Computational methods and graphical models for integrative proteogenomics

Dissertation zur Erlangung des Grades
eines Doktors der Naturwissenschaften (Dr. rer. nat.)
am Fachbereich Mathematik und Informatik
der Freien Universität Berlin

von

Franziska Zickmann

Berlin
Februar 2015

Erstgutachter:	PD Dr. Bernhard Y. Renard
Zweitgutachter:	Prof. Dr. Oliver Kohlbacher
Tag der Disputation:	29.04.2015

Abstract

Proteogenomics describes the integration of genomic, transcriptomic, and proteomic data. The combination of this multi-omics information offers unprecedented possibilities for more accurate and sample-specific gene and protein identification. Further, the advent of high-throughput technologies has led to a wealth of studies aiming at a deeper understanding of protein function and interaction. Hence, methods analyzing proteogenomic data, and particularly integrating various data types, are strongly demanded.

In this thesis, we present new proteogenomic approaches for the integration of next-generation sequencing and mass spectrometry data in form of DNA and RNA-Seq and tandem mass spectra. These contributions can be divided into three main projects: First, we developed the method GIIRA (Gene Identification Incorporating RNA-Seq data and Ambiguous reads) for the construction of gene models and transcript prediction based on RNA-Seq data. GIIRA analyzes RNA-Seq mappings on prokaryotic and eukaryotic reference genomes in order to identify expressed genes on the reference. Unlike other RNA-Seq analysis methods, it does not exclude ambiguously mapping reads, but rather explicitly includes all mappings to perform a more comprehensive prediction. It first extracts candidate regions based on the complete RNA-Seq mapping and represents all connections of reads and candidates in a network. This network is optimized in a maximum-flow approach to resolve ambiguous mappings and identify the most likely origin of each read. The optimization is realized by an integer linear program formulation. In several experiments we show that GIIRA is well suited for RNA-Seq-based gene identification and improves the accuracy of existing methods. For instance, on an *Escherichia coli* data set GIIRA showed up to 15% improved identification accuracy in comparison to other prediction methods.

The second main project builds on the output of GIIRA and post-processes gene prediction results in order to improve prediction accuracy. We developed IPred (Integrative gene Prediction), a computational approach that explicitly combines the results of *ab initio* gene finders and evidence-based methods. *Ab initio* approaches employ machine learning techniques and predict genes exclusively based on a given reference sequence. Hence, their results are accurate for standard gene structures, but they are not sample-specific. Thus, IPred provides an automated simplistic framework to integrate the results of varying evidence-based predictions to *ab initio* identifications. Thereby, it excludes false positives and allows support

for sample-specific mutations. Predictions combined by IPred show improved accuracy in comparison to results from single method gene finders and other combination methods. In particular the specificity of single method results is increased by up to 30%.
The third project extends the former two methods and combines RNA-Seq-based predictions with tandem mass spectrometry. We introduce MSProGene (Mass Spectrometry and RNA-Seq-based Protein and Gene Identification), a new proteogenomic method that performs protein identification beyond reference protein databases or six-frame translations. It constructs customized transcript databases (for instance using GIIRA or IPred) and analyzes peptide spectrum matches with the help of a network representation. In particular, MSProGene explicitly resolves shared peptides for protein inference using RNA-Seq information in a linear program optimization. Resulting peptide spectrum matches are controlled by an expectation-maximization-based false discovery rate. We performed an exhaustive comparison to reference dependent and independent proteogenomic approaches and demonstrate that MSProGene facilitates a reliable database independent prediction on gene and protein level and additionally identifies novel genes. For instance, on a *Litomosoides sigmodontis* data set it identified twenty times as many proteins verified by BLAST search than a standard six-frame analysis.
With these projects we developed new methods for automated and accurate proteogenomic analysis. The introduced approaches successfully integrate genomic data with RNA-Seq and mass spectrometry experiments to enable a better understanding of protein function and interaction.

Zusammenfassung

Das Feld der Proteogenomik verbindet genomische, transkriptomische, und proteomische Daten und ermöglicht so die Kombination von Genexpressionsinformationen für akkuratere und experimentspezifische Gen- und Proteinidentifikation. Zusätzlich hat auch die Entwicklung von Hochdurchsatzverfahren zu einer Vielzahl von Studien geführt, mit dem Ziel, ein besseres Verständnis von Proteinfunktion und -interaktion zu erlangen. Daher ist es sehr wichtig, automatisierte Methoden für die Analyse von proteogenomischen Daten, insbesondere der Integration von verschiedenen Datentypen, bereitzustellen.
In dieser Doktorarbeit stellen wir proteogenomische Ansätze für die Integration von Daten aus der DNA- und RNA-Sequenzierung und der Tandemmassenspektrometrie vor. Die Beiträge der Arbeit können in drei Hauptprojekte unterteilt werden: Erstens, die Entwicklung der Methode GIIRA (Gene Identification Incorporating RNA-Seq data and Ambiguous reads) für die Erstellung von Genmodellen und die Vorhersage von Transkripten basierend auf RNA-Sequenzierung. Dazu analysiert GIIRA die auf prokaryotischen und eukaryotischen Referenzen alignierten RNA-Sequenzen um expremierte Gene auf der Referenz zu identifizieren. Im Gegensatz zu anderen Methoden zur Analyse von RNA-Sequenzierungsdaten entfernt GIIRA dabei nicht die mehrdeutig alignierten Sequenzen, sondern verwendet stattdessen explizit alle Alignments um eine umfassendere Vorhersage treffen zu können. Hierzu werden zunächst Kandidatenregionen extrahiert, basierend auf dem kompletten RNA-Alignment. Anschließend werden alle Verbindungen von RNA-Sequenzen und Kandidaten in einem Netzwerk repräsentiert. Dieses Netzwerk wird mit einem Maximum-Flow Algorithmus optimiert, um für jede mehrdeutige Alignierung die wahrscheinlichste Ursprungsposition zu bestimmen. Dabei basiert die Optimierung auf der Formulierung und Lösung eines Linearen Programms. Wir zeigen in verschiedenen Experimenten, dass GIIRA sehr gut zur Genidentifizierung basierend auf RNA-Sequenzierung geeignet ist und die Genauigkeit bestehender Methoden übertrifft. Beispielsweise zeigt GIIRA auf einem *Escherichia coli* Datensatz bis zu 15% höhere Vorhersagegenauigkeit als andere Genidentifizierungsmethoden.
Das zweite Hauptprojekt baut auf den Ergebnissen von GIIRA auf und prozessiert Genvorhersagen, um deren Genauigkeit weiter zu verbessern. Dazu entwickelten wir IPred (Integrative gene Prediction), eine Methode, die explizit Resultate von *ab initio* Genidentifizierungsmethoden und evidenzbasierten Genidentifizierungs-

methoden verbindet. *Ab initio* Ansätze benutzen Maschinelles Lernen um Gene direkt auf gegebenen Referenzsequenzen vorherzusagen. Damit sind sie akkurat für bekannte Genstrukturen, aber nicht experimentspezifisch. Daher bietet IPred eine automatisierte Methode um die Resultate von evidenzbasierten Identifizierungsmethoden mit *ab initio* Vorhersagen zu vereinen. Dabei entfernt die Methode falsche Identifikationen und erlaubt die Detektion von experimentspezifischen Mutationen. Die kombinierten Vorhersagen von IPred zeigen verbesserte Genauigkeit, sowohl im Vergleich zu Vorhersagen von einzelnen Genidentifizierungsmethoden als auch anderen Kombinationsmethoden. Insbesondere die Spezifität konnte um bis zu 30% verbessert werden.
Das dritte Projekt erweitert die vorherigen zwei Methoden und kombiniert RNA-Sequenzierung mit Tandemmassenspektrometrie. Wir entwickelten die neue proteogenomische Methode MSProGene (Mass Spectrometry and RNA-Seq-based Protein and Gene Identification), welche Proteinidentifikation unabhängig von Referenzproteindatenbanken und six-frame Translationen durchführt. MSProGene generiert maßgeschneiderte Transkriptdatenbanken (zum Beispiel mit Hilfe von GIIRA und IPred) und analysiert Peptididentifikationen mit Hilfe einer Netzwerkdarstellung. Insbesondere integriert MSProGene dabei RNA-Sequenzierungsdaten um mit Hilfe einer linearen Optimierung mehrdeutig zugeordnete Peptide zum korrekten Protein zuzuordnen. Die resultierenden Peptididentifikationen unterliegen einer Qualitätskontrolle basierend auf einem Expectation-Maximization Algorithmus. In einem umfangreichen Vergleich zu referenzabhängigen und referenzunabhängigen, proteogenomischen Analysemethoden zeigen wir, dass MSProGene eine verlässliche datenbankunabhängige Identifikation von Genen und Proteinen ermöglicht und zusätzlich neue Gene detektiert. Beispielsweise identifiziert MSProGene auf einem *Litomosoides sigmodontis* Datensatz zwanzig mal so viele BLAST verifizierte Proteine wie eine standard six-frame Analyse.
Mit diesen Projekten stellen wir neue Methoden für die automatisierte und akkurate proteogenomische Analyse bereit. Die vorgestellten Methoden integrieren erfolgreich genomische Daten mit RNA-Sequenzierungs- und Massenspektrometrieexperimenten und tragen so zu einem besseres Verständnis von Proteinfunktion und -interaktion bei.

Acknowledgements

First and foremost, I want to thank my supervisor Bernhard Renard for his advise and support during the last years. I am very thankful that he always gave me the chance to pursuit own ideas and at the same time had helping input at hand when I struggled.
Further, I would like to thank Oliver Kohlbacher for agreeing to review this thesis.
A deep thank you also to several people working at the Robert Koch-Institute: Most of all to Wojtek Dabrowski, who was not only the godfather of solving server issues and partner in inspiring discussions, but more importantly also became a good friend. Further, I want to thank Sébastien Calvinac-Spencer for an enjoyable collaboration that allowed me to stay in touch with phylogenetics, next to my main thesis topic.
I want to thank the students who worked with me in various projects, in particular Sven Giese for his work on specificity control for read mappings, Annkathrin Bressin for the implementation of an RNA-Seq simulator, Jakob Schulze for his work on RootAnnotator, Stephan Knorr for his dedicated master thesis on influences of target and decoy databases, Paul Schäpe for implementing the phylogenetic analysis pipeline for viral families, and finally Yoonjeong Cha for her work on the pipeline to access virus evolution. I enjoyed working with you and thank you for your invaluable contribution to all these side projects throughout my PhD.
Of course one of the biggest thanks goes to my colleagues in the NG4 Bioinformatics group. I hope that I can always work in a group with such great people and nice atmosphere. Particularly I want to thank Martin, first for his valuable contribution to the GIIRA publication and second for being my best coffee break partner and oldest ally in the PhD student force in our group. A big thank you also to my fellow PhD students Martina and Kathrin for all the good conversations and lots of funny moments, Mathias for great gaming sessions, and also to Vitor and the postdocs Carlus and Robert and all alumni for making these four years really enjoyable, not only at work, but also afterwards.
Special thanks also to my friends for being such wonderful people and for the amazing time. Last but not least, I want to express deep gratitude to my parents and grandparents and particularly to Jochen for their steady love and support. Knowing that I can always rely on you is an invaluable source of strength for me.

Abbreviations

Throughout the thesis we use different abbreviations common in the genomic as well as in the proteomics field. The following table presents a list with abbreviations and a short explanation for each term.

List of abbreviations

Abbreviation	Explanation
AUC	Area Under the Curve
CDS	Coding Sequence
DNA	Deoxyribonucleic Acid
EMBOSS	European Molecular Biology Open Software Suite
ENCODE	Encyclopedia of DNA Elements at UCSC
EST	Expressed Sequence Tag
FDR	False Discovery Rate
GTF	Gene Transfer Format
GUI	Graphical User Interface
HMM	Hidden Markov Model
MS	Mass Spectrometry
MS/MS	Tandem Mass Spectrometry
NCBI	National Center for Biotechnology Information
NGS	Next-Generation Sequencing
ORF	Open Reading Frame
PSM	Peptide Spectrum Match
PTM	Post-translational Modification
RNA	Ribonucleic Acid
ROC	Receiver Operating Characteristic
rRNA	Ribosomal Ribonucleic Acid
SAM	Sequence Alignment/Map format
SNP	Single Nucleotide Polymorphism
UTR	Untranslated Region
VCF	Variant Call Format

Contents

1. Introduction

1.1. Proteogenomics

High-throughput technologies in genomics and proteomics have facilitated ongoing advances in the analysis of the mechanisms of gene expression and the function and interaction of proteins. Next-generation sequencing (NGS) in form of DNA and RNA sequencing (RNA-Seq) enables the assembly of genomic sequences (Metzker, 2009) and measures the transcriptome as an intermediate step during gene expression (Wang et al., 2009). In proteomics, mass spectrometry (MS) allows the identification and quantification of proteins that were expressed (Nesvizhskii et al., 2007; Nilsson et al., 2010). These fields are directly connected and allow mutual verification since genes encoded on the genomic sequence are first transcribed to RNA sequences and then translated to proteins (Crick et al., 1970). Particularly in the area of genome annotation, which describes the prediction of expressed regions and their regulation system on the genome, this correlation has a high impact (Ansong et al., 2008). Studies focusing on the analysis of genomic DNA sequences can predict genes, but only transcriptome analysis can determine if these genes are actually transcribed under a certain condition. Moreover, only proteomic measurements can validate if a gene is indeed translated to a protein. Thus, in the field of proteogenomics this multi-omics data is combined to allow a deeper understanding in genome analysis (Ansong et al., 2008; Castellana and Bafna, 2010; Nesvizhskii, 2014). In doing so, proteogenomics targets numerous aspects of genome annotation, such as the detection of novel genes, the verification of predicted genes, the search and validation of correct start and stop positions and exon boundaries, the analysis of post-translational modifications, and the analysis of splice variants.
However, the analysis and integration of data in proteogenomic studies is challenging. With decreasing sequencing costs a plethora of proteogenomic data is generated, which demands efficient methods to analyze and integrate measurements from different instruments. From early on, proteogenomic studies have been focused on better genome annotation using mass spectrometry in addition to the standard sequencing-based annotation (Yates III et al., 1995; Link et al., 1997). Here, genomic sequences are six-frame translated to create an amino acid sequence database for spectra search. Resulting spectra support for specific regions then indicates unannotated genes. In addition, also previously predicted gene models are translated for spectra search to verify and revise existing genes (Küster et al.,

2001; Jaffe et al., 2004). Further, expressed sequence tag (EST) databases are used to include information from expressed genomic regions for improved identification (Mann, 1996; Choudhary et al., 2001).
The success of proteogenomics to improve the annotation even of well-studied model organisms has promoted numerous studies using proteomics data for verification of genomic analysis (Kalume et al., 2005; Fermin et al., 2006; Tanner et al., 2007; Kelkar et al., 2011; Safavi-Hemami et al., 2014). Efforts are dedicated to comparative studies of multiple bacterial genomes (Gupta et al., 2008), revising model organisms (Oshiro et al., 2002; Castellana et al., 2008) or identifying post-translational modifications (Gupta et al., 2007). Recently, also standardized workflows have become available for automated proteogenomic analysis (Kumar et al., 2013; Jagtap et al., 2014).
The advances in genomic annotation have facilitated the construction of comprehensive databases, such as dbSNP (Sherry et al., 2001) including known single nucleotide polymorphisms (SNPs) or ENCODE (Consortium, 2004), a database collecting information on the human genome. Proteogenomic approaches use these databases, for instance by creating more sophisticated genomic databases for spectra search by including SNPs to existing protein annotations (Ahn et al., 2013; Krug et al., 2014). Since these efforts are database-dependent and rely on *a priori* information, studies in addition frequently employ transcriptome data. Although also the integration of EST libraries improves proteogenomic analysis, ongoing advances in high-throughput sequencing shifted the focus to the integration of RNA-Seq information. EST sequences are short and do not represent all tissues or cell types (Küster et al., 2001; Schurch et al., 2014). In contrast, RNA-Seq measures the complete transcriptome and in addition advanced sequencing techniques provide high coverage information. Thus, various studies include RNA-Seq evidence in proteogenomic analyses (Ning and Nesvizhskii, 2010; Fanayan et al., 2013; Mohien et al., 2013; Wang et al., 2014). Recently, also metabolomics and interactome data are integrated in proteogenomic workflows, further extending the spectrum of possible sources of evidence (Wang and Zhang, 2014; Meierhofer et al., 2014).
An exemplary overview of the correlation and the measurements of the three primarily integrated research fields genomics, transcriptomics, and proteomics is shown in Figure 1.1. The figure illustrates that different experiments can be combined, but that also the analysis within each research field is of importance. This is detailed in the next sections, where we first introduce concepts for gene prediction, which is followed by an overview of protein identification. These two tasks are key challenges in proteogenomic workflows since the basis for all integrative approaches is the combination of comprehensive genomic annotation with sophisticated protein identification methods.

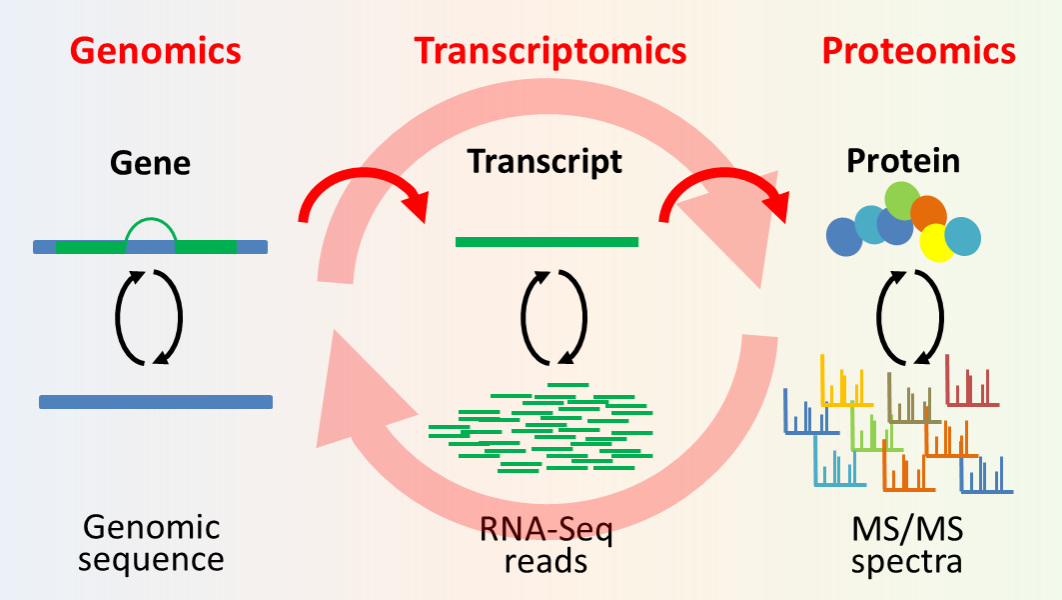

Figure 1.1.: Exemplary illustration of the correlation of different fields and experiments connected by the central dogma of gene expression and integrated in proteogenomics. Transcriptomic measurements, in this example in form of RNA-Seq reads, and proteomic measurements in form of tandem mass spectra (MS/MS spectra), benefit the annotation of genes present on genomic sequences. Simultaneously, genomic and transcriptomic information can be used to identify translated proteins.

1.2. Gene prediction

In proteogenomic studies, gene finding is often a central focus to provide a basis for the construction of spectra search databases. Even sophisticated proteogenomic approaches need a comprehensive and meaningful interpretation of genomic sequences, either in form of complete gene models, transcribed sequences or simplistic open reading frame (ORF) predictions. However, the annotation of expressed regions and their structures is a challenging research area (Claverie, 1997; Yu et al., 2014). As a consequence, numerous studies focus on revealing the structure of genes and their controlling mechanisms (Schrimpe-Rutledge et al., 2012; Wang et al., 2012; Wijaya et al., 2013; Fawal et al., 2014).
Often, ORF prediction or genomic six-frame translation is the first step to analyze unannotated organisms. Widely used programs for ORF prediction are for instance *getorf* from the EMBOSS package (Rice et al., 2000) or the *ORF finder* from NCBI (Wheeler et al., 2003). However, these simplistic prediction methods are not suited to predict complex gene structures or regard organism specific characteristics, such

as splicing events, the presence of pseudo genes, non-standard coding schemes, or sample-specific variations. As a consequence, much effort is dedicated to more sophisticated annotation and the construction of reliable gene models. The resulting gene prediction methods can be categorized into *ab initio* as well as evidence-based and comparative gene finders (Goodswen et al., 2012).

