive proteomics dataset was generated. This was achieved by dissection of snails and subsequent digestion with complementary proteases (i.e. trypsin, GluC, LysN, subtilisin) followed by high-pH RP-fractionation and LC-HR/AM MS analysis with *de novo* peptide sequencing-optimized settings as described above. This way, over 3 million high-resolution peptide spectra were acquired. These were fed to a customized tripartite data analysis pipeline. First, all spectra were searched against the first draft of the *R. auricularia* protein database, annotated based on a combination of in silico gene prediction, transcript-level evidence and homology –driven algorithms (MAKER pipeline, see above). As appropriate control of the FDR is a crucial factor in proteogenomics search approaches and can easily be underestimated, this was taken particular care for. Only non-matching spectra of the first search, additionally meeting certain quality criteria were passed on to a second search against a 6-frame translation of the assembled *R. auricularia* genome and results were filtered to meet a 0.1% FDR cutoff. Thirdly, all spectra were annotated by the above-mentioned de novo peptide sequencing pipeline. This combined effort led to the identification of 769 additional protein coding genes and the substantial refinement of the gene structure *R. auricularia*. Additionally, N-terminal peptides were enriched by charge-based fractional diagonal chromatography (ChaFRADIC)Venne, Vogtle et al. (2013) leading to the confirmation of 1,265 predicted translational initiation sites (TIS), while 373 alternative TIS could be identified this way.

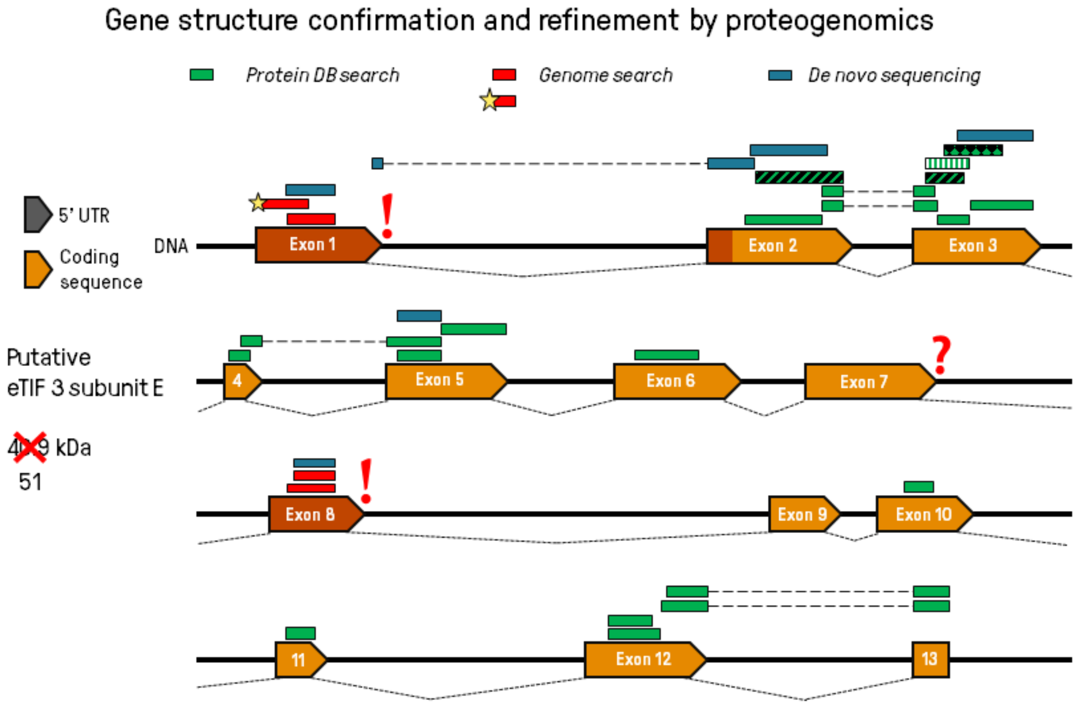

Figure 1 Example of *a R. auricularia* gene refined by proteogenomics methods, identifying two additional exons in a putative eTIF 3 subunit E gene. Green bars represent peptides identified in the predicted protein database search, red bars represent peptides found on genome-level only and blue bars depict de novo sequenced peptides

De novo peptide sequences for annotation improvement

It was one of the major goals of the proposal to use *de novo* peptide sequences to improve genome annotations. On the basis of *Radix auricularia* samples of different developmental stages from the same inbred line the genome sequencing was performed on, the ISAS partners were able to retrieve several tens of thousands *de novo* peptide sequences with MS/MS spectra and consecutive database searches. These peptide sequences were then searched for in the genome sequence by retranslating them using suitable software bioinformatically into potential DNA-sequences, respectively and vice versa.

