Bernhard Blank-
Landeshammer

Development and application of proteogenomics methods to refine genome annotations and detect single amino acid variants

SHAKER
VERLAG

Development and application of proteogenomics methods to refine genome annotations and detect single amino acid variants

Zur Erlangung des akademischen Grades eines

Dr. rer. nat.

von der Fakultät Bio- und Chemieingenieurwesen
der Technischen Universität Dortmund
genehmigte Dissertation

vorgelegt von

Bernhard Blank-Landeshammer, MSc

aus

Kirchdorf a.d. Krems, Österreich

Tag der mündlichen Prüfung: 18.12.2020

1. Gutachter/-in: Prof. Dr. Sickmann
2. Gutachter/-in: Prof. Dr. Nett

Dortmund 2020

Berichte aus der Biologie

Bernhard Blank-Landeshammer

Development and application of proteogenomics methods to refine genome annotations and detect single amino acid variants

D 290 (Diss. Technische Universität Dortmund)

Shaker Verlag
Düren 2021

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at http://dnb.d-nb.de.

Zugl.: Dortmund, Technische Univ., Diss., 2020

Printed in Germany.

ISBN 978-3-8440-7854-1
ISSN 0945-0688

Shaker Verlag GmbH • Am Langen Graben 15a • 52353 Düren
Phone: 0049/2421/99011-0 • Telefax: 0049/2421/99011-9
Internet: www.shaker.de • e-mail: info@shaker.de

Abstract

Genome annotation procedures can benefit from information provided by proteomics in various ways. Detection of sample- or patient specific alterations of protein sequences in the form of single amino acid variants (SAAVs) require dedicated analysis and validation procedures. In this work, proteogenomics methods were established that enable proteomics analysis of organisms void of reference sequence databases and that can universally be used to refine genome annotations based on protein level information as well as for identification of SAAVs.

First, a *de novo* peptide sequencing method was established to improve database-free peptide identification. By integration of multiple algorithms, the reliability of the results could be improved and the number of identified sequences at stringent conditions was tripled compared to a reference analysis. The applicability of *de novo* peptide sequencing to differentiate species of foraminifera (unicellular protists) based on their proteome was assessed and a homology-based strategy was employed to quantify the proteomic response of the foraminifera *A. lobifera* to heat stress.

Next, the proteomic landscape of the freshwater snail *R. auricularia* was charted and a combination of *de novo* peptide sequencing and searches against the newly sequenced genome enabled an automated refinement of the annotation. This led to the additional identification of almost one thousand genes and major alterations of the underlying gene models. The genome annotation of the model fungus *S. macrospora* was subjected to similar refinements. Remarkably, *de novo* peptide sequencing enabled to our knowledge the first ever confirmation of A-to-I RNA editing in fungi on the protein level.

Finally, a proteogenomics approach was employed to analyze liver metastases retrieved from patients suffering from colorectal cancer. This enabled the identification of multiple SAAVs such as the prognostic KRAS G12V mutation. These discoveries were further validated and quantified by a targeted MS assay using of stable-isotope labeled standard (SIS) peptides.

Abstract (in German)

Genomannotation stellt in der Form von Protein Sequenzdatenbanken die Basis für LC-MS basierte Proteomanalysen zur Verfügung. Proteomische Daten können umgekehrt aber auch dazu beitragen, diese Annotationsvorgänge zu verbessern, indem Genmodelle verfeinert und Kodierungspotentiale besser bewertet werden können. In dieser Arbeit wurden proteogenomische Methoden etabliert um Proteomanalysen von Organismen mit fehlender Referenzdatenbank durchzuführen, die Annotation von Genomen mithilfe von Proteomdaten zu verbessern und Einzel-Aminosäureaustausche (SAAVs) sicher zu identifizieren.

Zuerst wurde eine kombinatorische *de novo* Peptidsequenziermethode etabliert, um eine Datenbank-freie Identifikation von Peptiden zu ermöglichen. Durch die Integration mehrerer Algorithmen konnte die Zuverlässigkeit der Ergebnisse verbessert werden und die Anzahl der identifizierten Peptide bei relativ stringenten Kriterien (d.h. 5% FDR) konnte im Vergleich zum besten Einzel-Algorithmus verdreifacht werden. Die Anwendbarkeit dieser *de novo* Sequenziermethode wurde anhand eines Speziesvergleichs von Foraminiferen (einzellige Protisten) getestet und mittels einer Homologie-basierten Suchstrategie wurde die Proteomantwort der Foraminiferen-Spezies *A. lobifera* auf Hitzestress quantifiziert.

Ferner wurde das Proteom der Frischwasserschnecke *R. auricularia* charakterisiert und eine Kombination aus *de novo* Sequenzierung und Datanbanksuchen gegen das translatierte Genom dazu genutzt, diese Informationen automatisiert in die Annotation einfließen zu lassen. Dies erlaubte die Identifikation von annähernd 1,000 Genen und eine beträchtliche Abänderung der zugrundeliegenden Genmodelle. Dieser Arbeitsablauf wurde weiter dazu genutzt, die Genomannotation des als Modelorganismus genutzten Pilzes *S. macrospora* zu verbessern. Hier ermöglichte der *de novo* Sequenzier-Ansatz den unseres Wissens nach ersten Nachweis von A-zu-I RNA Editierung in Pilzen auf Proteinebene.

Schließlich wurde ein proteogenomischer Ansatz dazu verwendet, Gewebeproben von Lebermetastasen zu analysieren, die Patienten mit kolorektalem Karzinom entnommen wurden. Dabei wurden mehrere SAAVs auf Proteinebene nachgewiesen, unter anderem die KRAS G12V Mutation. Eine zielgerichtete MS Untersuchung mit stabil isotopenmarkierten Standardpeptiden wurde etabliert um den Mutationsgrad zu quantifizieren.

Table of contents