Ab initio gene finders identify genes exclusively based on genomic sequences. They predict start and stop codons of ORFs and in case of eukaryotes identify intron-exon structures indicated by known splice sites (Goodswen et al., 2012). Typically, these approaches are based on statistical or machine learning techniques, such as Hidden Markov Models (HMMs), and they require training data to evaluate the probability for each gene and gene structure (Sleator, 2010). The training set is used to learn general characteristics, such as the distribution and frequency of GC nucleotides. Then, the learnt features are expressed in a model that is used for prediction on the data set of interest (Brent, 2007). Popular methods include GeneMark (Lukashin and Borodovsky, 1998; Besemer et al., 2001) and GLIMMER3 (Delcher et al., 2007) for prokaryotic gene prediction, and GlimmerHMM (Majoros et al., 2004), SNAP (Korf, 2004), and GeneMark.hmm (Lomsadze et al., 2005) for eukaryotic gene prediction.

In contrast to *ab initio* methods, evidence-based and comparative gene finders make use of additional information to identify genes and their structures. Sources of evidence include EST libraries, mRNA, or protein sequences. This additional information is compared to the genome of interest to identify regions showing similarity to the given evidence (Wei and Brent, 2006; Savidor et al., 2006; Allen and Salzberg, 2005).

Comparative methods use annotations on closely related species for gene prediction on the sequence of interest, based on the assumption that general structures, such as introns or coding sequences, are subject to similar evolutionary selective pressures (van Baren et al., 2007).

Hybrid approaches, such as AUGUSTUS (Stanke et al., 2006) and JIGSAW (Allen and Salzberg, 2005), combine *ab initio* predictions with additional evidence. This strategy allows a more accurate verification of predicted genes (Guigó et al., 2006).

A class of methods related to hybrid approaches are prediction combination programs. These methods combine the output of different gene prediction strategies to complement the strengths of single method predictions (Yok and Rosen, 2011; Ederveen et al., 2013). For instance, evidence-based predictions are used to validate gene models predicted by *ab initio* approaches to improve the overall prediction accuracy (Pavlović et al., 2002; Elsik et al., 2007; Haas et al., 2008).

Independently of the strategy used for gene prediction, the resulting identified gene models are further processed to provide suitable databases for proteogenomic analysis. As detailed in the next section, often subsequent proteomic searches strongly depend on the quality of these databases. Thus, obtaining reliable gene

predictions is a key challenge in proteogenomics and prediction strategy as well as used software have to be selected carefully, with regard to the specific data set and research focus.

1.3. Protein identification

In proteogenomics, genomic data is combined with proteomic information. Thus, the search of reliable gene models is only one step in a proteogenomic workflow that is typically followed by the integration of shotgun proteomic data. Here, tandem mass spectrometry (MS/MS) is established as the method of choice for high-throughput proteogenomic analysis (Nesvizhskii, 2014; Branca et al., 2014). In recent years, instruments for shotgun MS/MS experiments have become more and more advanced, thereby providing deeper coverage of peptides and proteins. This resulted in significant progress towards the identification of complete proteomes (Nesvizhskii, 2010; Wilhelm et al., 2014).
In a typical MS/MS experiment a protein sample is first digested to peptide sequences since mass spectrometers are not as sensitive in detecting proteins as they are in peptide detection. The peptides are then ionized and scanned in the mass spectrometer that measures their mass to charge (m/z) ratio and signal intensity (Steen and Mann, 2004). In the end, an MS/MS analysis of a protein sample results in thousands of tandem mass spectra (MS/MS spectra), where each spectrum is supposed to represent one peptide sequence. Based on the spectra, first the sequence of the corresponding peptide needs to be reconstructed and finally the original proteins that gave rise to the peptides need to be inferred (Nesvizhskii et al., 2007).
The first objective, referred to as peptide identification, is most commonly realized by comparing the experimental measured spectra against theoretical spectra constructed from existing protein reference databases (Nesvizhskii, 2010). Typically, various criteria, such as the charge state of the peptide and a tolerance mass window, influence the peptide search and specify possible peptide candidates for a spectrum. Resulting identified peptide spectrum matches (PSMs) are scored based on the similarity of experimental and theoretical spectrum (Steen and Mann, 2004). Numerous methods performing this database-driven peptide search have been developed, which differ in their search strategies and PSM scoring methods. Popular search engines include SEQUEST (Eng et al., 1994), X!Tandem (Craig and Beavis, 2004), MASCOT (Perkins et al., 1999), and MSGF+ (Kim and Pevzner, 2014).
Other approaches to peptide identification are spectral library searches or *de novo* sequencing (Nesvizhskii, 2010). Spectral libraries include experimental measured spectra that have been associated to peptide sequences in previous experiments (Frewen et al., 2006; Lam et al., 2007). Novel unknown spectra are compared to

those previously identified, using methods such as X!Hunter (Craig et al., 2006) or SpectraST (Lam et al., 2008). Spectral library searches are usually faster and more sensitive than standard database searches, but they require peptides and also post-translational modifications (PTMs) to be measured in previous experiments (Lam et al., 2007).
In contrast, *de novo* approaches are independent of *a priori* derived protein databases or spectral libraries. These methods assemble the peptide sequence based on the differences between m/z ratios of the peaks observed in MS/MS spectra. Possible sequences explaining the observed differences are enumerated and the best matching series of amino acids is associated to the spectrum (Dancik et al., 1999; Seidler et al., 2010). Although several methods for *de novo* sequencing are available, for instance including PEAKS (Ma et al., 2003), PepNovo (Frank and Pevzner, 2005), or Vonode (Pan et al., 2010), currently results often need to be curated manually. Thus, current methods are usually not practical for standard shotgun proteomic analysis (Nesvizhskii, 2010).
Since in general peptide identification is a challenging task and results of different search engines can vary a lot, several methods aim at post-processing and integrating results of peptide identification engines to improve PSM accuracy (Käll et al., 2007; Nahnsen et al., 2011; Shteynberg et al., 2013). The evaluation of PSMs is one of the key challenges in proteomic experiments since they need to be carefully analyzed in order to exclude false positive identifications. The accepted standard for MS/MS quality control is the target-decoy approach to calculate a search-specific false discovery rate (FDR) (Bradshaw et al., 2006). Here, not only a database containing the protein sequences of interest (target) is provided for spectra search, but also a database containing artificial sequences (decoy), for instance derived by reversing or shuffling target proteins. Based on the assumption that a decoy identification is similarly likely as a false identification, the FDR can be estimated (Benjamini and Hochberg, 1995) and a predefined FDR threshold can be used as a quality filter on the original PSMs. The target-decoy approach is easily implemented for peptide identification, but it increases the spectra search time due to larger search database sizes. Thus, also decoy-free approaches for FDR calculation have been proposed that estimate the target and decoy distribution among a set of PSMs (Keller et al., 2002; Renard et al., 2010).
The second objective in typical shotgun proteomic workflows is the inference of the original proteins present in the sample, based on the identified peptides (Nesvizhskii and Aebersold, 2005; Huang et al., 2012). Protein inference is a difficult task, particularly due to so called shared peptides that not only map exclusively to one protein, but are present in multiple proteins with homologous subsequences. Thus, these peptides cannot be assigned in a straightforward way and the choice of proteins that are actually present in a sample is challenging (Huang et al., 2012; Li and Radivojac, 2012). Further, not all peptides can be measured equally well in the mass

spectrometer (Sanders et al., 2007) and additionally not all PSMs pass the quality thresholds in the PSM evaluation. Thus, usually proteins are not homogeneously covered, but instead their support can be limited. Proteins only supported by one single peptide hit are often denoted as one-hit wonders, and it is the accepted standard to exclude these proteins due to lack of reliability (Huang et al., 2012). Existing approaches for protein inference group proteins based on their shared peptide support and use parsimonious strategies to infer the smallest possible set of proteins explaining the observed peptides (Nesvizhskii et al., 2003; Serang et al., 2010). Other methods use additional information for protein inference, for instance in form of gene function networks and protein interaction networks (Li et al., 2009b; Ramakrishnan et al., 2009; Gerster et al., 2010).

1.4. Open problems

Although much effort is dedicated to the design of comprehensive proteogenomic studies, the field remains challenging and several key questions are not yet completely solved. For instance, the necessity to define and construct suitable databases for spectra search is one of the key problems in proteogenomics. In particular for unannotated organisms with unknown reference proteins, standard proteomic search techniques that require reference databases are not applicable. Even methods for error-tolerant database search cannot overcome this problem entirely (Renard et al., 2012), because variations between organisms might be too large to use related organisms as references. Once genomic sequences are available, six-frame translation can be used to create an initial search database that reflects potential ORF regions. However, a drawback of using complete six-frame translations is the artificial increase of the database used for spectra search, which introduces a bias in peptide identification and also increases the search time (Reiter et al., 2009). Further, simplistic ORF prediction and six-frame translation do not cover complex gene structures with splicing events, as they often occur in eukaryotic genomes (Nesvizhskii, 2014; Branca et al., 2014). Thus, methods that go beyond ORF prediction and six-frame translation are desirable.
With new sequencing technologies, such as RNA-Seq, simplistic genomic ORF prediction can be extended by including additional information. For instance, *de novo* transcript assembly with methods such as Trinity (Grabherr et al., 2011) can be used to assemble RNA-Seq reads to longer continuous sequences that serve as a basis for translation. However, RNA-Seq driven transcript assembly is a challenging problem in itself, and the resulting transcripts can contain many false contigs which bias and impede the correct spectra analysis (Schliesky et al., 2012).
Even if reference protein databases are available, sample-specific proteogenomic analysis is difficult. Mutations or novel proteins present in a data set might not

be covered by the database and thus search methods can fail to detect these proteins. Methods that make use of existing SNP databases, such as dbSNP (Sherry et al., 2001), can partly approach this problem. However, also here only variations known *a priori* can be integrated. Thus, additional sample-specific mutations are not available. Moreover, integrating SNPs does not overcome the problem of detecting completely novel proteins.
A more sophisticated way to derive better suited search databases is the application of gene prediction methods. These methods explicitly aim at the identification of gene structures on unannotated genomes. Thus, they can predict novel genes and respect prokaryotic as well as eukaryotic gene characteristics. However, despite numerous research efforts and the availability of advanced methods, gene identification still faces significant challenges handling complex gene structures, rare splice sites or mutations in genes (Goodswen et al., 2012; Ederveen et al., 2013). For instance, a general problem of *ab initio* prediction methods is their dependency on given training sets. The influence of training set choice can be considerable: Parameters trained on one data set might not be feasible for other data, and thus predictions might not be correct in case of insufficient training data. Further, *ab initio* methods have the disadvantage (i) of providing no information on whether the genes are indeed expressed under a certain condition or not, and (ii) of missing or incorrectly predicting genes that differ from the considered standard codon scheme (Yada et al., 2002; Mathé et al., 2002). However, reduced sequencing costs and new fields like metagenomics, where even organisms are sequenced that cannot be cultivated, lead to more and more organisms that employ gene structures and codon schemes different from the ones we presently know (Woyke and Rubin, 2014; Ivanova et al., 2014). Although evidence-based gene finders can include experiment-specific information to approach these challenges, they are prone to noise in the experiments and can be limited by incomplete or contradicting evidence (Mathé et al., 2002). These limitations also apply to hybrid gene finders because they also rely on evidence. Further, since hybrid gene finders, such as AUGUSTUS (Stanke et al., 2006), are *ab initio* in their core prediction strategy, they are additionally limited in case of insufficient training data.
However, the search and construction of suitable sequence databases is not the only key problem in proteogenomic studies. Also the accurate search of spectra in the given database and the evaluation of search results is an important issue. For instance, the suitability and interpretability of FDRs and target-decoy analyses to estimate the proportion of incorrect identifications is an ongoing debate (Cooper, 2012; Jeong et al., 2012; Bonzon-Kulichenko et al., 2014; Branca et al., 2014). Here, a general problem is the choice of target databases: They must be large enough to include all proteins present in a sample, otherwise the measured peptides cannot be identified. However, large database sizes can bias the peptide identification and FDR evaluation (Reiter et al., 2009; Blakeley et al., 2012). This is particularly

challenging for proteogenomic analyses, where the database is often not a standard protein reference, but rather based on genomic analysis including six-frame translation. These databases are often large, which can be difficult for PSM evaluation (Branca et al., 2014).
Further, not only peptide identification, but also the inference of the original proteins based on the identified peptides is a challenging key problem in proteomics (Huang et al., 2012). Particularly the allocation of shared peptides to their correct protein remains an unsolved problem, such that they are often discarded from the analysis. Approaches including shared peptides parsimoniously select a subset of proteins explaining all observed peptides or group proteins that share peptides (Nesvizhskii et al., 2003; Serang et al., 2010). However, this results in a level of uncertainty in identifications. Further, since not all peptides can be measured equally well in the mass spectrometer (Sanders et al., 2007), often not all parts of a protein are fully covered, which additionally complicates a unique identification. Thus, a method to distinctively select the proteins actually present in the sample is highly desirable.

1.5. Terminology

Throughout the thesis we often refer to the term *gene* and its corresponding parts and structures. Hence, here we introduce the basic terms used to describe genes encoded in prokaryotic and eukaryotic genomes (refer to Figure 1.2 for an illustrative example).
We denote a region on a genomic sequence as a gene if it contains transcribed and translated intervals. When a gene is transcribed, the resulting messenger RNA (mRNA) sequence is called a *transcript*. Due to splicing events, in eukaryotes these transcripts are often present on the genome as several *exons* that are separated by *introns*. Introns are spliced out after the initial transcription and are thus not part of the transcript sequence that is translated to an amino acid sequence. The series of introns connecting the exons of a transcript is called *intron-chain*. Contradicting splicing events lead to *alternative transcripts*, where each transcript can also be denoted as an *isoform*. All isoforms corresponding to a gene represent the gene *locus*.
In contrast to eukaryotes, prokaryotic genes are organized in operons. Here, multiple so called *structural genes* are simultaneously regulated and also transcribed as one continuous mRNA. Thus, one transcript can contain multiple, possibly overlapping, structural genes encoding for different proteins.

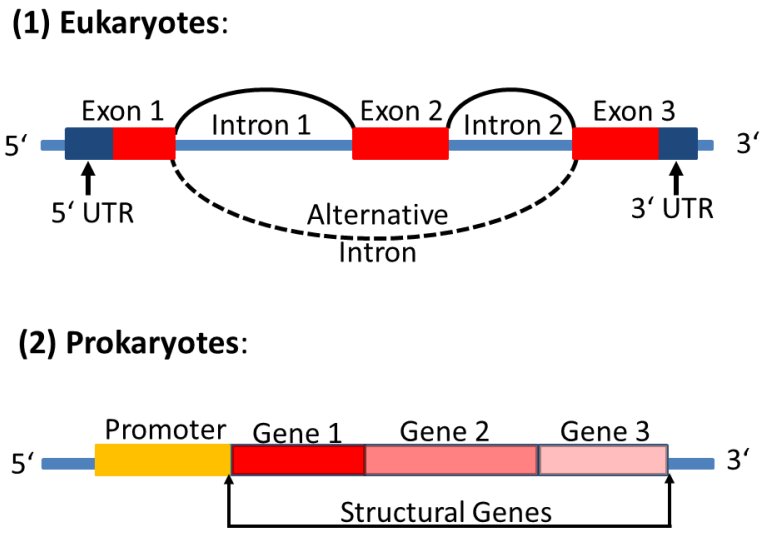

Figure 1.2.: Simplified examples of eukaryotic (1) and procaryotic (2) gene structures. Eukaryotes organize genes in exon intron structures and only exons are translated to amino acid sequences, after the introns are spliced out. It is possible that not only one transcript sequence is transcribed per gene, but multiple alternative transcripts (here indicated by an alternative intron illustrated with a dotted line). Note that not the complete exon sequence is translated: At the beginning and end of a gene we see an untranslated region (UTR). In contrast to eukaryotes, prokaryotes organize genes in operons, which can include multiple genes regulated by one promoter.

1.6. Thesis outline

This thesis introduces new computational methods to perform automated and accurate proteogenomic analysis and to overcome limitations described in the former sections. We integrate genomic and RNA-Seq data to construct reliable gene models, which are further refined in a post-processing that integrates evidence from additional gene predictions. Thereby, we approach accurate sample-specific gene identification and provide the basis for constructing customized databases for spectra searches. Further, we integrate tandem mass spectrometry and RNA-Seq information for tailored spectra search and improved shared peptide protein inference. This work is based on three publications and was undertaken under the supervision of Dr. Bernhard Renard, who is Co-author in each project.
Chapter 2 of the thesis describes the database construction with the evidence-based gene prediction method GIIRA, where Martin Linder participated in the development of the graphical model used in the prediction algorithm. The project is based on the following publication:

> **Zickmann F**, Lindner, MS, and Renard, BY : *GIIRA - RNA-Seq driven gene finding incorporating ambiguous reads*. Bioinformatics, 2014, 30(5), 606-613.

GIIRA addresses sample-specific and organism-independent gene identification. Unlike other gene finders, GIIRA includes the full information contained in RNA-Seq experiments by explicitly making use of ambiguously mapping reads. We describe the design and optimization of a read mapping network and evaluate the method in several experiments with prokaryotic and eukaryotic data.
In Chapter 3 we detail the method IPred that post-processes the output of GIIRA to combine gene prediction strategies and integrate additional evidence for more accurate combined predictions. IPred is based on work published in:

> **Zickmann F** and Renard, BY : *IPred - Integrating ab initio and evidence-based predictions for better gene identification*. BMC Genomics, 2015, 16(1), 134.

IPred targets the integration of prediction strategies to verify gene identifications. Since accurate gene models are strongly demanded in proteogenomics, IPred is designed as an easy-to-use intermediate step in genomic analysis pipelines. We describe the combination approach and show the superior prediction accuracy of IPred predictions in several experiments and comparisons to other combination methods.
In Chapter 4 the previously described approaches for accurate gene predictions are integrated in a proteogenomic analysis framework. We introduce the method MSProGene, which makes use of the sample-specific RNA-Seq-based gene model construction to construct customized databases for tandem mass spectra search. We address the problem of shared peptide protein inference by designing and optimizing a proteogenomic network. In a comprehensive comparison of different proteogenomic approaches in simulated as well as real data experiments we show that MSProGene facilitates an accurate proteogenomic analysis. This work is based on the following publication:

> **Zickmann F** and Renard, BY : *MSProGene - Integrative proteogenomics beyond six-frames and single nucleotide polymorphisms*. Bioinformatics, 2015, 31(12), i106-i115.

2. Constructing customized transcript databases - RNA-Seq driven gene prediction

A main focus in proteogenomic studies is the design and retrieval of tailored databases suitable for spectra search. Protein reference databases are not always available and are also not suited to identify novel or mutated proteins, whereas six-frame translations can introduce a bias in peptide identification (Reiter et al., 2009; Blakeley et al., 2012; Jeong et al., 2012; Branca et al., 2014) . Hence, sample-specific and reference-independent databases tailored to the experiment are required.
To approach this challenge, we designed the method GIIRA (**G**ene **I**dentification **I**ncorporating **R**NA-Seq and **A**mbiguous reads), a gene finder exclusively based on RNA-Seq information. The rationale to use a gene finder for database construction is (i) the independence of reference-protein databases and (ii) the integration of more sophisticated algorithms to predict likely expressed regions, in contrast to a simple ORF prediction or six-frame translation. Further, GIIRA is an evidence-based gene finder, which allows incorporating sample-specific information to gene prediction. In contrast to *ab initio* gene finders, such as GLIMMER3 (Delcher et al., 2007) or SNAP (Korf, 2004), this makes GIIRA ideal for predicting genes tailored to specific experiments and also tailored to detect mutated or novel genes.
RNA-Seq reflects the genes expressed in the current condition of the cell, which provides valuable information to identify novel genes or to confirm predicted genes. Although RNA-Seq experiments were included in various annotation studies (Martin et al., 2010; Palmieri et al., 2012; Tu et al., 2012; Pickrell et al., 2012; Sultan et al., 2008), so far only few gene finders directly incorporate RNA-Seq in gene prediction. Methods for gene expression analysis, such as iReckon (Mezlini et al., 2013), Cufflinks (Trapnell et al., 2010), Scripture (Guttman et al., 2010) and Erange (Mortazavi et al., 2008), perform a transcript assembly on RNA-Seq reads and thereby allow the identification of exons and splice sites, but they do not predict reading frames and start and stop codon for genes (Garber et al., 2011). The hybrid gene finder AUGUSTUS (Stanke et al., 2008) allows the integration of RNA-Seq experiments as an additional external source for eukaryotic gene identification, but the basis for the actual prediction is *ab initio* and relies on training data sets. The same holds for GeneMark (Besemer et al., 2001; Martin et al., 2010), a prokaryotic *ab ini-*

tio gene finder that can be combined with RNA-Seq analysis to identify operons. The gene finder G-Mo.R-Se (Denoeud et al., 2008) predicts gene models based on RNA-Seq reads, but does not identify mono-exonic genes and only incorporates non-ambiguous mappings.
Since for instance repetitive or highly similar regions or homologous genes lead to a substantial part of non-unique mappings, discarding ambiguously mapped reads from further analysis may result in a significant loss of prediction accuracy. This is exemplified in Table 2.1, where we show the proportion of ambiguous mappings for four different data sets. All RNA-Seq mappings (obtained using TopHat2 (Kim et al., 2013) with default settings) included a significant proportion of ambiguity. Thus, current RNA-Seq analysis methods that discard ambiguous reads (due to a lack of confidence in their mappings) remove a considerable proportion of information. In contrast, GIIRA explicitly targets the integration of ambiguity to assist gene identification with the complete information contained in RNA-Seq experiments: Based on the coverage observed in RNA-Seq read mappings, GIIRA first identifies candidate genes that are refined in further validating steps. These candidates are used to reassign ambiguous reads to their most likely origins using a maximum-flow approach formulated as a linear program. In addition, the identified candidates are completed into gene models by a search for start and stop codons as well as reading frame and strand prediction.
GIIRA is a gene finder that is primarily focused on prokaryotic gene prediction and in particular resolves genes within the continuously expressed region of an operon using a linear program optimization. However, GIIRA can also be applied to predict genes and alternative transcripts for eukaryotes and it leverages information from spliced reads for intron identification. Hence, it is also a useful addition to annotation pipelines, such as MAKER (Holt and Yandell, 2011), or a good complement to other eukaryotic gene finders.
In contrast to other approaches to ambiguous read assignment, such as the expectation maximization-based strategy introduced in (Chung et al., 2011) or ContextMap

	Human	*S. cerevisiae*	*E. coli*	*B. henselae*
NCBI accession	SRR032277	SRX187114	SRX180743	GSE44564
reads mapped (million)	29.0	5.8	10.1	51.7
ambiguous reads (%)	18.3	7.5	85.1	12.2
ambiguous hits (%)	53.1	19.0	97.4	35.1

Table 2.1.: Proportion of ambiguous reads for four different data sets, based on mappings with TopHat2 (Kim et al., 2013). Note that one ambiguous read has more than one hit on the reference genome. The values for the *E. coli* data set are based on the raw mapping without removal of rRNA contamination, which contributed to a particularly high number of ambiguous reads.