One of the major results of the project was a very good congruence between the transcriptome based "traditional" annotation and the aligned peptides. The by far largest part of the *de novo* peptide sequences was located in already annotated genes (94.2%). This approach provided thus most genes (65.1%) in the genome with actual protein evidence, which is apart from man and mouse, rarely the case for animal genomes. Moreover, the peptides improved in particular the exact exon/intron location annotation of many genes. To make full use of the few peptide sequences that were not aligned to already annotated genes, we used the peptides as additional hints in the training of a Hidden-Markov-Model that was then integrated into an existing annotation pipeline. This led to a slight improvement of the annotation: about 769 genes (+4.4%) were additionally found, some additional exons were identified for previously found genes and others excluded as potential artefacts. However, using the peptide sequences as sole source for informing gene annotation did not result in a reasonably complete annotation.

In conclusion, it could be shown that *de novo* proteome sequencing of unknown biodiversity samples can improve their genome annotation, it is, however, not (yet) possible to completely replace classical transcriptome sequencing of several developmental stages by proteomic approaches. A manuscript reporting these results and involving researchers from the project partners ISAS and SGN is currently under preparation.

3. Cluster analysis

We have developed and evaluated cluster methods for the analysis of MS data in biodiversity research. Alternatively to species identification by DNA barcoding, analyses of the protein composition of organisms were performed. Tandem mass spectra consisting of detected intensities of masses can be used to identify peptides by means of database search algorithms. Today, often unknown peptides are detected via error prone de novo peptide sequencing algorithms. As an alternative to annotation methods, we investigated the approach to directly cluster the tandem mass spectra. Two aspects, cluster analysis of the runs involving thousands of spectra of a protein sample, and cluster analysis of individual tandem mass spectra, were considered. The data are based on datasets created in the project (DISMS2, foraminifera, biodiversity- Q Exactive, biodiversity Orbitrap Elite) and on a public dataset (palmblad (Magnus and M. 2012)).

A cluster analysis of runs was performed for all data sets using the new method DISMS2 (Rieder et al., 2017a), which determines distances between MS/MS runs without annotation. Thus, it is an alternative to comparing peptide lists based on the identification of spectra in database searches. The parameters of DISMS2 can be freely chosen. The parameters address the selection of the highest peaks per spectrum (topn), the binning size in the binning (bin), a restriction in the comparison of spectra to those that are temporally close (ret) and those that have similar precursor mass (prec), and the distance measure for mass spectra (dist) with a freely selectable threshold (cdis). For parameter selection, an optimization procedure was used which uses the coefficient of determination R2 of a nonparametric analysis of variance. The resulting distances were represented by means of a hierarchical cluster analysis as dendrograms, resulting in phylogenetic trees.

For the DISMS2 datasets, some organisms have annotations derived from a database search, and the foraminifera dataset contains *de novo* annotations. The optimization of the parameter settings in the DISMS2 algorithm lead to a high R2 value for the data set DISMS2 (R2 = 0.923). In the corresponding dendrogram, which contains three technical replicates each of human, mouse, yeast, roundworm, fruit fly, foraminifera and freshwater snail samples, a clear separation of the groups can be seen. In the Palmblad dataset, which includes blood sera from apes and other primates, the distances are much larger and the distinction between some types is not possible.

The present annotations in the DISMS2 dataset were used to validate the DISMS2 distances. Differences between annotations and DISMS2 distances were observed, but the algorithm steps are not comparable. In a fair comparison, where the steps are the same except for the method (database or distance search), the differences between the results were only small. Using the cosine distance of the mass spectra, a binary classifier was constructed that distinguishes between the same and different peptides. For the threshold cdis = 0.3, the sensitivity and specificity were 0.923 and 0.867, when two runs are compared. The competitiveness of

DISMS2 compared to annotation distances was confirmed by the foraminifera data set. However, distinguishing closely related foraminifera species, each of which has three biological replicates, is more difficult.