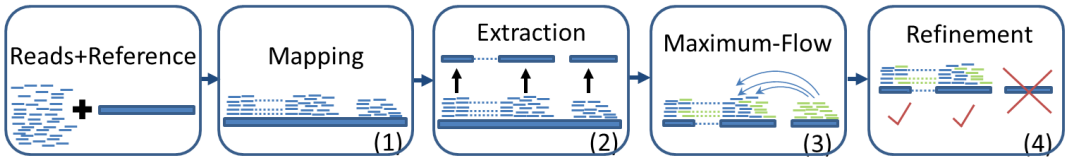

Figure 2.1.: Workflow of GIIRA: Given a genomic sequence and a set of RNA-Seq reads, reads are mapped to the reference (1) and the resulting alignment is then analyzed by GIIRA. Candidate genes are extracted (2) and ambiguous reads are reassigned using a maximum-flow optimization (3). Finally, candidate genes are evaluated based on the reallocated reads (4).

(Bonfert et al., 2012), our approach can integrate information on the likelihood of a read alignment not only from a fixed context (interval of specified length) or a context exclusively based on the mapping. Instead, we directly integrate the information gained in the process of identifying gene candidates and further the linear program ensures a convergence to an optimal solution.
We show the prediction accuracy of GIIRA and the advantage of integrating ambiguity in three simulations and on two real data sets. We compare our approach to the widely used transcript prediction method Cufflinks as well as the gene finders GeneMark (Besemer et al., 2001), GLIMMER3 (Delcher et al., 2007) and AUGUSTUS (Stanke et al., 2008).
Figure 2.1 illustrates the four main steps of the proposed algorithm. The input of GIIRA is a set of RNA-Seq reads which are aligned to a reference genome using an external alignment method (Fig. 2.1 (1)). Based on the alignment, GIIRA identifies regions on the genome that are likely to be expressed genes, in the following called *gene candidates* (Fig. 2.1 (2)). The identification regards the nucleotide coverage as well as splicing events indicated by the RNA-Seq reads. For prokaryotes, these candidates are regarded as expressed regions that might contain more than one gene. Hence, they are refined to determine the correct gene structure using an additional optimization step. Finally, ambiguously mapped reads are reallocated to their most likely origin using a maximum-flow optimization approach (Fig. 2.1 (3)). Based on this reassignment, the candidate genes undergo a refinement leading to the removal of candidate genes and isoforms without a sufficient number of remaining supporting reads (Fig. 2.1 (4)).

2.1. Candidate search

2.1.1. Alignment analysis

GIIRA is based on an alignment of reads from a RNA-Seq experiment to the DNA sequence of interest. For eukaryotes it is advisable, although not strictly necessary, to use a split read mapper for this alignment to obtain support for splicing events. The software is preconfigured to call either TopHat2 (Kim et al., 2013) or BWA (Li and Durbin, 2009) for read mapping, but can include the results of any read mapper with output in SAM format (Li et al., 2009a). GIIRA takes all mappings reported in the resulting SAM file into account, including ambiguous reads. For performance reasons, we only store the start positions of reads and their differences to the reference, as well as read quality and potential splice sites.
Note that during mapping analysis GIIRA additionally identifies mappings that are likely sequencing artifacts or (particularly for prokaryotic genomes) rRNA contamination. In order to do so, we calculate the average and median nucleotide coverage of the RNA-Seq mapping. A significant difference of these measures indicates high variation in the mapping coverage. However, although we expect different expression levels among transcripts, extreme coverage differences are unlikely because the overall coverage is dependent on the used sequencing protocol and sample preparation, which should equally affect all transcripts. Thus, regions with exceptionally high coverage indicate sequencing artifacts or contamination. These regions are excluded from the mapping with a simplistic iterative filtering: As long as average and median differ by more than one order of magnitude (i.e. the average is more than ten times as high as the median), the mappings that lead to the highest observed coverage are removed (threshold chosen based on the analysis of various independent experiments, data not shown).

2.1.2. Extraction

As illustrated in Figure 2.1 (2), regions with sufficient support of mapped reads are extracted to serve as *candidate genes*. The algorithm poceeds through all start positions of read alignments and tests if the coverage at these positions exceeds a *minimum coverage* threshold. Since the coverage threshold is an important parameter in the analysis, it can either be estimated from the given data without any *a priori* knowledge or be defined by the user. If the mapping coverage exceeds the minimum coverage threshold, a new candidate gene is opened and all following reads are assigned to the currently open region. This process is continued until the coverage falls below the *end coverage*, a threshold either user-specified or calculated from the minimum coverage. The currently open candidate gene is closed and the so called *core region* of a candidate is extracted (see Figure 2.2 for an example). The

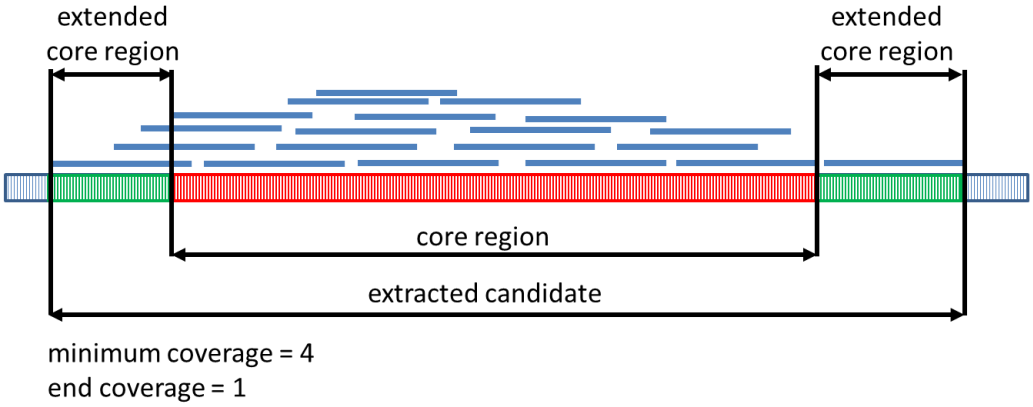

Figure 2.2.: An example for a gene candidate extraction with minimum coverage of 4 and end coverage of 1. The core region of the extracted candidate is marked in red. The complete extracted candidate after search of start and stop codon is marked in green.

core region defines the initial interval of a candidate gene, which is extended by the search for suitable start and stop codons.
We distinguish between minimum and end coverage because the mapping coverage observed in RNA-Seq experiments can vary significantly throughout the region of a gene and also between genes (Schurch et al., 2014; Garber et al., 2011). Usually, the overall coverage at beginning and end of a gene is smaller than in the middle and the ends of a gene are not accurately defined. This is also the reason why we search for start codon and stop codon in an interval exceeding the core candidate region. If the minimum coverage threshold is too small, the risk for false positive candidates increases because of possible incorrect read mappings. In contrast, if the end coverage is high (i.e. as high as the minimum coverage), we risk to loose parts of the gene due to low coverage ends and - in particular for genes with low overall coverage - to split one gene into several parts due to variations within the overall coverage. Thus, we distinguish minimum and end coverage to account for the coverage variations and their implications.
Once the core region of a candidate is identified, GIIRA aims at predicting the correct reading frame and strand of the corresponding gene. We expect reads that partly overlap with the core region to be also part of the gene and we expect to find start codon and stop codon not within the core region but in an interval before the beginning and after the end of the core, respectively (refer to Figure 2.2). We choose the length of this interval to be one read length to account for all overlapping reads. Within the specified interval, we search for start and stop codons regarding forward as a well as reverse strand (because at this point both directions are equally likely). Note that since not all organisms follow a standard codon usage, GIIRA can be pro-

vided with a list of alternative start and stop codons to be used for frame detection. Overall, we distinguish three different cases in frame prediction:

(i) Pairs of start codons and stop codons are found for only one direction, then we assume this to be the direction of the gene and choose the pair with the smallest possible interval including the core region. If a gene has no introns, a pair is suitable if it is in frame (separated by multiple triplets of nucleotides). In case of introns the pair can appear to be not in frame (but the pair is suitable after introns are spliced out).

(ii) We find pairs of start and stop codons for both directions. If we have information from the XS tag of the SAM file (that indicates the direction of split reads that span introns, see SAM format specification in (Li et al., 2009a)), then we prefer the direction supported by a higher number of reads. Otherwise we choose the smallest possible interval that explains the core region.

(iii) No pair of start and stop codon is detected. In this case we mark this candidate as incomplete, which means that GIIRA tries to merge it with neighbors (details explained below). As in case (ii), if we have XS tag information, we use it to identify the most likely direction of the candidate.

Note that in case (iii) an incomplete candidate is still reported as a candidate region as long as it is supported by read mappings after the reassignment of ambiguous reads and as long as it passes the subsequent filtering steps (similar to all other candidates, see Section 2.3). Hence, discrepancies between the RNA-Seq reads and the reference do not impede the identification of coding regions but are rather reflected in the resulting prediction, e.g., of partial or novel open reading frames.
Candidate genes without a suitable pair of start and stop codon are likely artifacts due to a low coverage because depending on the chosen coverage thresholds one gene with an overall small coverage is likely to be split into several candidate genes. Hence, we merge neighboring candidate genes in case their extended core regions overlap or if they are not farther apart than a user specified threshold (default: one read length).
If two candidate genes are merged, a new gene spanning both former candidate genes is defined, which inherits all properties as well as assigned reads and introns of the former candidates. Then we search a reading frame for the new candidate.

2.1.3. Splicing

In case of splicing events the basic extraction procedure is extended: Not only one continuous core region of a candidate gene is constructed, but depending on splice sites several regions connected by introns are extracted.

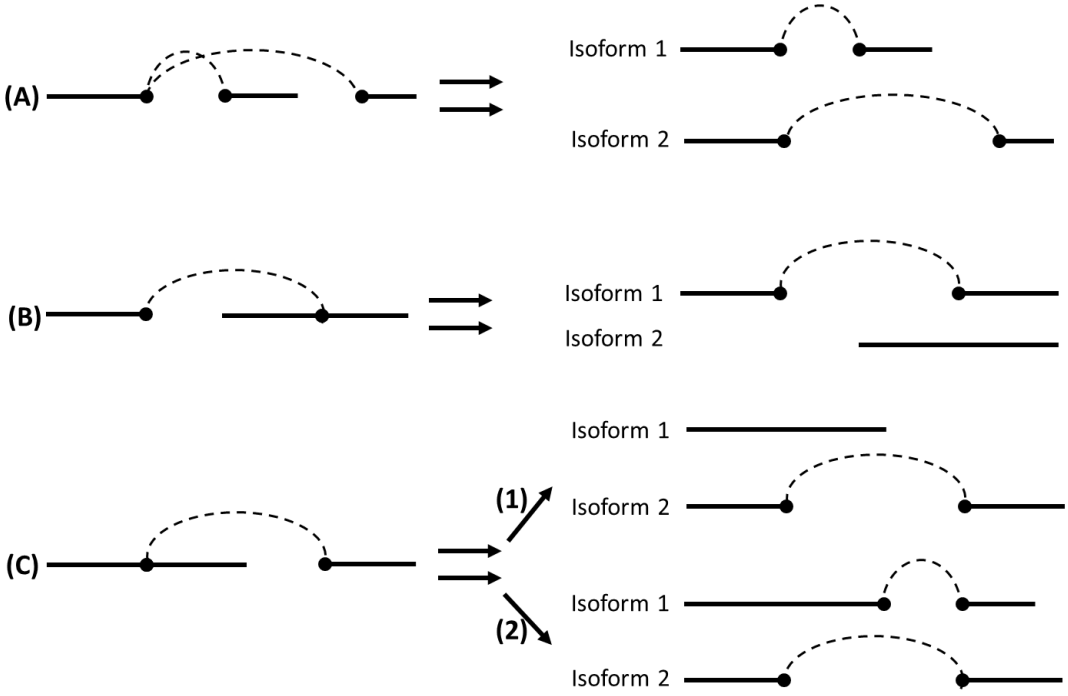

Figure 2.3.: Schematic representation of the three cases leading to alternative isoforms. In (A) more than one intron starts at the same position. In (B) a new exon starts within an ongoing intron and in (C) an intron starts within an ongoing exon.

A splice site is only considered as a non-erroneous site if it has a sufficient support of reads. By default the threshold for splice site acceptance is set equal to the overall desired minimum coverage. In case reads overlap an accepted splice site, they are assigned to their corresponding isoform, e.g., an intron starting at this splice position or an ongoing exon. During the initial extraction all isoforms with sufficient support by reads are taken into account, even if they indicate alternative or contradicting events. The refinement and exclusion of erroneous alternative isoforms is performed in subsequent steps (see Section 2.3).
An alternative splicing event can be indicated in several ways, which are illustrated in Figure 2.3: First, one splice start can lead to multiple endings, i.e. corresponds to multiple introns (Fig. 2.3 (A)). If more than one of these introns has sufficient read support, the splice start results in multiple alternative transcripts (according to the number of supported introns).

Second, an alternative isoform can start within the region spanned by an intron, indicated by reads mapping to positions within introns (Fig. 2.3 (B)). If the region exceeds the minimum coverage and if start or stop codon can be identified, we accept this alternative as a new transcript belonging to the same gene, only with a start codon downstream of the start codon of the gene.
Third, an intron starts within an ongoing exon, i.e. the position of the intron start is supported by a sufficient number of reads but other reads do not support any intron but indicate a normal exon (Fig. 2.3 (C)). This scenario can occur in two alternative ways: (i) The isoform without a splice site ends with the exon (and in contrast, the other isoform proceeds with the next exon) (see Fig. 2.3 (C.1)). (ii) The isoform without a splice site proceeds downstream to an alternative splice site (that is spanned by the first intron) (Fig. 2.3 (C.2)).

2.1.4. Prokaryotic gene structuring

Prokaryotic candidates undergo an additional extraction step since prokaryotic operons contain a continuously expressed region that can include several so called structural genes. For a given operon, we need to identify these genes respecting the present open reading frames (ORFs). To determine the most likely gene structure, we iteratively select sets of ORFs based on a linear program optimization.
First, all forward and reverse ORFs of the candidate sequence are enumerated. Second, the direction is selected that provides a set of ORFs that covers a large number of bases in this operon while restricting the overall number of ORFs. To achieve a trade-off between these two goals we adopt and alter a scoring metric from alignment evaluations (Vingron and Waterman, 1994): The set of all possible ORFs in a candidate sequence with length L is denoted as O. An ORF $o_i \in O$ contributes with its length l_i to the number of covered bases; hence, it is assigned a positive ("match") score $m_i = l_i$. If two ORFs o_i and o_j overlap, the overlap region is assigned a negative score ov_{ij} that equals the negative of the length l_{ov} of the overlap. This ensures that no region is counted twice. To avoid the suboptimal solution of simply selecting all ORFs present in O, we enforce sparsity by introducing an *ORF open penalty* p_i for each ORF o_i:

$$p_i = -\left(\frac{L}{l_i} \cdot \frac{l_{max}}{l_i}\right),$$

with l_{max} denoting the length of the longest ORF included in O. This penalty is smaller for longer ORFs since these are preferable to short ones because they cover more bases. Further, p_i reflects whether o_i is comparably short or long in relation to the ORFs present in O.

These metrics can be combined in a linear program that maximizes the sum of all scores:

$$max \sum_{i \in O} (m_i + p_i) + \sum_{i \neq j} ov_{ij}.$$

To integrate the above dependencies as constraints into the linear program, we introduce a variable $y_i \in \{0, 1\}$ for each ORF o_i. This variable indicates whether the corresponding ORF is chosen in the final solution ($y_i = 1$) or not ($y_i = 0$). This way we can write the different scores as follows:

$$\begin{aligned} m_i &= l_i \cdot y_i \\ p_i &= -\left(\tfrac{L}{l_i} \cdot \tfrac{l_{max}}{l_i} \cdot y_i\right) \\ ov_{ij} &= -(l_{ov} \cdot y_i \cdot y_j). \end{aligned}$$

The overlap constraints are quadratic to ensure that the overlap penalty is only applied if indeed both overlapping ORFs are selected. Note that if three or more ORFs overlap we also regard all pairwise overlaps of these ORFs. Hence, for more than two overlapping partners we subtract more than the originally counted region, thereby additionally penalizing highly overlapping ORF combinations.
We formulate the linear program for the set of all forward and reverse ORFs, respectively; and use an optimization method such as CPLEX (CPLEX, 2011) to obtain the best selection of ORFs for each direction. For each selection, the solution maxima reported by the optimization method are compared and the direction with the higher score is selected as the direction of the entire candidate sequence.
However, in cases of nearby or overlapping operons the extracted candidate sequence might span more than one operon in different directions. Since microorganisms often have densely packed genomes with many overlapping genes and operons, we integrate the former described procedure in an iterative process to ensure that potential overlapping coding sequences are detected and the corresponding genes are identified. This iteration process is illustrated in Figure 2.4.
The first iteration results in a set of ORFs that best explains the given candidate sequence. In a second iteration, we formulate a linear program similar to the initial one, with the difference that now we fix the previously selected ORFs and in addition pass the complete set of ORFs from the other direction. The idea is that now an ORF from the initially not chosen direction can be additionally selected if it enhances the overall alignment score (i.e. if it explains a part of the coding sequence that has not been explained by the previously chosen ORFs).
If this iteration step selects ORFs from the opposite strand, a final iteration ensures that if previously selected ORFs are less likely than newly selected ones (e.g., because they are contained in newly selected ORFs), they are discarded from the final set of chosen ORFs. Thus, in the third iteration we fix the newly selected ORFs and

formulate the linear program including all so far selected ORFs.
After this iteration procedure, the former gene candidate is split into several candidates in case two or more operons are identified.

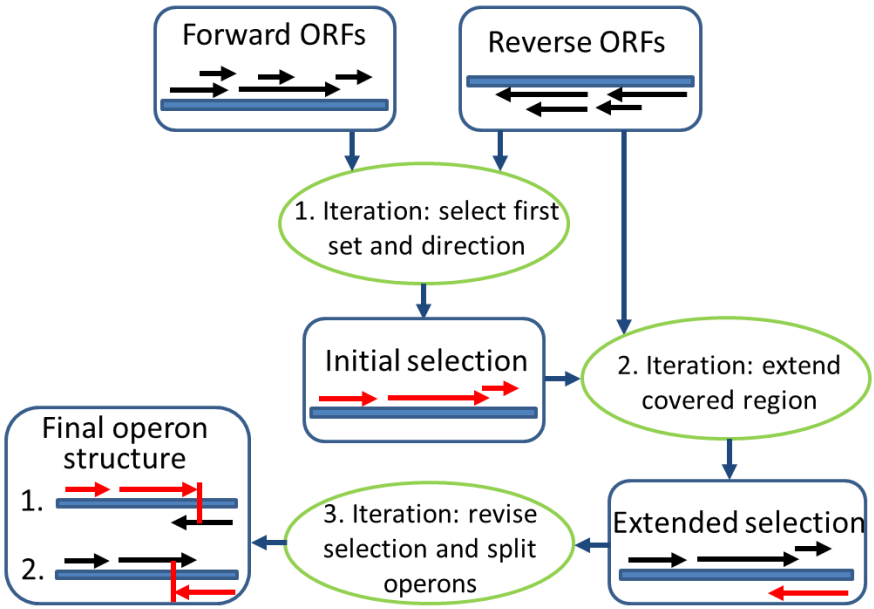

Figure 2.4.: Illustration of the iteration process of the prokaryotic operon structure optimization. Initially, forward and reverse open reading frames (ORFs) are processed separately and the direction with the highest optimization score is selected as the initial set of ORFs. In a second optimization, this set is extended with ORFs from the other direction. In a third iteration the overall selection is revised by excluding ORFs that became unsuitable because of the selection of the second iteration. Finally, the regions and structure of the resulting operons are defined.

2.2. Maximum-Flow optimization

Up to this point, all read mappings contributed equally to the extraction of candidate regions, even if a read had multiple mappings with similar quality. However, as each read can only arise from one genomic locus, we aim at reassigning ambiguously mapped reads to their most likely origin. To do so, we perform a maximum-flow (Ford and Fulkerson, 1956) optimization using the information of extracted gene candidates.

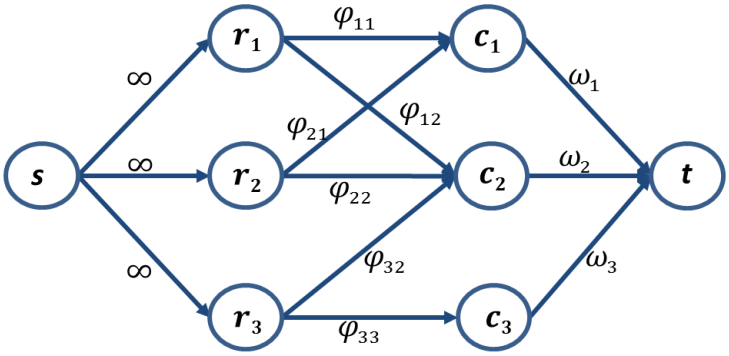

Figure 2.5.: Simplified example for a maximum-flow network representation propagating information from source node s to target node t. The source node is connected to the nodes representing reads (r_i), which are connected with all genes they are mapped to (c_j). The edge labels indicate the capacity for the throughput that is allowed to be passed from one node to the other (representing the support of the read to the corresponding candidate gene).

The rationale behind this approach is that if several genes compete for the same read, their overall read coverage and the presence of support from unique reads indicates the most likely origin of this read. Both factors do not only enhance the probability for a candidate to be chosen, but also decrease the chances of the competitors such that the number and quality of the competitors directly affects the choice for the best origin. Further, also the ambiguity of the read itself is taken into account by weighting the influence of reads on candidate quality by the number of their alignment positions. The more alignments a read has, the less it supports each single gene it is mapped to.
The problem of assigning each read to exactly one gene candidate is formulated as a graphical model, as illustrated in Figure 2.5. We define a network $G = \{N, E\}$ with edge set E and node set $N = R \cup C \cup s \cup t$ with nodes $r \in R$ representing reads and nodes $c \in C$ representing gene candidates, respectively. Source node s and target node t are defined for technical reasons. Further, all edges are directed and an edge $e_{ij} \in E$ between two nodes represents that read $r_i \in R$ is assigned to gene $c_j \in C$. Note that each edge has a capacity, which can be understood as the maximal input that can pass through this edge. In contrast, nodes have an unlimited throughput.
The aim of the maximum-flow is to set all capacities φ_{ij} (belonging to edges e_{ij} connecting a read r_i to a candidate c_j) in a way that the flow passing from source

to target node is maximized:

$$max \sum_{e_{ij} \in E} \varphi_{ij}.$$

Each edge originating from the source has an unlimited capacity. The capacity φ_{ij} of the edges connecting reads and their possible corresponding genes is restricted by the following condition:

$$0 \leq \varphi_{ij} \leq y_{ij} \;\; \forall e_{ij} \in E,$$

where $y_{ij} \in \{0, 1\}$ are the binary variables that denote whether the read r_i is assigned to gene c_j ($y_{ij} = 1$) or not ($y_{ij} = 0$). In other words, if a read is assigned to a gene, the corresponding edge connecting both nodes has a capacity with a maximal value of 1. If the read is not assigned, the capacity is zero.
In addition, we require all ambiguously mapping reads to be assigned to exactly one candidate, as reflected in the constraint:

$$\sum_{j} y_{ij} = 1 \;\; \forall \, i | r_i \in R.$$

Further, each gene has a maximal number of reads that can be assigned, depending on the support of reads for this gene and the support for its competitors. Since for each node the input flow has to equal the output flow, this maximum is given by the capacity ω_j of the edges connecting gene nodes to the target node:

$$\sum_{i | e_{ij} \in E} \varphi_{ij} \leq \omega_j \;\; \forall j | c_j \in C,$$

where ω_j is calculated as follows:

$$\omega_j = \frac{b_j}{\sum\limits_{c_k \in P_j} b_k^u}.$$

Here, b_j is the average base coverage of gene c_j derived by all its mapping reads, where in contrast b_j^u is the coverage derived only by reads that map uniquely to the corresponding gene. The set P_j contains all genes that directly compete with c_j for ambiguously mapped reads, or in other words, that share reads with gene c_j. For illustration, refer to Figure 2.5: here P_2 consists of c_1 and c_3, whereas P_1 only includes c_2 because c_1 only shares reads with c_2.
Allowing genes to influence their competitors with the help of their own likeliness ensures that not only genes with an overall high coverage are preferred over genes with less coverage. Otherwise genes with no or only few unique reads could be

preferred over genes with a high unique coverage, as long as they have enough multiple hits.
The maximum-flow problem is formulated as an integer linear program including the constraints described above. This program is solved using for instance the IBM CPLEX academic version V12.4 (CPLEX, 2011) or, as a slower alternative, the open source GLPK solver (GLPK, 2006).

2.3. Candidate refinement and scoring

The maximum-flow optimization identifies a unique position for each read such that the previously extracted gene candidates have to be refined according to the new assignment of reads. If a gene candidate or an alternative isoform lost all of its supporting reads, it is regarded as an artifact of ambiguous read mappings and is thus erased. All remaining genes are evaluated in a scoring process according to their exon length l_j, their read coverage and the quality of their assigned reads. It is also of relevance whether the corresponding reads are mapped ambiguously since ambiguity implies more uncertainty for the gene and thus leads to a smaller score. The final gene score s_j for gene c_j is calculated as:

$$s_j = \frac{1}{l_j} \cdot \sum_{i|e_{ij} \in E} \frac{l_i \cdot q_i}{M_i},$$

where q_i denotes the quality of read r_i, l_i its length and M_i its total number of mappings. GIIRA reports the identified genes and transcripts in GTF annotation format, including additional information on coverage and ambiguous read support. This allows an easy post-processing to verify genes for follow-up analyses. GIIRA also provides a filter script that can be used to perform the post-processing.

2.4. Implementation

GIIRA is implemented as a Java program (http://www.java.com). Further, it uses helper scripts written in Python (http://www.python.org/), including the python packages SciPy, NumPy, and PySam. For optimization, GIIRA relies on the CPLEX Optimizer (CPLEX, 2011) (free for academic use) or the open source alternative GLPK (GLPK, 2006). Note that GLPK cannot be applied to quadratically constrained linear programs; hence, only CPLEX can be used as the solver for prokaryotic gene structuring. However, to perform a gene prediction on prokaryotes without installed CPLEX, the gene structuring can be turned off by not setting the parameter "-prokaryote". Then the prokaryotic genome is treated as a eukaryotic genome and the coding sequences are interpreted as exons instead of resolved

into structural genes (hence, no quadratically constrained linear program is formulated). GIIRA is open source and the source code or a precompiled version can be downloaded from http://sourceforge.net/projects/giira/.

2.5. Experiments

GIIRA was evaluated on three simulated and two real data sets, on prokaryotic as well as eukaryotic organisms. The details of the different experiments and comparisons to other methods are presented in the following. Note that for all data sets the RNA-Seq reads used as evidence for gene prediction were mapped to the respective reference sequences, using TopHat2 (Kim et al., 2013) (version 2.0.8) with default settings. The resulting alignment served as the starting point for all compared methods. We also analyzed the mapping with regard to the proportion of ambiguous reads and the number of resulting ambiguous hits to investigate the influence of ambiguity in our data sets.

2.5.1. Simulations

To evaluate GIIRA given a known ground truth, we use three different simulated data sets. We generated a prokaryotic simulated data set based on *Escherichia coli* (NCBI-Accession: NC_000913.3) and two eukaryotic simulations based on chromosome 15 of the human genome (NC_000015.9) and chromosome 4 of *Saccharomyces cerevisiae* (NC_001136.10), respectively (the simulation setup is explained below). Based on this data, we compare GIIRA to Cufflinks (Trapnell et al., 2010), GLIMMER3 (Delcher et al., 2007), and GeneMark (Besemer et al., 2001) in the prokaryotic simulation and to Cufflinks and AUGUSTUS (Stanke et al., 2008) in the eukaryotic simulations.
As GeneMark is originally an *ab initio* gene prediction method that does not include RNA-Seq information, we used the framework proposed by Martin et al. (2010) that combines GeneMarkS (Besemer et al., 2001) *ab initio* predictions with the program ParseRnaSeq to include RNA-Seq evidence. Then, we generated a pile-up count file based on the read mappings to indicate the number of reads covering each position of the genome. This information is combined with standard GeneMarkS (version 4.6b) gene predictions (GeneMarkS was applied with default settings). Note that in this framework the resulting predictions cover operons rather than structural genes. GLIMMER3 (version 3.02) and Cufflinks (version 2.0.2) were applied with default settings. GLIMMER3 predicted genes directly on the unannotated reference sequences, whereas Cufflinks was applied on the mapping file obtained with TopHat2. AUGUSTUS (version 2.7) can incorporate information from RNA-Seq experiments in form of "external hints". We followed the pipeline recommended

on the AUGUSTUS website[1] for RNA-Seq integration and filtered the RNA-Seq mapping to only contain uniquely mapped reads. GIIRA was applied with default settings on the RNA-Seq mapping in SAM format, sorted by read names. CPLEX was used to solve the maximum-flow optimization, and for prokaryotic data sets we specified the parameter "-prokaryote". In addition, to demonstrate the influence of ambiguous mappings on the prediction accuracy we configured and compared a second version of GIIRA that excludes ambiguous mappings from the analysis. To ensure a fair comparison between methods, we masked all direction information in our evaluation since Cufflinks does not report any strand information in case no splicing events occur.

Simulation setup

The simulation uses the read simulator Mason (Holtgrewe, 2010) applied to the NCBI reference annotation for each organism of interest. In this annotation the coding sequence of each known isoform appears as a consecutive sequence. Hence, the simulated reads show similar characteristics as real RNA-Seq reads since they cover alternative isoforms, span introns (if existing in the data set), and show a coverage profile typical for gene expression. The simulation setup is illustrated in Figure 2.6 and information on the simulated data sets is summarized in Table 2.2. We illustrate the process on the example of the human data: As an overall sequencing depth we intended to achieve a coverage of 20. To obtain different gene expression levels, before applying Mason the set of annotated coding sequences for human chromosome 15 was divided into three parts with almost equal overall exon lengths. For each part, reads were simulated with different coverages of 10, 20 and 30, respectively. After the simulation we merged the reads from all parts and obtained expression levels ranging from below 10 to over 30 of coverage depth.

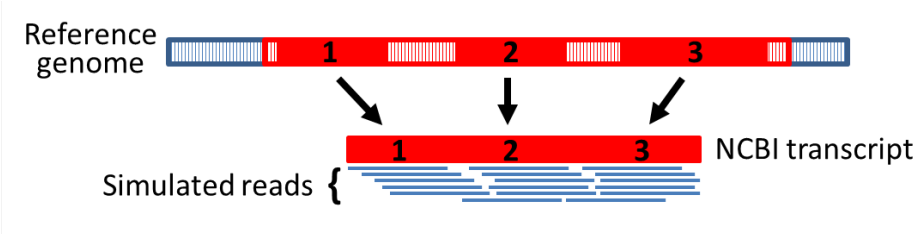

Figure 2.6.: Idea of the simulation study: The NCBI annotation contains the coding sequences of each transcript as consecutive sequences without introns. Hence, we simulate reads directly from these sequences and obtain reads spanning introns and reflecting different isoforms.

[1] http://bioinf.uni-greifswald.de/bioinf/wiki/pmwiki.php?n=Incorporating RNAseq.Tophat

	E. coli	*S. cerevisiae*	Human chromosome 15
number reads	1,500,000	570,000	500,000
read length	36	50	50
simulated cov.	15	25	20
experimental average cov.	14.0	28.5	18.7

Table 2.2.: Statistics on the three simulated data sets. For each simulation the simulated coverage (simulated cov.) and the average coverage observed in the actual mapping (experimental average cov.) are presented.

2.5.2. Real data sets

We applied GIIRA, Cufflinks, GLIMMER3, and GeneMark on a real data set of 11 million reads (NCBI accession: SRX180743) from *E. coli*. This data set contains a large proportion of ambiguous mappings as well as high coverages in the areas coding for ribosomal RNA, posing a challenge to distinguish false from correct gene loci. Since GIIRA is also applicable to eukaryotic organisms, an experiment with a real *S. cerevisiae* data set comprising 6 million reads (SRX187114) was performed comparing GIIRA and Cufflinks. The settings and versions of all compared methods are equal to the specification used in the simulation experiments.
In contrast to the simulations, where we can compare predictions to a specific ground truth, for the real data experiments a known ground truth is not available. Thus, we evaluated the compared methods against the complete annotation of *E. coli* and *S. cerevisiae*, respectively. However, this evaluation can only be regarded as a relative comparison between methods.
Since not all genes of an organism are necessarily expressed at the same time, we performed an additional evaluation for the real *E. coli* data set, based on the observed RNA-Seq evidence for this experiment. The evaluation is based on the comparison against a reference subset including likely expressed genes. To obtain the subset, we analyzed the TopHat2 mapping of the RNA-Seq reads to the reference genome. We counted all reads mapping to each annotated region and then sampled a subset of reference genes comprising all annotations with a minimum overall mapping coverage greater than one. This resulted in a sample of 2,002 reference genes instead of the original 4,146 annotations.

2.5.3. Evaluation

To evaluate the compared methods following accepted standards, gene predictions were analyzed using the Cuffcompare framework (Trapnell et al., 2012), providing the annotated coding sequences of NCBI as a reference transcript set. Cuffcompare follows the guidelines presented in (Burset and Guigó, 1996). Here the gene

predictions are evaluated on several levels, namely the base, exon, intron, intron-chain, transcript, and locus level. The base level reflects the per-base accuracy by distinguishing the following four categories for each base prediction. Each prediction can be a correct prediction as part of a coding sequence ("True positive", or *TP*) or as non-coding ("True negative", or *TN*), or vice versa a false prediction as coding ("False positive", or *FP*) or non-coding ("False negative", or *FN*). Based on these definitions, prediction sensitivity (Sn) and specificity (Sp) can be obtained by calculating the proportion of true predictions on the set of all possible coding bases and the set of all predicted bases, respectively:

$$Sn = \frac{TP}{TP + FN}$$
$$Sp = \frac{TP}{TP + FP}.$$

Similarly, the other levels are separated into the four different categories (TP, TN, FP, FN) and the corresponding Sn and Sp can be calculated. In contrast to the base level, on exon level an exact overlap of predicted and ground truth exon is required to be counted as a true positive. Further, for a correct intron-chain (which is the series of introns explaining a transcript) all introns belonging to this chain have to be reported (where an intron is defined by the interval between two connected exons). The transcript level directly corresponds to the intron-chain level, which means that a transcript is considered to be identified if the corresponding intron-chain is correct and if no additional exon has been assigned to this transcript. Finally, a locus is considered as correctly identified if at least one of its transcripts is found.
The conditions for an exact match are very strict, because in particular gene annotations depend on the exact prediction of start and stop codons to preserve the correct reading frame. However, to also account for exons and introns (and the corresponding other levels) that only slightly differ from the exact match, Cuffcompare introduces the "fuzzy" match. This fuzzy measure counts exons as a match even if they show a very small variation to the correct exon boundaries.
This way, the fuzzy measure indicates whether correct identifications were found in proximity even though the precise location might have been missed. This is particularly important for RNA-Seq-based gene predictions since RNA-Seq mappings show lower coverage at beginning and end of genes, which is challenging for correct start and stop codon prediction.
As an overall measure of prediction accuracy, for each compared level we combined Sn and Sp in the well-known F-measure F (van Rijsbergen, 1979):

$$F = 2 \cdot \frac{Sp \cdot Sn}{Sp + Sn}.$$

In addition, we generated receiver operating characteristic (ROC) curves that illustrate the base level accuracy of the compared methods. The ROC curves are generated by sorting all predicted exons according to their assigned score. Then for each exon we calculate the sum of correctly predicted exonic bases and the sum of all incorrectly predicted bases (false positives and false negatives). To not only include the nucleotide level in the ROC analysis but also the exon level accuracy, we only take nucleotides of exons with at least partial overlap to reference exons into account (i.e. they cover a reference exon, or one reference exon covers a predicted exon, or they share an interval larger than one read length). If a predicted exon does not fall into this category, it counts as a false positive.
In addition, for the two real data sets and the human simulation we performed an alternative evaluation study based on sampling a fixed number of predictions for all compared methods. This way the measure of accuracy is independent of the overall number of predictions of each tool. To sample a fixed number of gene predictions, first all predictions were sorted according to their score. Note that GIIRA, GLIMMER3 and GeneMark yield such a score for each predicted gene, whereas for Cufflinks we used the provided coverage score associated to each transcript as the quality measure closest to the GIIRA score. For AUGUSTUS we utilized the score associated to *"% of transcript supported by hints (any source)"* (from the AUGUSTUS output file) as the reported measure closest to prediction reliability.

2.6. Results

2.6.1. Mapping and ambiguity

In our study we intend to demonstrate the applicability of GIIRA on different organism types and the effect of including ambiguous mappings in the analysis. Thus, a crucial point is the proportion of ambiguously mapped reads in the alignment. The details of the mappings resulting from TopHat2 are listed in Table 2.3. For all data sets except the real *E. coli* experiment we see that the main proportion of ambiguous mappings has its source in a comparably small number of actual ambiguous reads. The reason for this observation is that most of the ambiguous reads do not only map two times, but rather several times to the reference sequence.
All mappings showed ambiguity, although in varying levels: with 6.6% the *E. coli* simulation has the lowest proportion of ambiguous mappings, while the real *E. coli* experiment shows the highest proportion with 97%. The human simulation shows 22.8% and the yeast data sets 19% ambiguous hits, respectively. Note that high ambiguity in the real *E. coli* data is due to a high level of rRNA contamination within the sample, as is often observed in prokaryotic RNA-Seq experiments (Sorek and Cossart, 2010). Without contamination, the ambiguity is approximately 5%, simi-

lar to the mapping proportion of the simulated *E. coli* data set that did not include contaminants.

	E. coli Sim	*S. cer* Sim	Human Sim	*E. coli* Real	*S. cer* Real
reads mapped	1,448,779	551,596	472,969	10,052,045	5,754,018
ambiguous reads(#)	20,395	30,967	31,769	8,555,561	430,389
ambiguous reads(%)	1.7	5.6	6.7	85.1	7.5
hits total	1,529,558	638,869	571,815	57,769,265	6,569,842
ambiguous hits(#)	101,174	118,240	130,615	56,272,781	1,246,213
ambiguous hits(%)	6.6	18.5	22.8	97.4	19.0

Table 2.3.: Mapping results and the proportion (in percent) of ambiguous reads and ambiguous hits for the TopHat2 mapping of the three simulated and two real data sets with *E. coli* and *S. cerevisiae* (*S. cer*), respectively.

2.6.2. Simulations - *E. coli* data set

Table 2.4 shows the Cuffcompare comparison between Cufflinks, GIIRA, GeneMark, and GLIMMER3 for the *E. coli* simulation. The reads were simulated directly from the complete set of annotated genes. Thus, in this simulation no operon resolution was necessary, but rather the identification of expressed regions and the resolution of overlaps between genes. Overall, GIIRA shows the best prediction accuracy for all evaluated categories. For instance, the accuracy on the exact measure on the locus level is increased by 9% compared to GLIMMER3, the second best method. Only on the exact base level, the sensitivity of GLIMMER3 (96.7) is slightly higher than the sensitivity of GIIRA (96.5). But due to the better specificity of GIIRA, also on this level its overall accuracy is still higher compared to all other methods. This is also illustrated in Figure 2.7: Cufflinks, GLIMMER3, and GIIRA show a high accuracy on the base level, with GIIRA being more specific than Cufflinks and GLIMMER3, whereas GLIMMER3 is slightly more sensitive than GIIRA. Compared to GeneMark, all methods show a sensitivity and specificity increased by more than 20%.
Since only GIIRA and GLIMMER3 focus on extracting structural genes rather than operons or expressed areas, it is as expected that on exon and locus level both methods show significantly better accuracy than the competing methods. We note that for Cufflinks only the fuzzy exon and locus level are of relevance since Cufflinks does not predict start and stop codons and thus regularly misses bases at the start and end of genes. The fuzzy category covers these bases because here not only a perfect match, but also a match in a range around the correct result is accepted.