The generation and the pre-processing of the mass spectra influence the DISMS2 distances. Prior to application of the algorithm, 2000 spectra with high total ion intensity were selected for the foraminifera data set. The resulting dendrogram could be confirmed by a de novo method. A clear separation of the foraminifera species *A. gibbosa* and *M. vertebralis* was observed. The species *A. lessonii* and *A. lobifera* showed a higher similarity.

The use of different mass spectrometers has a significant influence on the DISMS2 distances. The Biodiversity datasets contain *Radix* specimens and other specimens as reference. Technical replicates were analysed using Q Exactive HF and Orbitrap Elite mass spectrometers (both Thermo Scientific). In the dendrograms a clear separation between the *Radix* and other samples was observed, but not between individual *Radix* species or between the origin respectively different parts of the snails.

Runtime and memory consumption of the implementation of the R algorithm were examined during the optimization for the DISMS2 data set. Memory usage was low (below 3GB) as distances are calculated only when needed after checking multiple conditions. The runtime was almost 16 hours for the optimal settings. Especially an analysis of little explored species benefits from the new method DISMS2. A match between de novo respectively database annotations and the DISMS2 algorithm was observed.

For the cluster analysis of individual mass spectra, a comprehensive comparison of algorithms was performed, including established algorithms for tandem mass spectra (CAST, MS-cluster, PRIDE cluster) and for large data sets (hierarchical cluster analysis, DBSCAN, connected components of a graph) as well as a new algorithm (neighbor clustering). The evaluation was done on the DISMS2 dataset, applying several quality measures.

First, annotations were compared with cluster results of the seven algorithms and different parameter settings, applied to the tandem mass spectra of the DISMS2 data set. It was shown that established methods and the recently introduced neighbor clustering algorithm (N-Cluster) are at least as good as the proteomics-based MS-Cluster and PRIDE Cluster. A clear improvement of the cluster solutions is achieved by the DISMS2 filter, which takes into account precursor charge, precursor mass and retention time. First, the parameter settings were optimized with respect to the highest mean adjusted Rand index, when cluster results were compared to the annotations. On average, none of the methods is optimal with respect to all quality measures considered. For PRIDE, the proportion of clusters with more than one spectrum is highest and the number of clusters in relation to the number of spectra is smallest. DBSCAN performs best in terms of ARI values and regarding the proportion of remaining annotations. The smallest proportion of spectra without the most frequent annotation is provided by MS-Cluster. There are only slight differences between the algorithms with respect to variation of the values (standard error below 0.02).

The cluster analysis was then extended to multiple runs of three technical replicas of the DISMS2 dataset. The proteome of a species is better represented when evaluating multiple measurements. Hierarchical clustering did not provide solutions for roundworm, mouse, and human, because the memory limit was exceeded. Compared to the evaluation of individual runs, a drastic improvement in the values of the evaluation measures can be observed.

Finally, clustering of mass spectra was linked to the DISMS2 algorithm. Using cluster analysis (hclust or PRIDE) prior to applying the DISMS2 algorithm did not improve the original algorithm, but even increases the distances within the groups. One application of clustering spectra is the peptide assignment of spectra with no annotation. In view of the high number of unannotated spectra, the subsequent assignment could be successfully performed only in a few cases.

Compared with the runtime and memory consumption of solutions from different implementations of clustering, MS-cluster and PRIDE clusters are best. Cluster solutions were generated in less than 10 minutes with a storage requirement of 0.15 GB or less. For the remaining algorithms, implementations in R whose input is a distance matrix vector were used. Clustering multiple runs required a maximum of more than 60 GB of RAM and approximately 67 minutes.

4. Presentation of the achieved results and discussion with regard to the relevant state of research, possible perspectives of applications and sequential projects

See above. More projects already resulted from this such as with Justin Ries, papers in progress (see list of publications), DFG proposal in cooperation with ISAS, TU Dortmund, Israel and South Africa (rejected for special call but planned to be resubmitted in adapted form), new proposal in progress. The tripartite genome annotation workflow developed at ISAS in this project is currently being applied in collaboration with Prof. U. Kück (RUB Bochum) in order to refine the genome annotation of the well-studied model organism *S. macrospora*. The FDR-adjusted database search strategies and combined *de novo* peptide sequencing approach are successfully utilized in a collaboration with Ines Teichert (RUB Bochum) in an effort to detect recoding RNA-Editing events in *S. macropsora* (DFG proposal submitted). Further, adaptions of the abovementioned developed search strategies are vital elements of ISAS' contribution to the Cancer Moonshot project (coordinator: NIH (National Institutes of Health), USA), an international effort to accelerate the development of new cancer treatments as well as in a collaborative research cluster on platelet function (DFG Transregio 240).