***E. coli* simulation**

	exact measure								
	Base			Exon			Locus		
	Sn	Sp	F	Sn	Sp	F	Sn	Sp	F
GIIRA	96.5	**97.7**	**97.1**	**76.5**	**69.9**	**73.1**	**78.3**	**81.1**	**79.7**
Cufflinks	91.1	92.5	91.8	0.1	0.2	0.1	0.2	0.2	0.2
GeneMark	69.2	66.5	67.8	0.0	0.0	0.0	0.0	0.0	0.0
GLIMMER3	**96.7**	94.6	95.6	71.9	67.8	69.8	72.3	69.4	70.8
	fuzzy measure								
	Base			Exon			Locus		
	Sn	Sp	F	Sn	Sp	F	Sn	Sp	F
GIIRA	-	-	-	**77.6**	**70.9**	**74.1**	**79.3**	**82.1**	**80.7**
Cufflinks	-	-	-	27.6	47.1	34.8	32.3	47.1	38.3
GeneMark	-	-	-	5.8	21.8	9.2	6.8	21.8	10.4
GLIMMER3	-	-	-	73.1	68.9	70.9	73.5	70.5	72.0

Table 2.4.: Cuffcompare analysis for the simulated *E. coli* data set. The highlighted numbers indicate the best results for each criterion for measures of sensitivity (Sn), specificity (Sp), and F-measure (F), respectively for GIIRA, Cufflinks, GeneMark, and GLIMMER3.

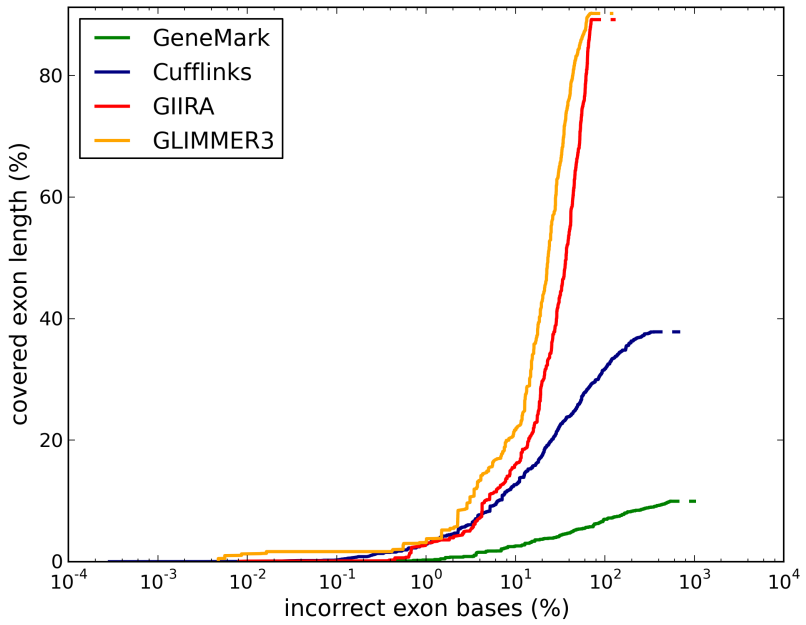

Figure 2.7.: ROC curve comparing the proportion of correctly and incorrectly predicted exonic bases for GeneMark, Cufflinks, GIIRA, and GLIMMER3 for the *E. coli* simulation. Dashes indicate the number of bases missed due to not identifying a reference exon. Note that the proportion of false predictions is reported on a logarithmic scale.

For GIIRA the fuzzy sensitivity and specificity are only slightly increased compared to the perfect level measures, indicating a high accuracy in predicting the correct frame for an expressed region. Also in the fuzzy categories GIIRA shows significantly improved results compared to the other prediction methods, in particular to Cufflinks and GeneMark.

2.6.3. Simulations - Human data set

In the human simulation, we compared GIIRA to Cufflinks as well as AUGUSTUS as an example of a hybrid gene prediction approach. Table 2.5 (1) shows the overall accuracy of predictions, the detailed sensitivity and specificity of all compared categories are presented in Table A.1 in the appendix.
As illustrated in the tables and in Figure 2.8, GIIRA yields the most sensitive predictions on the base level as well as on the fuzzy exon and transcript level, while Cufflinks is more sensitive in predicting introns, in particular exact intron-chains. Further, on the exact exon and intron level GIIRA yields a sensitivity comparable to the best values (obtained by AUGUSTUS) while it is clearly more specific with an

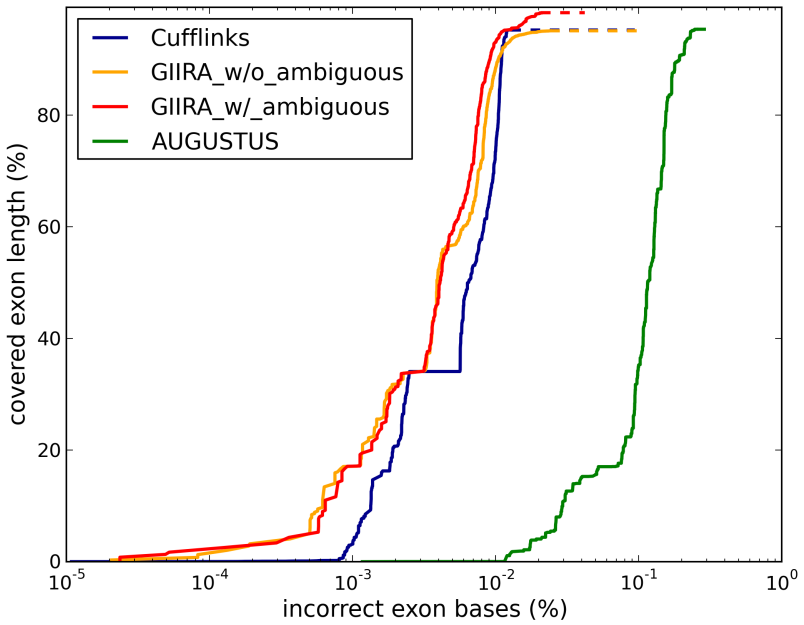

Figure 2.8.: Comparison for the human data set, showing correctly and incorrectly predicted exonic bases for Cufflinks and AUGUSTUS, and for GIIRA excluding ("GIIRA_w/o_ambiguous") and including ("GIIRA_w/_ambiguous") ambiguous reads, respectively. Dashes indicate the number of bases missed due to not identifying a reference exon. The proportion of false predictions is reported on a logarithmic scale.

increase of more than 7% and 11%, respectively. Thus, for the exact measure GIIRA shows the highest accuracy on base, exon, and intron level. In the fuzzy evaluation, it additionally achieves the highest transcript level accuracy.

(1) Human Simulation - complete set of predictions

Methods	Base	Exon	Intron	Intron-Chain	Transcript	Locus
F-measure - exact						
GIIRA_w/_ambiguous	**97.6**	**87.4**	**93.8**	43.9	36.6	50.4
GIIRA_w/o_ambiguous	95.9	84.0	91.6	43.9	35.9	46.9
Cufflinks	95.4	74.8	91.7	**50.2**	0.5	49.4
AUGUSTUS	87.5	84.8	88.5	47.2	**38.7**	**51.2**
F-measure - fuzzy						
GIIRA_w/_ambiguous	-	**91.6**	**94.5**	57.1	**42.6**	54.1
GIIRA_w/o_ambiguous	-	88.3	92.1	54.1	40.7	49.2
Cufflinks	-	88.9	92.2	65.0	35.7	53.0
AUGUSTUS	-	85.6	88.9	**72.9**	40.0	**63.1**

(2) Human Simulation - sampled set of predictions

Methods	Base	Exon	Intron	Intron-Chain	Transcript	Locus
F-measure - exact						
GIIRA_w/_ambiguous	**92.6**	83.2	88.3	45.5	38.8	54.4
GIIRA_w/o_ambiguous	91.1	81.7	86.9	45.9	39.2	53.1
Cufflinks	74.9	58.0	73.2	44.3	0.3	44.0
AUGUSTUS	90.4	**86.9**	**90.1**	**49.5**	**40.7**	**55.7**
F-measure - fuzzy						
GIIRA_w/_ambiguous	-	86.9	89.1	59.4	**45.4**	58.1
GIIRA_w/o_ambiguous	-	85.1	87.5	56.9	44.5	55.8
Cufflinks	-	71.7	73.4	56.7	31.0	47.4
AUGUSTUS	-	**87.7**	**90.5**	**76.5**	42.2	**69.4**

Table 2.5.: Excerpt of the Cuffcompare analysis for the simulated human data set showing the F-measures for GIIRA including ambiguous reads (GIIRA_w/_ambiguous), GIIRA excluding ambiguous reads (GIIRA_w/o_ambiguous), Cufflinks, and AUGUSTUS. Table (1) shows the evaluation on the complete set of predictions. In Table (2), a sample of 600 predictions for each compared method is evaluated against the 992 reference transcripts. The best result of each category is marked in bold.

AUGUSTUS predicted a high number of incorrect exons (data not shown), which is reflected in the low specificity of AUGUSTUS observed in Figure 2.8 and in the reduced exon and intron level specificity compared to other methods. On the locus level, AUGUSTUS shows higher accuracy than GIIRA and Cufflinks. However, in the exact evaluation the improvement is comparably small, while in the fuzzy evaluation GIIRA and Cufflinks are outperformed by approximately 10% in sensitivity and 7% in specificity.
The direct comparison between GIIRA including and excluding ambiguous reads shows that the prediction sensitivity is increased for all levels when ambiguous mappings are included (refer to Table A.1 in the appendix). The effect is particularly pronounced on the exon and intron level, where including ambiguous reads reduces the lack of sensitivity by up to one third. This leads to an overall improved prediction accuracy.
Table 2.5 (2) shows the evaluation for the simulated human data set based on predictions sampled to size 600. Details on sensitivity and specificity are presented in Table A.2 in the appendix. We see changes in terms of which method is marked best for a certain category (e.g., for the exon level, where GIIRA is best on the complete set of predictions and AUGUSTUS on the sampled set). Compared to the evaluation on the complete set of predictions, overall the sensitivity is decreased and the specificity is improved. Particularly AUGUSTUS shows an improved overall accuracy on the sampled set due to higher specificity values. Cufflinks shows decreased accuracy on all compared levels. GIIRA displays reduced accuracy on base, exon, and intron level, but increased accuracy in the other categories.

2.6.4. Simulations - *S. cerevisiae* data set

Table 2.6 shows the F-measure analysis and Figure 2.9 illustrates the corresponding ROC curves of the gene predictions for the simulated yeast data set. Details on sensitivity and specificity are presented in Table A.3 in the appendix.
Overall, on all levels except intron-chain GIIRA shows higher prediction accuracy than Cufflinks. On the intron-chain level it shows comparable results since it is less sensitive but more specific than Cufflinks.
In regard to the comparison between the two GIIRA configurations including and excluding ambiguous reads, we see that including ambiguous read mappings results in a higher sensitivity and specificity in gene predictions, although overall both configurations show comparable results.

Methods	Base	Exon	Intron	Intron-Chain	Transcript	Locus
F-measure - exact						
GIIRA_w/_ambiguous	**99.2**	**85.7**	**75.9**	73.7	**86.1**	**86.7**
GIIRA_w/o_ambiguous	98.3	84.9	**75.9**	73.7	85.2	85.9
Cufflinks	95.6	1.9	75.4	**74.6**	0.6	3.6
F-measure - fuzzy						
GIIRA_w/_ambiguous	-	**86.6**	**75.9**	73.7	**86.9**	**87.1**
GIIRA_w/o_ambiguous	-	85.9	**75.9**	73.7	86.1	86.5
Cufflinks	-	74.6	75.4	**74.6**	73.8	74.7

Table 2.6.: Excerpt of the Cuffcompare analysis for the simulated yeast data set. The F-measure accuracy is shown for GIIRA including ambiguous reads (GIIRA_w/_ambiguous), GIIRA excluding ambiguous reads (GIIRA_w/o_ambiguous), and Cufflinks. The best result for each category is marked in bold.

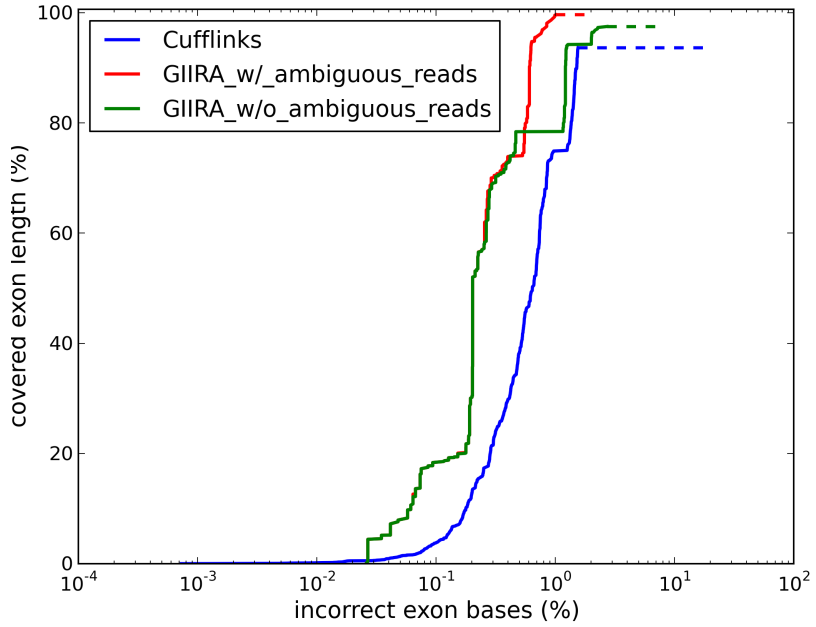

Figure 2.9.: ROC curve comparing the proportion of correctly and incorrectly predicted exonic bases for the yeast simulation. GIIRA was applied in two configurations: including ("GIIRA_w/_ambiguous_reads") and excluding ("GIIRA_w/o_ambiguous_reads") ambiguous reads. Dashed lines indicate the number of bases missed due to not identifying a reference exon. The proportion of false predictions is reported on a logarithmic scale.

2.6.5. Real data sets - *E. coli*

Table 2.7 shows the overall prediction accuracy expressed by the F-measure for the real *E. coli* experiment. Details on sensitivity and specificity are presented in Table A.4 in the appendix. For this data set the prediction accuracies are only relative measurements to compare the four methods, but cannot be regarded as absolute numbers since not all of the genes in *E. coli* are necessarily expressed at the same time. Thus, we included an additional alternative evaluation based on comparison against a reference subset including likely expressed genes (Table 2.7 (2)). We note that this subset does not necessarily reflect an exact ground truth since it is based on the RNA-Seq mapping for this specific experiment. Hence, we also show the evaluation against the complete reference.
As shown in Table 2.7 (1), as expected GLIMMER3 as the only compared method that exclusively predicts *ab initio* has the highest prediction accuracy in all compared categories on the complete reference. In comparison, the sensitivities of the RNA-Seq-based methods are significantly decreased. However, this changes in Tables 2.7 (2) and A.4 (2) for the comparison against the reference subset. Now, GIIRA shows a sensitivity comparable to GLIMMER3, which is accompanied by higher specificity. Thus, on base und locus level GIIRA achieves the best overall accuracy. Cufflinks and GeneMark both show low exon and locus level accuracy on

(1) *E. coli* real - complete reference

	exact measure			fuzzy measure	
	Base	Exon	Locus	Exon	Locus
GIIRA	74.1	42.3	47.3	42.9	47.9
Cufflinks	52.1	0.0	0.0	0.0	0.1
GeneMark	51.7	0.0	0.0	0.0	0.0
GLIMMER3	**95.6**	**69.8**	**70.8**	**70.9**	**72.0**

(2) *E. coli* real - reference subset

	exact measure			fuzzy measure	
	Base	Exon	Locus	Exon	Locus
GIIRA	**77.6**	43.3	**49.0**	43.9	**49.6**
Cufflinks	62.1	0.0	0.0	0.0	0.1
GeneMark	33.5	0.0	0.0	0.0	0.0
GLIMMER3	59.5	**44.5**	47.1	**45.3**	47.8

Table 2.7.: Excerpt of the Cuffcompare analysis showing the F-measure accuracy for the real *E. coli* data set compared against the complete annotated reference (1) and a subset of reference genes (2). The best result for each category is marked in bold.

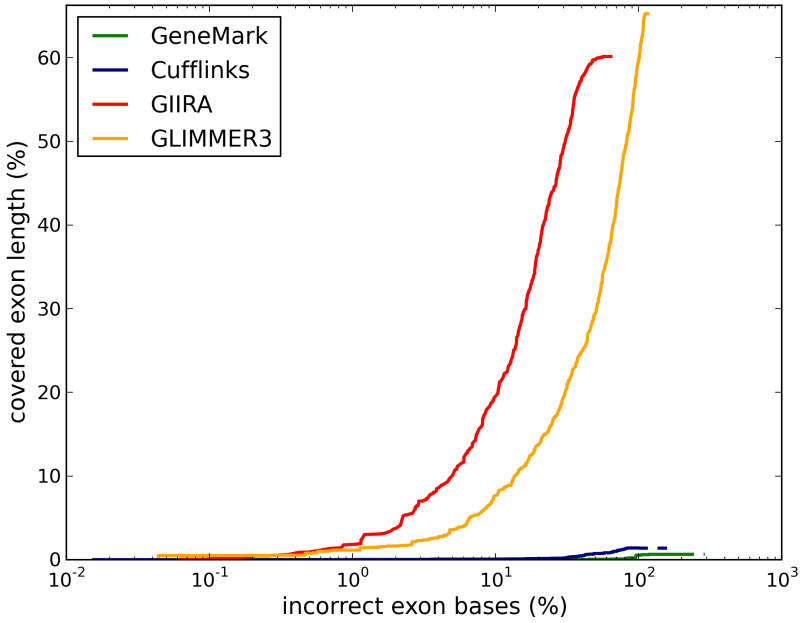

Figure 2.10.: ROC curve comparing the accuracy of predicted exonic bases of the four evaluated methods for the real *E. coli* data set, compared against the reference subset. Dashed lines indicate the number of bases missed due to not identifying a reference exon. The proportion of false predictions is reported on a logarithmic scale.

this data set, as illustrated in both Table 2.7 and Figure 2.10. They show comparable results on the base level, but fail to resolve the structural genes in the identified expressed regions.
GeneMark and GIIRA yield comparable results on the exact base sensitivity level compared against the complete reference. However, GIIRA is more specific since GeneMark covers large parts of the *E. coli* genome with operons without indicating the correct locus of the included genes. This is also reflected when comparing against the reference subset, where GeneMark showed reduced sensitivity and specificity (see Table A.4 (2)). Hence, overall the accuracy of GIIRA predictions is higher, it shows F-measures of 74.1 and 77.6 in contrast to 51.7 and 33.5 for GeneMark (for the complete reference and the subset, respectively).
As also illustrated in Figure 2.11, GIIRA outperforms Cufflinks and GeneMark on exon and locus level. GIIRA achieves a good prediction accuracy of the reference genes, while Cufflinks only predicts the expressed regions without indicating the included genes. GeneMark predicts operons, although these predicted regions also cover not expressed areas and can also span more than one operon (indicated by reference genes in different directions). GLIMMER3 performs well for actually ex-

Figure 2.11.: Exemplary excerpt of the gene predictions of GIIRA, Cufflinks, GLIMMER3, and GeneMark for the gene region starting at position 87,000 of the *E. coli* genome, illustrated in Geneious (Kearse et al., 2012). GIIRA (transcripts in red and genes in orange) achieves a good prediction accuracy of the grey reference genes (which overlap when shown in different rows), while Cufflinks (blue) only predicts expressed regions without distinguishing genes. GLIMMER3 (yellow) achieves a good prediction accuracy for actually expressed genes, although it also predicts not expressed genes (e.g., on the right-hand side) since it does not consider RNA-Seq information. GeneMark (green) predicts operons, although these predicted regions also cover non-expressed areas.

pressed genes, but since it is exclusively *ab initio*-based, it incorrectly predicts not expressed genes as well.
Table 2.8 shows an additional evaluation for the real *E. coli* data set for a sample size of 500 predictions for each compared method. Details on sensitivity and specificity are presented in Table A.4 (3) in the appendix. Note that the small sample size results from a low number of predictions from GeneMark (569) compared to prediction numbers above 1500 for the other methods. As expected, for all compared methods the overall sensitivity is smaller for the sampled subset than for the complete set of predictions. In contrast, the specificity increases, which also follows the expectation (because we sampled predictions according to their reliability). However, the overall accuracy is decreased compared to the accuracy observed for the complete set of predictions.

***E. coli* real - sampled set of predictions**

	exact measure			fuzzy measure	
	Base	Exon	Locus	Exon	Locus
GIIRA	29.1	**20.1**	**21.5**	**20.6**	**21.9**
Cufflinks	27.6	0.0	0.0	0.0	0.0
GeneMark	**47.7**	0.0	0.0	0.0	0.0
GLIMMER3	26.5	17.8	19.4	17.9	19.4

Table 2.8.: Excerpt of the Cuffcompare analysis showing the F-measure accuracy for the real *E. coli* data set compared against the complete annotated reference of 4,146 genes. The comparison is based on a sample of 500 predictions for each method. The highlighted numbers indicate the best results for each category.

Filtered and non-filtered analysis

Included in the files provided for the download of GIIRA is a script for filtering the predicted genes according to the information on coverage and ambiguous read support provided in the GTF result file. There are several filtering options available, which can be applied in different combinations, depending on the intended follow-up analysis. For instance, this allows to filter predictions that are exclusively or mainly supported by ambiguously mapping reads. In Table 2.9 the filtered and non-filtered results are compared for the real *E. coli* data set. Note that here "filtered" denotes that we applied the strictest possible filter.
Overall, for exon and locus level we see a significant improvement in specificity with applied filtering. In contrast, the sensitivity is reduced only slightly, for instance by 1.6% on the locus level. The differences in sensitivity are more pronounced on the base level, but also here the specificity is improved by filtering.

	Sensitivity		Specificity		fuzzy Sensitivity		fuzzy Specificity	
	not filt	filt	not filt	filt	not filt	filt	not filt	filt
Base	**70.8**	61.4	92.7	**93.3**	-	-	-	-
Exon	**43.7**	42.3	24.6	**42.3**	**44.3**	42.9	24.9	**42.9**
Locus	**46.3**	44.7	29.4	**50.2**	**46.9**	45.3	29.8	**50.8**

Table 2.9.: Cuffcompare analysis of the filtered (filt) and not filtered (not filt) gene predictions of GIIRA for the real *E. coli* data set. The highlighted numbers indicate the best results for each criterion (and for sensitivity and specificity, respectively) compared between both filtering options.

2.6.6. Real data sets - *S. cerevisiae*

Table 2.10 shows the F-measure analysis for the gene predictions of Cufflinks and the two configurations of GIIRA on the complete *S. cerevisiae* genome, compared to the NCBI reference annotation. Detailed values of sensitivity and specificity are presented in Table A.5 in the appendix.
GIIRA is the most sensitive of the three compared methods. However, on the locus level Cufflinks is more specific and shows higher overall accuracy. Overall, the two configurations of GIIRA show the best accuracy in all categories except locus level. As also shown in Figure 2.12, a loss in identifications can be observed when ambiguously mapped reads are disregarded, in particular the sensitivity in correctly predicting exonic bases is reduced by 8%. Interestingly, excluding ambiguous reads results in an improved specificity in intron predictions at the cost of slightly decreased sensitivity. With more than 80% correctly predicted exonic bases GIIRA shows the highest sensitivity, while both Cufflinks and GIIRA are comparable in specificity. Although overall the compared methods obtain a very low prediction accuracy on exon, transcript, and locus level, Table 2.10 (2) shows that the actual number of missed reference annotations is only 10% for GIIRA on exon and locus level compared to 20% for Cufflinks. Hence, most exons have been predicted at least partially. However, we observe a higher proportion of missed reference annotations on the intron level ($\approx$ 42%).
Table 2.11 shows the analysis for the real yeast data set with a sample size of 4,200 (a more comprehensive analysis is shown in Table A.6). As expected, the sensitivity of predictions is decreased compared to the complete prediction set. Since overall the prediction accuracy is comparably low, the effect of prediction sampling is not pronounced. The overall trends are similar as for the evaluation on the complete gene set. On the locus level the accuracy is slightly improved, while it is decreased on the base level.

(1) Accuracy

Methods	Base	Exon	Intron	Intron-Chain	Transcript	Locus
F-measure - exact						
GIIRA_w/_ambiguous	**81.0**	**0.3**	59.0	57.9	**0.1**	2.0
GIIRA_w/o_ambiguous	79.1	0.2	**62.4**	**61.2**	**0.1**	2.0
Cufflinks	76.2	0.0	58.3	57.8	0.0	**2.5**
F-measure - fuzzy						
GIIRA_w/_ambiguous	-	**0.3**	59.8	59.0	**0.1**	2.0
GIIRA_w/o_ambiguous	-	**0.3**	**63.1**	**62.3**	**0.1**	**2.1**
Cufflinks	-	0.0	59.6	0.0	0.0	0.0

(2) Missed and novel exons (in percent)

Methods	Missed exons	Novel exons	Missed intron	Novel intron	Missed loci	Novel loci
GIIRA_w/_ambiguous	**10.6**	11.5	**41.7**	35.7	**9.7**	11.3
GIIRA_w/o_ambiguous	12.6	10.5	42.1	**27.1**	11.5	10.4
Cufflinks	20	**5.5**	43.1	35.5	18.7	**4.6**

Table 2.10.: Cuffcompare analysis for the real yeast data for GIIRA including ambiguous reads (GIIRA_w/_ambiguous), GIIRA excluding ambiguous reads (GIIRA_w/o_ambiguous), and Cufflinks. Table (2) shows the proportions of completely missed and completely novel predictions. Best values for each category are marked in bold.

Methods	Base	Exon	Intron	Intron-Chain	Transcript	Locus
F-measure -exact						
GIIRA_w/_ambiguous	**64.8**	**0.2**	59.1	57.9	**0.1**	**3.1**
GIIRA_w/o_ambiguous	64.0	**0.2**	**61.0**	**59.7**	**0.1**	**3.1**
Cufflinks	63.0	0.0	59.0	58.3	0.0	2.9
F-measure - fuzzy level						
GIIRA_w/_ambiguous	-	**0.3**	59.8	59.1	**0.1**	**3.2**
GIIRA_w/o_ambiguous	-	**0.3**	**61.7**	**60.9**	**0.1**	**3.2**
Cufflinks	-	0.0	60.1	60.2	0.0	3.0

Table 2.11.: Overall prediction accuracy for the real yeast data set on a sample of 4,200 predictions for each compared method, evaluated against 5,905 reference transcripts. GIIRA was applied in two configurations: including ambiguous reads (GIIRA_w/_ambiguous), and excluding ambiguous reads (GIIRA_w/o_ambiguous). The best values for each category are marked in bold.

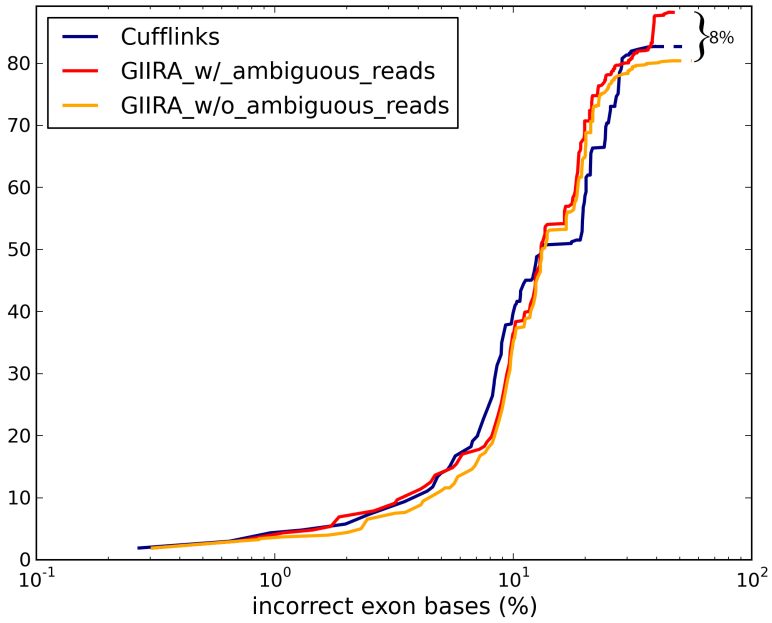

Figure 2.12.: ROC curve comparing the proportion of correctly and incorrectly predicted exonic bases for yeast chromosome 1, with GIIRA including ("GIIRA_w/_ambiguous_reads") and excluding ambiguous reads ("GIIRA_w/o_ambiguous_reads"). Including ambiguous reads increases the sensitivity by up to 8%, at constant specificity. Dashed lines indicate the number of bases missed due to not identifying a reference exon. Note that the proportion of false predictions is reported on a logarithmic scale.

2.6.7. System requirements

Table 2.12 shows the run time and peak memory required by GIIRA to predict the genes on the different data sets. The software was tested on a linux system with 48 threads and 256GB of available memory. We see that the run time and also the required memory increases with higher numbers of RNA-Seq reads (for exact read numbers see Table 2.3). Overall, due to the necessity to resolve operons, the system requirements are higher for prokaryotic data sets than for eukaryotic experiments.

	E. coli Sim	*S. cer* Sim	Human Sim	*E.coli* Real	*S.cer* Real
time	240sec	70sec	60sec	2.2h	3h
threads	1	1	1	5	15
RAM (GB)	6.4	2.6	1.8	62.0	20.0

Table 2.12.: Table representing the time and memory requirement of GIIRA for the three simulated data sets and for the two real data sets with *E. coli* and *S. cerevisiae* (*S. cer*), respectively.

2.7. Discussion

We introduced GIIRA as a gene finder that identifies potential coding regions exclusively based on mappings of reads from RNA-Seq experiments. Unlike other gene prediction methods, GIIRA also includes ambiguously mapped reads in the analysis, which improves on the prediction accuracy as demonstrated for various data sets with different levels of ambiguity. As shown in Section 2.6 in Table 2.3, already a comparably small number of ambiguous reads can substantially contribute to the ambiguity of a mapping. Disregarding this information leads to a loss in sensitivity, e.g., for genes sharing homologous regions or present in high copy numbers (refer to Section 2.6.3, Tables 2.5 and A.1, where including ambiguous reads increased the sensitivity of exon predictions by up to one third). We observe two interesting facts when comparing the results derived including and excluding regarding ambiguous reads: First, the intron predictions become more specific when ambiguous mappings are excluded, indicating that a number of erroneous introns is due to ambiguous split reads (refer to Tables A.1 and A.5). Second, as shown in Figure 2.8, the difference between prediction accuracy of the two configurations of GIIRA is more pronounced for lower scored genes. GIIRA calculates the prediction score according to the overall coverage, where each read contributes to the likeliness of the gene. Since ambiguous reads have less weight than unique reads (the number of ambiguous mappings determines the weight of read, the more mappings, the less weight is associated), exons with a high score are likely to have a high support of unique reads instead of ambiguous reads. This explains the almost identical results for the two configurations of GIIRA, including and excluding ambiguous reads, respectively. In contrast, exons with low scores are likely to be supported by a high number of ambiguous mappings. Thus excluding these mappings leads to incorrect identifications of only parts of genes or the loss of complete genes, such that in the lower score range we see decreased sensitivity when excluding ambiguous mappings.

GIIRA accurately predicts the correct structural genes for prokaryotic transcripts, as demonstrated in the two prokaryotic experiments. It identifies the most likely set of genes explaining the expressed region using an alignment scoring adaptation coupled with a linear program formulation. Thus, in comparison with existing approaches facilitating RNA-Seq integration, GIIRA has two major benefits: (i) it shows an overall increased prediction accuracy and (ii) it predicts structural genes themselves rather than focusing on operons such as GeneMark or transcripts without indicating start and stop codons such as Cufflinks.

Although GIIRA was primarily designed for prokaryotic gene prediction, it can also be applied to eukaryotic gene prediction as an addition to existing annotation pipelines or a complement to other gene finders. Eukaryotic data poses challenges different from prokaryotic data; instead of distinguishing operons and determining

gene structures, here many genes have alternative splice sites and various alternative isoforms are present. The complexity of alternative splicing events poses a critical challenge because GIIRA does not work with splice graphs to combine exons, but evaluates each splice site independently from others. As illustrated in Section 2.6.3, compared to the other methods GIIRA is very sensitive in predicting exons and transcripts. It also yields a high accuracy in predicting introns, but is less accurate in combining them to the correct intron-chain. For instance, a challenge arises for GIIRA if two alternative isoforms share an exon where one isoform ends with this exon and the other isoform proceeds with other exons. For GIIRA both isoforms appear to be continued with other exons and it assigns an incorrect intron-chain. Since Cufflinks uses a graphical approach to evaluate splice sites, it is less affected by this phenomenon. Hence, on the intron-chain level it shows higher prediction accuracy than GIIRA. AUGUSTUS, as a hybrid gene prediction method using non-ambiguous RNA-Seq mappings as external evidence, is less specific than the compared methods in regard to exon prediction but is superior in locus prediction.

Since GIIRA is exclusively based on RNA-Seq information, it predicts genes currently expressed in the organism of interest and thus does not necessarily provide a complete annotation of all encoded genes. Thus, unlike *ab initio* gene finders, it facilitates a sample-specific analysis. This is particularly shown in the real *E. coli* experiment (refer to Table 2.7). Here the reference gene set used for comparison resulted in a significant difference in prediction accuracy of the compared methods. Hence, when comparing to the complete reference, RNA-Seq-based gene finders are never as sensitive as *ab initio* gene finders. However, since not necessarily all genes are expressed at the same time, only evidence-based gene finders, such as GIIRA, are suitable to predict genes in a sample-specific way. Thus, when comparing predictions to a subset of likely expressed genes, GIIRA performed favorably to the other prediction methods, including the *ab initio* gene finder GLIMMER3. This demonstrates that GIIRA is particularly suitable for proteogenomic analyses since it provides a sample-specific and accurate gene model prediction.

GIIRA provides two frameworks to control the number of false positive predictions: (i) to filter contaminants and sequencing artifacts and (ii) to verify the reported gene predictions. It can identify regions with an extremely large coverage compared to the average coverage to be sequencing artifacts or other errors such as contaminants. In case of the real *E. coli* data set, this outlier identification filtered out most of the rRNA contaminants. Further, GIIRA reports additional information on coverage and ambiguous read support for each prediction. This enables an easy post-processing of the output allowing a trade-off of sensitivity and specificity adjusted to the intended follow-up analysis.

As reflected in the filtering experiment, a conservative filtering helps to control the number of false predictions and therefore increases the specificity. However, the

sensitivity might be reduced by the filtering approach, such that users interested in a highly sensitive prediction should rather prefer the non-filtered result over the filtered one. For the example shown in Section 2.6.5, the decrease in sensitivity is small in comparison with the increase in specificity, showing the high accuracy of the filters to select false predictions. However, users should carefully decide whether they want to prefer sensitivity over specificity or vice versa. Depending on the desired follow-up analysis it can be helpful to first select as many predicted genes as possible and then refine the result with other methods.
Note that although GIIRA is independent from any *a priori* information, it is possible to utilize such information (if present) to improve the prediction accuracy. For instance, if a reference annotation is already available, different runs of GIIRA can be compared using the Cuffcompare framework to identify an optimal parameter setting.

3. Postprocessing of gene predictions - towards more accurate identifications

Numerous efforts are dedicated to perform comprehensive gene model prediction, which is one of the basic steps in most genomic and proteogenomic analysis pipelines. The more accurate and tailored the underlying gene structures used for the construction of a spectra search database, the better the peptide and protein identification. However, despite sophisticated prediction methods gene identification still faces significant challenges handling complex gene structures, rare splice sites or mutations in genes (Goodswen et al., 2012; Ederveen et al., 2013). Thus, no single method exclusively provides a perfect and comprehensive prediction and each approach has advantages and disadvantages, making it suitable for certain analyses, but insufficient for other questions (Guigó et al., 2006; Goodswen et al., 2012).

For instance, sensitive *ab initio* methods are strongly dependent on training data and disregard experiment-specific mutations or expression levels. In contrast, sample-specific evidence-based methods are often limited by inaccurate and noisy measures (Mathé et al., 2002). Thus, combining predictions offers a possibility to complement the strengths of different strategies and balance their weaknesses. Hybrid approaches, such as AUGUSTUS (Stanke et al., 2006) and JIGSAW (Allen and Salzberg, 2005), therefore perform *ab initio*-based gene prediction assisted by empirical evidence. Types of evidence for instance include junction information from RNA-Seq experiments or protein alignments. However, the nature of prediction is still *ab initio* and adding raw evidence as information can only assist prediction but cannot count as a complete and meaningful prediction itself.

Hence, several methods have been developed that focus on the combination of gene models previously predicted by other gene identification software. This allows complementing the strengths of single method predictions to obtain the sensitivity of *ab initio* approaches, while incorporating other evidence to complete and verify identifications, as for example shown in Yok and Rosen (2011) and Ederveen et al. (2013). Amongst others, these approaches include methods by Allen et al. (2004), Elsik et al. (2007), Liu et al. (2008), and Haas et al. (2008).

Prior to this work, to the best of our knowledge, approaches combining predic-

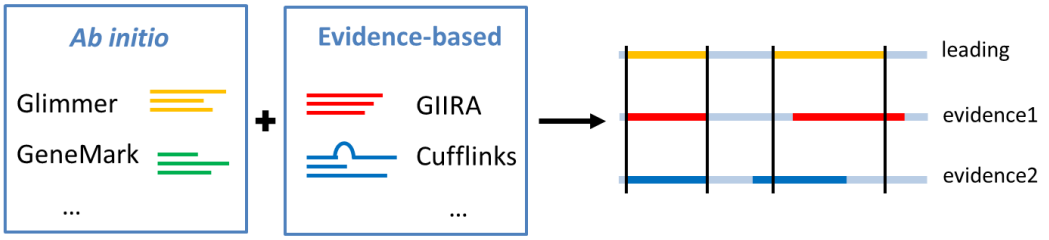

Figure 3.1.: The overall idea of the IPred method: Two lists of gene predictions are combined that include the output of *ab initio* and evidence-based gene finders, respectively. With default settings, the *ab initio* predictions are the leading predictions, which are complemented and evaluated based on the output of evidence-based gene finders.

tions treat all identifications independently of their prediction strategy and predominantly introduce weighting schemes to score different predictions. Hence, the full complementary potential of the combination of different prediction strategies is not fully tapped. Further, previous methods often focus on the integration of a specific set of gene finders (Shah et al., 2003). In addition, not all of the older methods facilitate the explicit integration of gene predictions based on all types of input data existing today, such as RNA-Seq. In recent years particularly RNA-Seq has become very popular since it offers new possibilities for the verification and revision of predictions with high coverage transcriptome information. Software that allows the integration of RNA-Seq information and that explicitly incorporates characteristics of RNA-Seq data (e.g., coverage variation at the ends of genes) is therefore highly desirable. To the best of our knowledge, other methods for prediction combination were developed before the main advent of RNA-Seq. Hence, they do not offer the explicit integration of RNA-Seq-based gene predictions and cannot take full advantage of its benefits (Murakami and Takagi, 1998; Pavlović et al., 2002; Yada et al., 2002).

Thus, we developed **IPred** (**I**ntegrative gene **Pred**iction), a method to integrate *ab initio* and evidence-based gene identifications to complement the advantages of different prediction strategies. As illustrated in Figure 3.1, IPred builds on the output of gene finders and generates a new combined set of gene identifications, representing the integrated evidence of the single method predictions.

In particular, IPred is independent of the evidence used to assist gene predictions. It incorporates prediction outputs based on the full plethora of evidence sources, for instance from EST libraries, protein alignments, sequence comparison, or from increasingly popular RNA-Seq runs. True positive identifications, for instance highly conserved genes, are likely to be present throughout different types of evidence, whereas false positive identifications are expected to only have weak support (only

by individual methods) and can thus be filtered out. IPred is a flexible and robust method that, in contrast to other methods, works independently from weighting schemes and does not require any *a priori* knowledge. Further, if a reference annotation is available, all predictions can be automatically evaluated using the framework provided by Cuffcompare (Trapnell et al., 2012).