5. Statement on the commercial exploitability of the project results and to which extend this is has been undertaken or is to be expected; details regarding potential patents and industry cooperations

After further development of the bottom-up proteomics approach, this will likely be used for further scientific projects, a commercial use of the analytical pathway is possible. Maybe for monitoring purposes, if specific proteins are targeted or identified as indicators for specific conditions for environmental assessments, if successfully applicable to other photosymbiotic calcifiers such as corals this may be used for aquarium economy and coral farming etc. Generally the overall results valuable for environmental management of marine habitats, the fundamental elements of the research provide basis for further projects .No patents or cooperation with industry so far.

6. Statement on the contribution of potential national or international cooperation partners

Michal Kucera, University of Bremen, Germany, provided support and laboratory facilities at the Center for Marine Environmental Sciences (MARUM) to conduct thermal stress experiment for Stuhr et al. (2017, 2018 and in press)

Pamela Hallock, University of South Florida, USA, supported the collection of foraminifera from Tennessee Reef at the Keys Marine Laboratory (KML), and hosted M. Stuhr twice in her working group "Bioindicators" at the College for Marine Science in St. Petersburg, Florida, for preparing the collected samples for transport to Bremen for Stuhr et al. (2017, 2018 and in press); also member in thesis (dissertation) committee of M. Stuhr.

Christopher A. Muhando, University of Dar es Salaam, Tanzania, supported foraminifera sampling efforts of G.R. Narayan and C.E. Reymond at the Insititute for Marine Sciences (IMS) in Stonetown, Tanzania, and provided unpublished bottom water temperature data from Zanzibar for Stuhr et al. (in press).

Justin B. Ries and Louise Cameron, Northeastern University, Boston, USA, conducted combined ocean acidification and warming experiment in the marine experimental aquarium facility

(MAREE) at ZMT, with corals and foraminifera, in collaboration with M. Stuhr and C.E. Reymond for Stuhr et al. (in prep.)

Amatzia Genin, Inter-University Institute for Marine Sciences (IUI), Eilat, Israel, supported foraminifera sampling efforts in the Red Sea and hosted M. Stuhr and C. E. Reymond at the IUI diving center and laboratories for Stuhr et al. (in prep).

7. Dissertations in connection to the project

Dissertation of Vera Rieder. Clustermethoden für Massenspektren in proteomweiten statistischen Analysen. Fakultät Statistik, Technische Universität Dortmund, Dortmund. 2018, Retrieved from https://doi.org/10.17877/de290r-18840

Dissertation of Marleen Stuhr: "Disentangling the effects of thermal stress on symbiont-bearing coral reef foraminifera – from populations to proteins", University of Bremen, submitted in December 2017, defended on January 31st 2018 (available online: http://nbn-resolving.de/urn:nbn:de:gbv:46-00106487-18)

The PhD thesis of Tilman Schell will be submitted within the next month; a substantial delay was caused by the fact that TS was hired as Bioinformatician in the LOEWE-TBG project of Senckenberg.

The PhD thesis of Bernhard Blank-Landeshammer is under preparation and will be submitted by end of 2018.

8. List of publications related to the project

a. Scientific publications (in preparation)

Blank-Landeshammer B, Schell T, Kollipara L, Biß K, Zahedi RP, Pfenninger M, Sickmann A, Refinement and validation of gene prediction in R. auricularia by proteomics, in preparation for submission to Genome Biology and Evolution

Stuhr M, Blank-Landeshammer B, Cameron L, Ries JB, Sickmann A, Westphal H, Reymond CE (in prep.) Adaptive physiology of reef-associated foraminifera to the combined effects of ocean acidification and warming; in preparation for submission to Frontiers in Marine Science

Stuhr M, Hallock P, Blank-Landeshammer B, Rieder V, Meyer A, … (in prep.) Comparison of *Amphistegina* phylogenies based on genotyping, proteome-wide distance measure DISMS2 and *de novo* peptide sequencing; in preparation for submission to Journal of Foraminiferal Research

b. Scientific publications (published)