3.1. Prediction combination

IPred accepts prediction output files in the commonly used GTF annotation format and provides converter scripts for a range of further file formats, for example the AUGUSTUS GFF format or the GLIMMER3 Predict format. The interpretation of GTF format styles can differ among methods. We decided to use the format supported by the Cufflinks/Cuffcompare suite (Trapnell et al., 2012) because we use Cuffcompare in the IPred pipeline.
When providing the output of gene finders, the user needs to categorize the different outputs into either *ab initio* or evidence-based (including comparative-based) predictions since IPred was particularly designed for combining complementary strategies. Also hybrid prediction methods and the results of annotation pipelines can be incorporated into IPred. For instance, if a hybrid method is *ab initio* in its nature, it should be specified as *ab initio*. When evidence has been integrated in the annotation pipeline, the result can be specified as evidence-based.
Note that it is not recommended to combine *ab initio* with *ab initio* methods since the underlying information, i.e. training sets or employed statistical models, might be very similar and thus could bias the combination of predictions. However, if an integration of two *ab initio* predictions is desired, one method can be classified as evidence-based. Here, it is necessary to keep in mind that potentially novel genes that are predicted by the *ab initio* method (that is classified as evidence-based) are genes that are not verified by external evidence.
Based on the categorization of each method, IPred first processes the loci of the predicted genes separately and then combines the loci of *ab initio* and evidence-based methods. IPred proceeds through the predicted *ab initio* loci (also called "leading" loci) and tests if an evidence-based prediction supports this identification. Per default, *ab initio* gene models are regarded as leading predictions, but it is also possible to instead use the evidence-based predictions as leading.
As illustrated in Figure 3.2, IPred distinguishes different types of prediction overlaps. Supported *ab initio* predictions are categorized into genes that perfectly overlap with at least one evidence-based prediction (Fig. 3.2 (1)) and weaker supported predictions that only show partial overlap (Fig. 3.2 (2)). Note that IPred per default accepts an overlap as a supporting overlap only if it is greater than a threshold of 80% of the length of the original *ab initio* prediction (calculated as the sum of

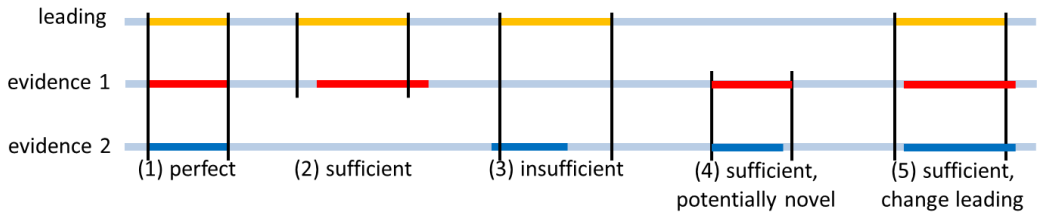

Figure 3.2.: IPred distinguishes different types for prediction combination: (1) a perfect overlap between *ab initio* and evidence-based predictions, (2) partial overlaps that are sufficiently verified depending on the specified overlap threshold, (3) partial overlaps that lack sufficient support, (4) novel annotations present in more than one evidence-based prediction, and (5) in case the evidence disagrees with the leading prediction, but is validated by other evidence, the combined gene model is changed according to the stronger support.

the number of nucleotides of its exons). The rationale for allowing also partially overlapping genes is that evidence-based methods might only incompletely predict a gene, e.g., due to low coverage in RNA-Seq experiments. Hence, requiring a perfect overlap could result in missed predictions. The threshold for overlap acceptance can be set by the user and is also adjustable to only accept perfect overlaps. In Section 3.6.5 we show that IPred is robust to different threshold settings.
If at least two evidence-based prediction outputs are available, the previously described merging process can be extended by also reporting genes that are not predicted *ab initio*, but instead have support from different evidence-based gene finders (Fig. 3.2 (4)). This way, potentially novel genes can be identified with greater confidence and also with respect to different approaches and sources for including external information (e.g., RNA-Seq evidence vs. EST evidence).
IPred scores the reported gene predictions depending on the quality of the overlap with other predictions. For each prediction g_i the score s_i is calculated as the number of overlapping bases l_i^{ov} divided by the total length of the prediction l_i:

$$s_i = \frac{l_i^{ov}}{l_i}.$$

Thus, a gene prediction with perfect overlap receives the highest score of 1. Genes that are only predicted by one of the compared methods, i.e. potentially novel genes, are written to additional output files corresponding to their prediction strategy and receive the lowest score of 0.

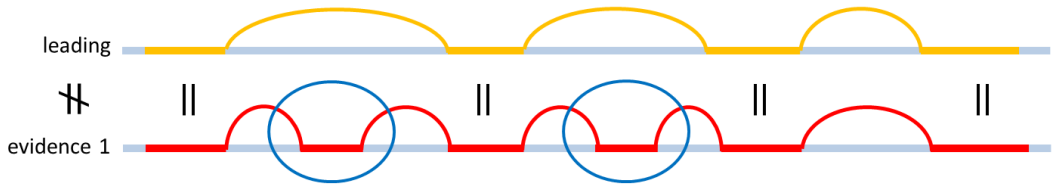

Figure 3.3.: Figure exemplifying the importance of similar exon chains for transcript combination. All exons of the leading transcript are covered by the evidence-based prediction. However, both transcripts differ in their exon chain due to additional exons in the second transcript (indicated by blue circles). Hence, both are likely alternative isoforms and do not support each other.

3.2. Alternative isoforms

IPred distinguishes between combinations of prokaryotic predictions and eukaryotic predictions since the structure of gene loci can differ significantly depending on the organism type. In contrast to prokaryotes, eukaryotes show splicing events and also alternative splicing resulting in alternative isoforms. This needs to be respected when merging eukaryotic gene predictions. Hence, for each gene locus all corresponding transcripts are processed separately. In addition, it is not only important that individual exons of a predicted transcript are supported by other methods, but that also the *exon chain* - all neighboring exons - is similar for compared transcripts (because differences indicate an alternative isoform).
Hence, IPred only considers a given exon as supported if the overlapping exon is part of a similar exon chain from a second prediction method (see Figure 3.3 for an example). Thus, a transcript is classified as perfectly supported only if all exons are matched perfectly by a different transcript. If all exons of a transcript are matched, but with minor differences (specified by the overlap threshold), the transcript is still regarded as supported, but it receives a lower score to indicate less agreement. If only a part of the exons of a transcript is matched, IPred analyzes if the overlapping transcripts predicted by other methods have stronger support (i.e. they differ from the leading transcript, but agree with each other). If this is the case, the leading transcript is regarded as incorrect and instead the overlapping transcripts with stronger support are taken into account. If the overlapping transcripts also disagree, the leading transcript is accepted only if the chosen overlap threshold is met by the number of matched exons (for the leading transcript as well as the overlapping other prediction). Since the original overlap threshold is defined as a percentage of nucleotides that need to be covered, the definition of the transcript overlap threshold t is adapted: The number of overlapping exons k must exceed

the fraction t of the total number of exons n that are part of the current transcript:

$$k \geq \lfloor t \cdot n \rfloor.$$

3.3. Output

IPred outputs a prediction file in GTF format that includes all genes supported by both prediction strategies, categorized by the reliability of each prediction. In addition, a tracking file reports the original gene predictions that generated each combined IPred prediction. Further, additional files reporting genes that were only supported by one strategy are provided, e.g., to allow the analysis of potentially novel or not expressed genes. In case a reference annotation is available, all predictions can be automatically evaluated using the framework provided by Cuffcompare (Trapnell et al., 2012) to allow for an easy comparison of different combinations of gene finders.
Currently, IPred returns predictions following the GTF format as interpreted in the Cufflinks suite, e.g., it does not specify untranslated regions (UTRs) or coding sequences (CDS). This is because currently the output formats of individual gene finders differ substantially, and often no UTRs or CDS are reported. Thus, to ensure a broad applicability, we decided to disregard these features and concentrated on gene loci and their corresponding transcripts and exons.

3.4. Implementation

IPred is implemented in Python (http://www.python.org/) and is an open source software that can be downloaded from http://sourceforge.net/projects/ipred/. For easy usability, IPred is available as a precompiled executable for Linux, Windows, and Macintosh.
In addition, we developed a GUI written in Java (http://www.java.com) to make IPred available to users that are not experienced in the use of command line software. A screenshot of the GUI is shown in Figure 3.4. The user can directly choose the directory of the input prediction files and log messages corresponding to each IPred run are directly visible in the GUI screen.
Since currently the GTF file format can differ significantly between gene finders (see above), the IPred suite also offers various converter scripts (also written in Python). These scripts convert the output of frequently used gene finders, such as GLIMMER3, GeneMark, or AUGUSTUS, to a GTF file format readable by IPred.

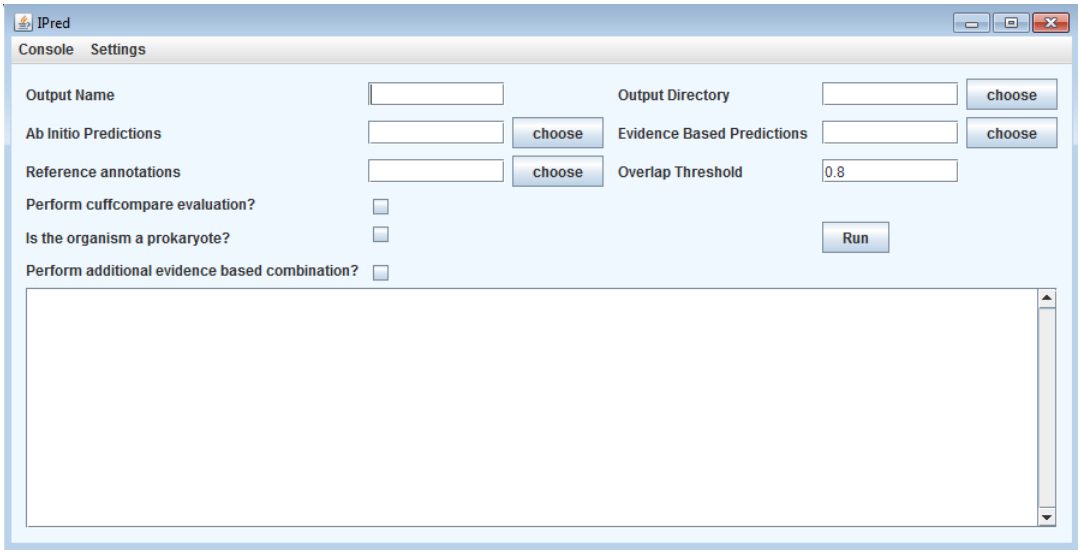

Figure 3.4.: Screenshot of the IPred GUI. Output directory and input files can be chosen with the help of drop-down menues, and log messages of IPred runs are printed directly to the GUI screen.

3.5. Experiments

We evaluated IPred in four experiments on *E. coli* (NCBI accession: NC_000913.3) and human data (NCBI accession: GRCh37). To compare the different methods on well-defined ground truth data, we not only used real but also simulated data sets in our evaluation. In the two experiments based on *E. coli* we combined predictions of the widely used *ab initio* gene finders GeneMark (Besemer et al., 2001) and GLIMMER3 (Delcher et al., 2007) and the evidence-based gene finders GIIRA (Zickmann et al., 2014) and Cufflinks (Trapnell et al., 2010).
In the first experiment we simulated RNA-Seq reads based on the NCBI reference annotation of *E. coli* as evidence information (for details see Section 3.5.1). In the second experiment we used real *E. coli* RNA-Seq reads (SRA accession: SRR546811) as evidence. The reads were mapped to the *E. coli* reference genome using the mapper TopHat2 (Kim et al., 2013) (for details see Section 3.5.2).
The eukaryotic experiments were also analyzed with Cufflinks and GIIRA, and additionally with AUGUSTUS (Stanke et al., 2006), a hybrid gene finder that facilitates the integration of evidence into its *ab initio* predictions. In the eukaryotic simulation we again used simulated RNA-Seq reads as additional evidence. Further, real RNA-Seq reads (SRA accession: SRR1654792) served as evidence for the human real data experiment.
GeneMark and GLIMMER3 were applied directly on the genomic sequence. To

generate GeneMark (GeneMark.hmm PROKARYOTIC, version 2.10f) predictions, we first applied the script "gmsn.pl" provided in the GeneMark installation and converted the resulting *ab initio* gene predictions to GTF format using the script "convertGeneMark.py" that is part of the IPred suite. To obtain GLIMMER3 (version 3.02) predictions, we used the script "g3-from-scratch.csh" provided in the GLIMMER3 installation that automatically defines a set of training genes that is used for prediction. The resulting .predict file was converted to GTF format using the IPred script "convertGlimmer.py". Both Cufflinks (version 2.0.2) and GIIRA were applied directly on the mapped RNA-Seq reads, using default settings. For the prokaryotic data sets, the prokaryotic mode of GIIRA was specified. To obtain hybrid gene predictions of AUGUSTUS (version 2.7), we followed the workflow recommended on the AUGUSTUS website[1] for integrating RNA-Seq evidence to AUGUSTUS, with specified "human" species model. Note that the use of pre-trained models might introduce a bias favoring the *ab initio*-based gene finders, due to possible similarities between training data and the data used in this study. However, the comparison of prediction combination methods is unaffected since all combinations are based on the same set of individual predictions.
The resulting single method predictions were combined by IPred and by the two state-of-the-art prediction combination methods Cuffmerge (Trapnell et al., 2012) (version 1.0.0) and EVidenceModeler (Haas et al., 2008) (version as of 25th June 2012). EVidenceModeler is an extension of the Combiner (Allen et al., 2004) idea and was shown to have superior performance to other existing combiners, such as GLEAN (Elsik et al., 2007) and JIGSAW (Allen and Salzberg, 2005).
In the prokaryotic simulation the predictions of GeneMark, GIIRA, and Cufflinks and additionally also GLIMMER3, GIIRA, and Cufflinks were combined. In the eukaryotic experiments AUGUSTUS was combined with GIIRA and Cufflinks. For the real *E. coli* data set, GLIMMER3 was combined with GIIRA and Cufflinks.
We applied IPred with default settings, specifying the prokaryotic configuration for the *E. coli* data sets. For the human real data set we specified an overlap threshold of 0.3 to balance variances of start and stop predictions between single methods. Cuffmerge was applied with default settings on an input file specifying the paths to the respective gene predictions.
Following the workflow recommended on the EVidenceModeler webpage[2], we created an evidence weights file to indicate the input predictions and their associated weights (with all weights set to be equal). The type of GeneMark and AUGUSTUS was specified as "ABINITIO_PREDICTION" and Cufflinks and GIIRA predictions were designated as "OTHER_PREDICTION" (because EVidenceModeler provides no explicit type for evidence-based predictions but instead recommends to

[1] http://bioinf.uni-greifswald.de/bioinf/wiki/pmwiki.php?n=IncorporatingRNAseq.Tophat
[2] http://evidencemodeler.sourceforge.net/

use "OTHER_PREDICTION" for complete gene predictions other than *ab initio*).
We additionally evaluated the influence of different weight settings for EVidence-Modeler: For each simulation experiment we performed two runs with EVidence-Modeler, one with equal weights (= 1) for each of the methods, and one with higher weights for the evidence-based predictions (5 for the prokaryotic simulation, 3 for the eukaryotic simulation because AUGUSTUS also received RNA-Seq hints) to consider their presumably higher reliability due to the use of RNA-Seq information.

3.5.1. Simulation Setup

As evidence information for the *E. coli* data set we simulated Illumina RNA-Seq reads with a length of 36bp based on the NCBI reference annotation. In this annotation the coding sequence of each known isoform appears as a consecutive sequence. Thus, the simulated reads also cover alternative isoforms and span introns (if existing in the data set). For an illustrative example, see Figure 2.6 in Section 2.5.1.
Note that only 70% of the annotated genes were used for evidence generation, simulating that not all genes are expressed at the same time. Therefore, we randomly picked 2,902 out of the present 4,146 annotations and used the chosen fasta sequences as input for the next-generation sequencing read simulator Mason (Holtgrewe, 2010). Before applying Mason, the set of annotated coding sequences was divided into 3 parts with 1,016, 1,451, and 435 genes, respectively. These three subsets were separately simulated with different coverages (5, 20, and 25, respectively), to obtain different gene expression levels in the subsequently combined set of simulated RNA-Seq reads.
Similar to the prokaryotic simulation, we simulated Illumina RNA-Seq reads with a length of 50bp based on the NCBI reference annotation for GRCh37 chromosomes 1, 2 and 3 (NCBI accessions: NC_000001.10, NC_000002.11, NC_000003.11). Also here we only simulated approximately 70% of the genes as expressed and generated varying coverage levels ranging from nucleotide coverage 5 to 20. This resulted in 5,318 genes that received RNA-Seq evidence (2,482, 1,488 and 1,348 for chromosome 1, 2 and 3, respectively), out of 7596 annotated reference genes (3,545, 2,126 and 1,925 for chromosome 1, 2 and 3, respectively).
Originally we intended to use the same read length for both simulated experiments, but due to few very short exons in the *E. coli* data set it was not possible to simulate 50bp reads for *E. coli* (the read simulator Mason resulted in an error when the read length exceeded the length of the gene). However, 50bp is a better reflection of current RNA-Seq read lengths than 36bp, so we decided to not reduce the length of reads in the human simulation.

3.5.2. Read mapping

We applied the read mapper TopHat2 (Kim et al., 2013) (version 2.0.8) to obtain a mapping of the RNA-Seq reads on the *E. coli* genome and the human chromosomes, respectively. We first indexed the reference sequence with Bowtie2 (Langmead and Salzberg, 2012) (version 2.2.1) and then called TopHat2 with default settings on the reference and the corresponding RNA-Seq reads in fastq format. The details of the resulting mappings are shown in Table 3.1.
The RNA-Seq mappings were then analyzed by GIIRA and Cufflinks to obtain the evidence-based gene predictions. Further, the mapping for the human simulation was used to generate hints for AUGUSTUS gene predictions.

data set	reads mapped	ambiguous reads	hits total	average cov.
E. coli simulation	1,187,830	16,019	1,253,150	16.6
E. coli real	10,052,045	8,555,561	57,769,265	17.0
human simulation	3,122,322	140,749	3,497,908	22.14
human real	126,914,607	9,340,757	142,753,401	293.0

Table 3.1.: Table showing the general properties of the TopHat2 mapping of the simulated and real data reads to the *E. coli* genome and to the human data sets. The column "average cov." specifies the average mapping coverage obtained with TopHat2.

3.5.3. Ground truth and evaluation

In both the prokaryotic and eukaryotic simulation the sample of genes selected as expressed serves as the ground truth annotation. All genes that are predicted and do not match this ground truth are regarded as false positives (also called "novel exons" in the Cuffcompare analysis), independent of the fact that the predicted gene locus might be present in the remaining NCBI reference genes (that are unexpressed in our simulation). This way we simulate condition-specific experiments, where mainly the fraction of genes that is indeed expressed is of interest.
Since not all genes of *E. coli* and *H. sapiens* are necessarily expressed at the same time, for the real data experiments we performed the evaluation by comparing against a subset of likely expressed reference genes. We note that this subset does not necessarily reflect an exact ground truth, but is only intended as an approximation of the real ground truth and serves as a basis to evaluate the performance of the compared methods. To obtain the subset, we mapped the RNA-Seq reads against the NCBI reference transcripts, using Bowtie2 (Langmead and Salzberg, 2012) (version 2.2.1) with default parameters (it was not necessary to use TopHat2 because the reference transcripts contain no introns and thus no split read map-

ping is required). Then we counted all reads mapping to each annotated gene and sampled a subset of reference genes comprising all annotations with a minimum overall mapping coverage of one. For *E. coli* this resulted in a ground truth sample of 2,680 reference genes instead of the original 4,146 annotations. For the human data set this resulted in 19,124 instead of 34,074 genes.
For all experiments we used Cuffcompare (Trapnell et al., 2012) to evaluate all single method predictions and combinations against the ground truth reference annotations. Details on the Cuffcompare metrics are explained in Section 2.5.3.

3.6. Results

3.6.1. *E. coli* simulation

Figure 3.5 and Table 3.2 (1) show sensitivity, specificity and F-measure (representing the overall prediction accuracy) for the single method gene predictions and different combinations generated by IPred. Note that for better visibility we included the Cuffmerge results only in the table, but not in the accompanying figure. Overall, IPred combinations show a significant improvement in specificity (e.g., from 63.8% to 98.1% for GeneMark only and GeneMark combined with Cufflinks), while also resulting in improved or comparable sensitivity. Also the number of missed and novel (not annotated in ground truth, hence false positive) genes is reduced when combining methods. GeneMark and GIIRA originally resulted in a high number of non-annotated predictions. However, when integrating both methods, the merged result shows a considerably reduced number.
Further, we see different effects on prediction accuracy depending on the evidence-based method combined with GeneMark predictions. For instance, the combination with Cufflinks shows a higher sensitivity and fewer missed exons than GeneMark combined with GIIRA. Although the combination of two gene finders already results in improved accuracy, the combination of all three methods produced even more accurate results. Further, when also genes missed by GeneMark but supported by both of the evidence-based methods are taken into account, we note an additional increase in sensitivity while showing comparable specificity. Overall, this IPred setup performs best of all compared methods (F-measure=97.4).
Independently of the chosen combination IPred outperforms EVidenceModeler and Cuffmerge with considerable increased sensitivity and specificity. Cuffmerge and in some cases also EVidenceModeler even results in smaller sensitivity and specificity compared to the single method predictions.