Blank-Landeshammer B, Kollipara L, Biß K, Pfenninger M, Malchow S, Shuvaev K, Zahedi RP, Sickmann A (2017) Combining De Novo Peptide Sequencing Algorithms, A Synergistic Approach to Boost Both Identifications and Confidence in Bottom-up Proteomics. Journal of Proteome Research. Sep 1;16(9):3209-3218. doi: 10.1021/acs.jproteome.7b00198

Rieder, V., Blank-Landeshammer, B., Stuhr, M., Schell, T., Biß, K., Kollipara, L., Meyer, A., Pfenninger, M., Westphal, H., Sickmann, A., Rahnenführer, J. (2017a). DISMS2: a flexible algorithm for direct proteome- wide distance calculation of LC-MS/MS runs. BMC Bioinformatics, 18(1). doi:10.1186/s12859-017-1514-2

Rieder, V., Schork, K. U., Kerschke, L., Blank-Landeshammer, B., Sickmann, A., Rahnenführer, J. (2017b). Comparison and evaluation of clustering algorithms for tandem mass spectra. Journal of Proteome Research, 16(11), 4035–4044. doi:10.1021/acs.jproteome.7b00427

Schell T, Feldmeyer B, Schmidt H, Greshake B, Tills O, Truebano M, Rundle SD, Paule J, Ebersberger I, Pfenninger M (2017). An Annotated Draft Genome for Radix auricularia (Gastropoda, Mollusca). Genome Biol. Evol. 9:585–592. doi: 10.1093/gbe/evx032.

Stuhr M, Meyer A, Reymond CE, Narayan GR, Rieder V, Rahnenführer J, Kucera M, Westphal H, Muhando CA, Hallock P (2018) Variable thermal stress tolerance of the reef-associated symbiont-bearing foraminifera *Amphistegina* linked to differences in symbiont type; Coral Reefs pp1-14. doi:10.1007/s00338-018-1707-9

Stuhr M, Blank-Landeshammer B, Reymond CE, Kollipara L, Sickmann A, Kucera M, Westphal H (2018) Disentangling thermal stress responses in a reef-calcifier and its photosymbionts by shotgun proteomics; Scientific Reports 8:3524. doi: 10.1038/s41598-018-21875-z

Stuhr M, Reyond CE, Rieder V, Hallock P, Rahnenführer J, Westphal H, Kucera M (2017) Reef calcifiers are adapted to episodic heat stress but vulnerable to sustained warming; PLoS ONE 12(7): e0179753

c. Conference contributions (international)

Blank-Landeshammer B., Biß K.,Kollipara L., Rieder V., Stuhr M., Schell T., Zahedi R. P., Pfenninger M., Rahnenführer J., Westphal H., Sickmann A. (2016) Novel approaches in de novo peptide sequencing and proteogenomics as tools to explore uncharted organisms, Poster, 64th Conference on Mass Spectrometry and Allied Topics (ASMS),05-09.06.2016, San Antonio, USA

Blank-Landeshammer B., Schell T., Kollipara L., Biß K., Zahedi R. P., Pfenninger M., Sickmann A. (2017) Talk, Optimized de novo peptide sequencing and proteogenomics workflows to refine ab initio gene prediction, Proteomic Forum, 02.-05.04.2017, Potsdam, Germany,

Stuhr M, Achim Meyer A, Reymond CE, Narayan GR, Rieder V, Rahnenführer J, Kucera M, Westphal H, Hallock P (2018) Are variations in thermal stress tolerance of different Amphistegina species linked to differences in symbiont type? Talk, International Symposium on Foraminifera (FORAMS), 17. - 22.06., Edinburgh, Scotland

Stuhr M, Blank-Landeshammer B, Kollipara L, Reymond CE, Sickmann A, Kucera M, Westphal H (2018) Proteomics allow novel insights into stress responses of foraminifera and their photosymbionts; Poster, FORAMS, 17. - 22.06., Edinburgh, Scotland

Stuhr M, Blank-Landeshammer B, Reymond CE, Sickmann A, Ries JB, Westphal H (2017) Proteomic response of photosymbiont-bearing foraminifera to global impacts on ocean conditions; Talk, 52nd European Marine Biology Symposium, 25. - 29.9., Piran, Slowenien

Stuhr M, Reymond CE, Kucera M, Blank-Landeshammer B, Kollipara L, Sickmann A, Westphal H (2016) Understanding the molecular basis for stress response in foraminifera and symbionts by proteomics; Poster, Batsheva de Rothschild Workshop – Live foraminifera as a new model system for monitoring and reconstructing marine environments, 10.-16.9., Eilat, Israel

Stuhr M, Reymond CE, Kucera M, Blank-Landeshammer B, Kollipara L, Rieder V, Rahnenführer J, Sickmann A, Westphal H (2016) Application of MS-based Proteomics to study Larger Benthic Foraminifera and their responses to environmental changes; Talk, 13th International Coral Reef Symposium (ICRS), 19.-24.6., Honolulu, Hawai´i, USA

Stuhr M, Reymond CE, Kucera M, Westphal H (2015) Acclimatization potential of Amphistegina spp. and their symbionts to long- and short-term thermal stress; Talk, GSA 2015 – Geological Society of America Annual Meeting, 1.-4.11., Baltimore, Maryland, USA

d. Conference contributions (national)

Stuhr M, Reymond CE, Sickmann A, Kucera M, Meyer A, Westphal H (2014) (Reverse) Proteomics as tool for biodiversity research - Applications on Foraminifera; Poster, Computational Molecular Analysis Summer School, 29.9.-3.10., Wilhelmshaven, Germany

Blank-Landeshammer B., Feldmann I. (2016) De-Novo Sequenzierung unbekannter Peptide Talk, Mikromethoden in der Proteinchemie, 23. Arbeitstagung,28-30.06.2016, Dortmund, Germany

Blank-Landeshammer B., Schell T., Kollipara L., Biß K., Zahedi R. P., Pfenninger M., Sickmann A. (2017) Talk, Reverse proteomics as alternative tool to refine gene prediction and genome annotation", Annual Conference of the German Genetics Society (GfG), 26.-28.09.2017, Bochum, Germany

9. Statement on the actions regarding protection and availability of the produced scientific data

Open Access publications (PLoS ONE, Scientific Reports, Frontiers in Marine Science, etc.), raw data of all publications was additionally deposited and is online available in PANGAEA (physiological data from Stuhr et al. 2017 and in press: pangaea.de), The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository, genetic data of symbionts was deposited in the European Nucleotide Archive (sequences from Stuhr et al. 2018: https://www.ebi.ac.uk/ena). All data (also so far unpublished) will furthermore be deposited in the ZMT database and can be made available upon request.

10. List of press and media reports

Press release, 07.02.2018: Proteomik gibt neue Aufschlüsse über Reaktion von Rifforganismen auf Umweltstress; report on websites of ZMT (https://www.leibniz-zmt.de/de/neuigkeiten/nachrichten-aktuelles/archiv-news/foraminferen-proteomik.html) and Wirt oder Gast? Proteomik gibt neue Aufschlüsse über Reaktion von Rifforganismen auf Umweltstress (https://idw-online.de/de/news689764)

ZMT Expedition report: Eilat, Israel, 28.3.16 - 31.3.16 (https://www.leibniz-zmt.de/de/forschung/expeditionen/eilat-israel-maerz-2016.html)

ZMT Newsletter 1/2018: Successful: New proteomics methods (https://www.leibniz-zmt.de/images/content/pdf/Infomaterial/ZMT_Newsletter_1_2018.pdf).

ZMT Newsletter 2/2016: UNIVERSE IN MICROCOSM - ZMT researchers collect tiny organisms – their miniscule cosmos can reveal how the submarine world will respond to climate change (http://lists.zmt-bremen.com/ZMT_Newsletter2-2016.pdf).

11. References

Bailie, J. E. M., C. Hilton-Taylor and S. N. Stuart (2004). IUCN Red List of threatened species. A global speceis assessment. . T. I. S. S. Commision. Gland and Cambridge.

Chi, H., H. Chen, K. He, L. Wu, B. Yang, R.-X. Sun, J. Liu, W.-F. Zeng, C.-Q. Song, S.-M. He and M.-Q. Dong (2013). "pNovo+: De Novo Peptide Sequencing Using Complementary HCD and ETD Tandem Mass Spectra." Journal of Proteome Research **12**(2): 615-625.

Cox, J. and M. Mann (2007). "Is proteomics the new genomics." Cell **130**(3): 395-398.

Darling, K. F. and C. M. Wade (2008). "The genetic diversity of planktic foraminifera and the global distribution of ribosomalRNA genotypes." Marine Micropaleontology **67**(3-4): 216-238.

Dorfer, V., P. Pichler, T. Stranzl, J. Stadlmann, T. Taus, S. Winkler and K. Mechtler (2014). "MS Amanda, a universal identification algorithm optimized for high accuracy tandem mass spectra." J Proteome Res **13**(8): 3679-3684.

Feldmeyer, B., C. Wheat, N. Krezdorn, B. Rotter and M. Pfenninger (2011). "De novo assembly of a non-model speceis transcriptome using Illumina data: assembly performance dependent on method used upon the snail transcriptome (Radix balthica, Pullmonata, Basommatophora)." BMC Genomics **12**: 317.

Frank, A. and P. Pevzner (2005). "PepNovo: De Novo Peptide Sequencing via Probabilistic Network Modeling." Analytical Chemistry **77**(4): 964-973.

Guillou, L. and D. Bachar (2013). "The Protist Ribosomal Reference Database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy." Nucleic Acids Research **41**(D1): D597-D604.

Hallock, P., L. B. Forward and h. H.J (1986). "Environmental influence of test shape in Amphistegina." Jorunal of Foraminiferal Research **16**: 224-231.

Haun, T., M. Salinger, A. Pachzelt and M. Pfenninger (2012). "On the process of shaping small scale population structure in Radix balthica (Linnaeus 1758)." Malacologia **55**: 219-233.

Jing Zhang, Lei Xin, Baozhen Shan, Weiwu Chen, Mingjie Xie, Denis Yuen, Weiming Zhang, Zefeng Zhang, Gilles A. Lajoie and B. Ma (2012). "PEAKS DB: De Novo Sequencing Assisted Database Search for Sensitive and Accurate Peptide Identification." Mol Cell Proteomics **11**(4): M111.010587.

John R. Yates, I., Jimmy K. Eng, Karl R. Clauser and A. L. Burlingame (1996). "Search of Sequence Databases with Uninterpreted High-Energy Collision-Induced Dissociation Spectra of Peptides." J Am Soc Mass Spectrom **7**: 1089-1098.

Ma, B. (2015). "Novor: real-time peptide de novo sequencing software." J Am Soc Mass Spectrom **26**(11): 1885-1894.

Magnus, P. and D. A. M. (2012). "Molecular phylogenetics by direct comparison of tandem mass spectra." Rapid Communications in Mass Spectrometry **26**(7): 728-732.

May, R. M. (1988). "How many species are there on earth?" Science (New York, N.Y.) **247**: 1441-1449.

Pauls, S., C. Nowak, M. Balint and M. Pfenninger (2012). "The impact of global climate change on genetic diversity within populations and species." Molecular Ecology **22**: 925-946.

Pawlowksi, J. and M. Holzmann (2002). "Molecular phylogeny of Foraminifera a review." European Journal of Protistology **38**(1): 1-10.

Pereia, J. C. and R. Chaves (2011). "An efficient method for genomic DNA extraction from different molluscs species
" Int J Mol Sci **12**(11): 8086-8095.

Pfenninger, M., M. Cordellier and B. Streit (2006). "Comparing the efficacy of morphologic and DNA-based taxonomy in the freshwater gastropod genus Radix (basommatophora, Pulmonata)." BMC Evolutionary Biology **6**: 100.

Pfenninger, M. and K. Schwenk (2007). "Cryptic animal species are homogenously distributed among taxa and biogeographical regions." BMC Evolutionary Biology **7**(121).

Teixeira, T., N. Rainha, J. Rosa, E. Lima and J. Baptistsa (2012). "Molluscicidal activity of crude water and hexane extracts of Hypericum speceis to snails (Radix peregra)." Environmental Toxicology and Chemisrty **31**: 345-356.

Venne, A. S., F. N. Vogtle, C. Meisinger, A. Sickmann and R. P. Zahedi (2013). "Novel highly sensitive, specific, and straightforward strategy for comprehensive N-terminal proteomics reveals unknown substrates of the mitochondrial peptidase Icp55." J Proteome Res **12**(9): 3823-3830.

Wisniewski, J. R., A. Zougman, N. Nagaraj and M. Mann (2009). "Universal sample preparation method for proteome analysis." Nat Meth **6**(5): 359-362.

Yoshida, M., Y. Ishikura, T. Moritaki, E. Shoguchi, K. K. Shimizu, J. Sese and A. Ogura (2011). "Genome structure analysis of molluscs revealed whole genome duplication and lineage specific repeat variation." Gene **483**: 63-71.