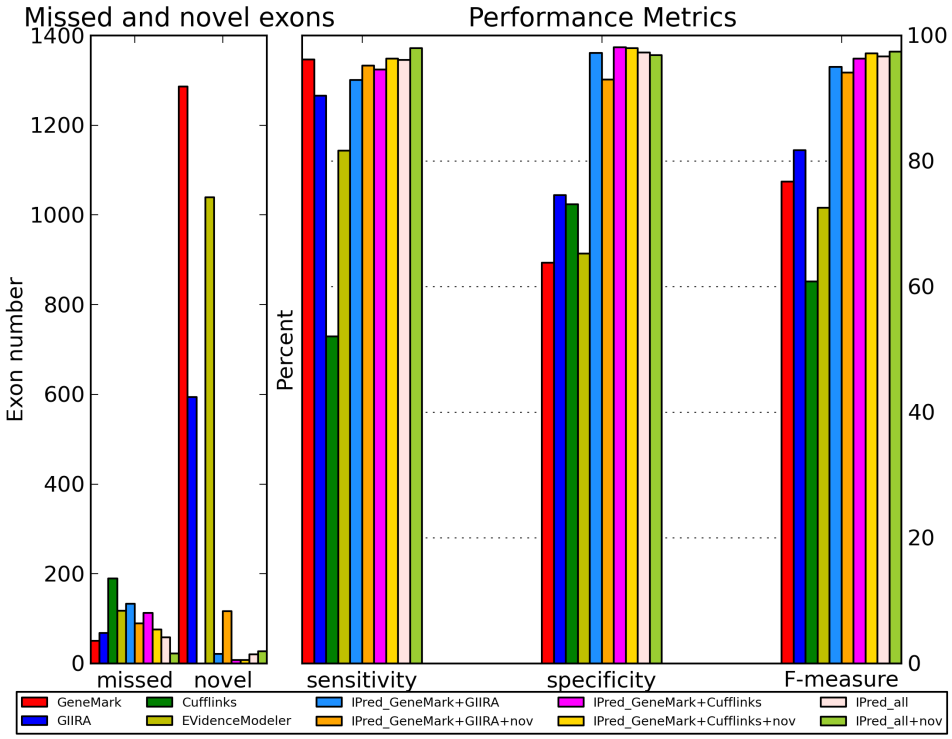

Figure 3.5.: Overview of Cuffcompare metrics for the predictions of single methods, EVidenceModeler and IPred combinations for the *E. coli* simulation based on GeneMark. Note that "IPred_all+nov" reports overall supported genes as well as genes missed by GeneMark, but supported by the evidence-based methods.

Table 3.2 (2) and Figure 3.6 show the combination of gene predictions based on the *ab initio* gene finder GLIMMER3. Note that for better visibility we included the Cuffmerge results only in Table 3.2 but not in the accompanying figure. Overall, this experiment shows the same trends as the GeneMark combinations, we see an improved prediction accuracy when combining different prediction strategies. We also note that combining three methods leads to more accurate results than combining two methods. Further, the compared methods EVidenceModeler and Cuffmerge are again outperformed by all IPred combinations. The sensitivity and specificity and thus also the F-measure are significantly higher for IPred predictions.
Compared to GeneMark, GLIMMER3 shows a slightly reduced prediction accuracy. This is also reflected in combinations with GLIMMER3, which have slightly lower F-measures than combinations based on GeneMark.

(1) GeneMark combinations

method	missed	novel	sensitivity	specificity	F-measure
GeneMark	51	1,286	96.2	63.8	76.7
Cufflinks	190	**0**	52.1	73.1	60.8
GIIRA	68	594	90.4	74.6	81.7
IPred_Cufflinks	113	8	94.6	**98.1**	96.3
IPred_Cufflinks+nov	76	8	96.3	98.0	97.1
IPred_GIIRA	133	21	92.9	97.2	95.0
IPred_GIIRA+nov	89	117	95.2	93.0	94.1
IPred_all	58	20	96.1	97.3	96.7
IPred_all+nov	22	27	**98.0**	96.9	**97.4**
EVidenceModeler	118	1,039	81.7	65.3	72.6
Cuffmerge	**3**	1,185	33.2	30.4	31.7

(2) GLIMMER3 combinations

method	missed	novel	sensitivity	specificity	Fmeasure
GLIMMER3	78	1,336	94.0	62.0	74.7
Cufflinks	190	**0**	52.1	73.1	60.8
GIIRA	68	594	90.4	74.6	81.7
IPred_Cufflinks	142	12	92.2	**96.5**	94.3
IPred_Cufflinks+nov	72	12	95.3	96.2	95.7
IPred_GIIRA	163	20	90.8	96.0	93.3
IPred_GIIRA+nov	87	119	95.0	91.8	93.4
IPred_all	88	20	93.9	96.1	95.0
IPred_all+nov	23	35	**97.6**	95.2	**96.4**
EVidenceModeler	127	944	81.4	66.9	73.4
Cuffmerge	**2**	1,218	32.6	29.6	31.0

Table 3.2.: Absolute numbers and percentages of the Cuffcompare evaluation of the exon level in the *E. coli* simulation. Note that all IPred combinations include either GeneMark (1) or GLIMMER3 (2). "IPred_all" denotes combinations with both Cufflinks and GIIRA. Note that combinations indicated with the tag "+nov" include genes exclusively predicted by the evidence-based methods. The best values for each category are marked in bold.

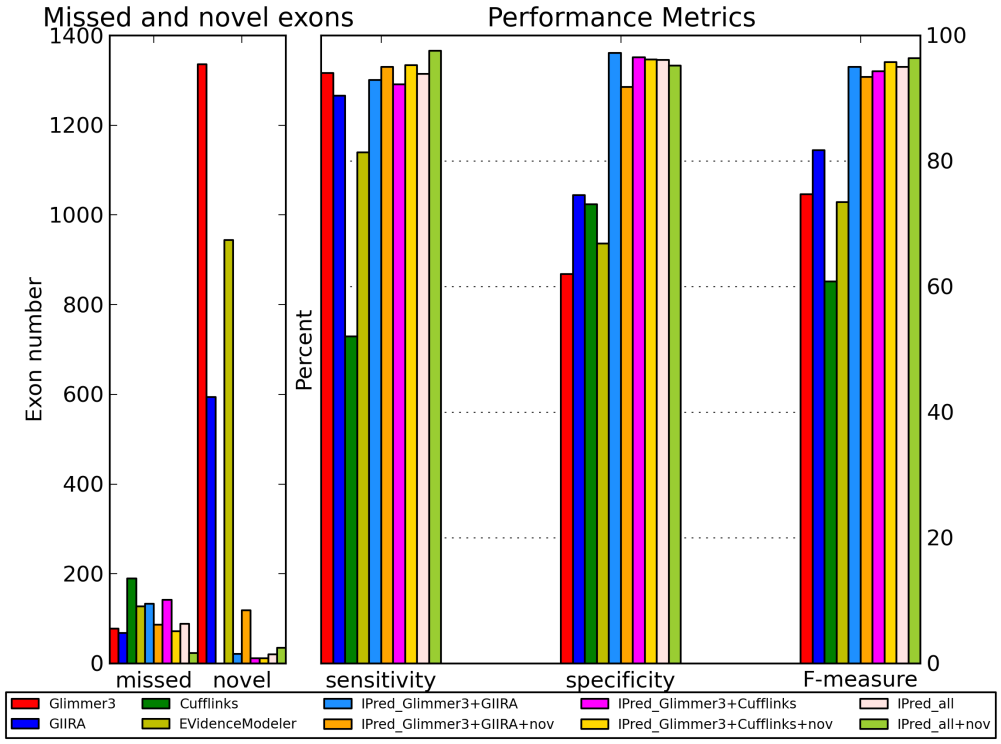

Figure 3.6.: Overview of Cuffcompare metrics for the predictions of single methods, EVidenceModeler and IPred combinations for the *E. coli* simulation based on GLIMMER3. Note that "IPred_all+nov" reports overall supported genes as well as genes missed by GLIMMER3, but supported by the evidence-based methods.

3.6.2. Human simulation

IPred was also evaluated on a simulation of a eukaryotic human data set. Figure 3.7 and Table 3.3 show the exon and transcript level comparison of the single method predictions and IPred, EVidenceModeler and Cuffmerge combinations. Overall, the performance on exon and transcript level significantly differs between methods. On the exon level, the sensitivity of IPred combinations strongly depends on the integration of novel predictions. If only predictions present in both AUGUSTUS and one or two of the evidence-based methods are taken into account, the sensitivity is considerably reduced compared to all other combinations. At the same time the specificity is on a comparable or higher level compared to other IPred combinations, and significantly higher than for EVidenceModeler and Cuffmerge. This results in an accuracy comparable to EVidenceModeler, but decreased in comparison to other IPred combinations and the single method predictions.

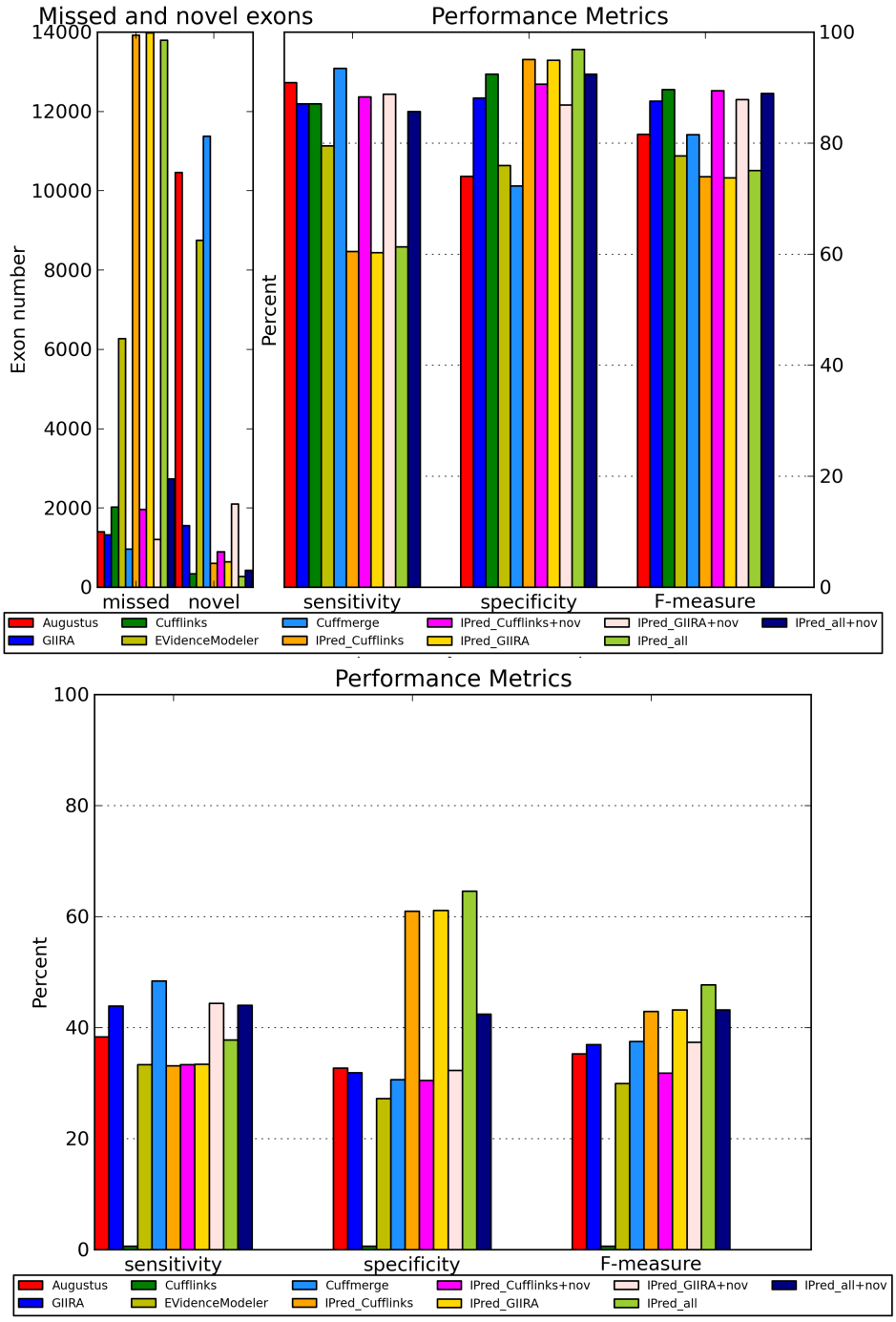

Figure 3.7.: Overview of Cuffcompare metrics for prediction accuracy on the simulated human data set. The upper figure shows the exon level comparsion, the lower figure the comparison on the transcript level. Note that "+nov" reports overall supported genes as well as genes missed by AUGUSTUS, but indicated by an evidence-based method.

Human simulation

			Exon			Transcript		
method	missed	novel	Sn	Sp	F	Sn	Sp	F
AUGUSTUS	1,401	10,458	90.9	74.0	81.6	38.3	32.7	35.3
GIIRA	1,320	1,551	87.1	88.1	87.6	43.9	31.9	37.0
Cufflinks	2,026	344	87.1	92.4	**89.7**	0.6	0.6	0.6
IPred_Cufflinks	13,921	600	60.5	95.1	74.0	33.1	61.0	42.9
IPred_Cufflinks+nov	1,965	890	88.3	90.6	89.4	33.3	30.5	31.8
IPred_GIIRA	13,977	640	60.3	94.9	73.7	33.4	61.1	43.2
IPred_GIIRA+nov	1,208	2,101	88.8	86.9	87.8	44.4	32.3	37.4
IPred_all	13,792	**275**	61.3	**96.9**	75.1	37.8	**64.6**	**47.7**
IPred_all+nov	2,736	430	85.7	92.4	88.9	44.0	42.4	43.2
EVidenceModeler	6,274	8,753	79.5	76.0	77.7	33.3	27.2	29.9
Cuffmerge	**965**	11,375	**93.5**	72.3	81.5	**48.4**	30.6	37.5

Table 3.3.: Cuffcompare evaluation of the exon and transcript level for the simulated human data set. Note that only missed and novel exons are reported by Cuffcompare, but not the numbers for the transcript level. Combinations indicated with the tag "+nov" include genes exclusively predicted by the evidence-based methods. All IPred combinations are based on AUGUSTUS predictions. "IPred_all" denotes the combination of AUGUSTUS with both Cufflinks and GIIRA. The best values for each category are marked in bold. Abbreviations: Sn = sensitivity, Sp = specificity, F = F-measure.

If predictions are included that do not overlap with AUGUSTUS identifications (indicated with the tag "+nov"), the sensitivity significantly increases, together with only slight decrease in specificity. Hence, these IPred combinations clearly outperform the result of EVidenceModeler. Also the Cuffmerge combinations are outperformed since the high sensitivity of Cuffmerge is accompanied with significantly lower specificity. Although including novel genes significantly increases the sensitivity, IPred shows a sensitivity only comparable to Cufflinks and GIIRA and thus performs comparable in regard to the overall accuracy.
On the exon level, IPred (including genes not fully supported by AUGUSTUS) provides more accurate results than EVidenceModeler and Cuffmerge and comparable results to the best single methods. On the transcript level, Cufflinks as the best performing method on the exon level shows almost no perfectly predicted transcripts. In comparison, IPred predictions show a significant increase in sensitivity and specificity. IPred again provides more accurate predictions than EVidenceModeler and Cuffmerge. Further, it also increases the accuracy of the single method predictions.

As an additional evaluation we compared the performance of the three gene prediction combination methods with regard to memory requirements and running time (tested on a linux system with 48 cores and 256GB of available memory). Table 3.4 shows the peak memory and overall time necessary to analyze and combine the single method predictions. IPred has the smallest memory and running time requirements of the three compared gene prediction combination methods.

Performance evaluation

combination method	overall time (s)	peak memory (MB)
EVidenceModeler	23,037	3,100
Cuffmerge	132	624
IPred	**59**	**215**

Table 3.4.: Overall running time (in seconds) and peak memory (in megabytes) for the compared gene prediction combination methods to analyze the simulated human data set.

3.6.3. *E. coli* real data set

We also evaluated IPred in an *E. coli* experiment based on real RNA-Seq evidence. Figure 3.8 and Table 3.5 show the results of the Cuffcompare evaluation against the subset of likely expressed reference annotations. Note that we excluded the Cuffmerge results from the figure to allow for better visibility.
Overall, IPred combinations show a pronounced increase in specificity and result in significantly improved prediction accuracy compared to all other methods. GLIMMER3 shows the highest sensitivity of all compared methods, which is accompanied by less specificity than IPred combinations. The single method predictions of Cufflinks and also the combinations of Cuffmerge show very low accuracy. We see that including predictions only indicated by one or more evidence-based gene finders results in an increase in sensitivity but also with a loss in specificity. Particularly the "Glimmer3+GIIRA+nov" combination shows significantly more novel exons than the combination excluding novel GIIRA predictions. However, although including novel evidence-based predictions reduces the specificity, also these IPred combinations are still more specific than all other prediction methods, including EVidenceModeler and Cuffmerge. Further, also the overall accuracy of IPred combinations is improved compared to other combination methods and the single method predictions.

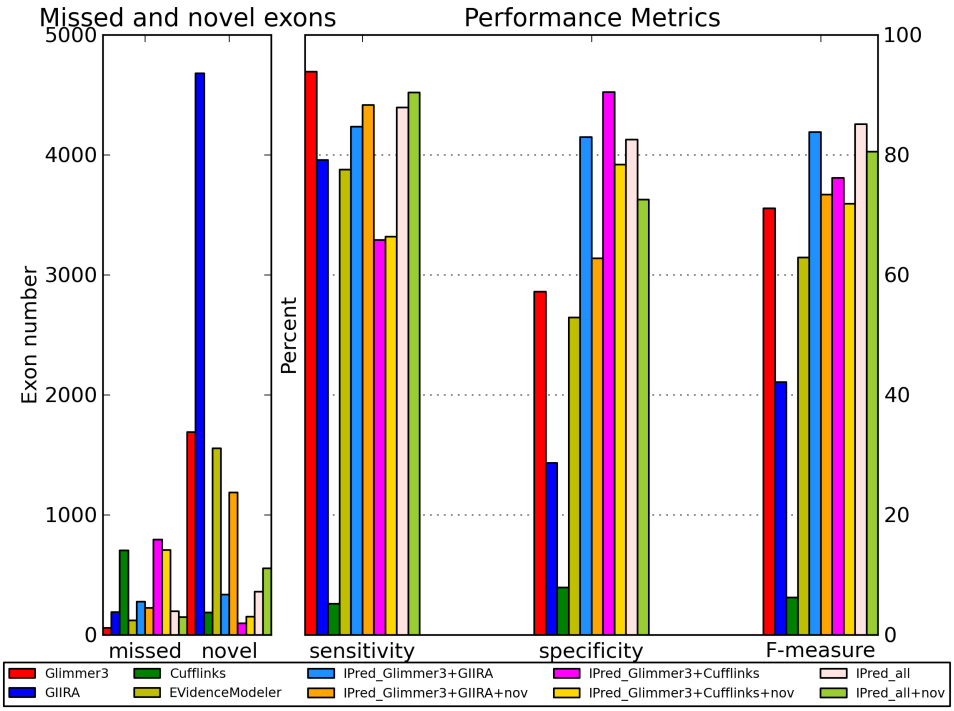

Figure 3.8.: Overview of Cuffcompare metrics for predictions on the *E. coli* real data set. Combinations indicated with "+nov" include genes exclusively predicted by the evidence-based methods.

***E. coli* real data set**

method	missed	novel	sensitivity	specificity	F-measure
GLIMMER3	59	1,692	**93.9**	57.2	71.1
Cufflinks	704	188	5.2	7.9	6.3
GIIRA	190	4,679	79.2	28.7	42.1
IPred_Cufflinks	796	**97**	65.8	**90.5**	76.2
IPred_Cufflinks+nov	709	154	66.4	78.4	71.9
IPred_GIIRA	279	338	84.7	83.0	83.8
IPred_GIIRA+nov	227	1,189	88.3	62.8	73.4
IPred_all	197	362	87.9	82.6	**85.2**
IPred_all+nov	151	556	90.4	72.6	80.5
EVidenceModeler	123	1,554	77.6	52.9	62.9
Cuffmerge	**11**	2,808	10.4	6.7	8.1

Table 3.5.: Cuffcompare evaluation for the *E. coli* real data set. Combinations indicated with "+nov" include genes exclusively predicted by the evidence-based methods. "IPred_all" denotes the combination of GLIMMER3 with both Cufflinks and GIIRA. The best values for each category are marked in bold.

3.6.4. Human real data set

The results of the evaluation on a complete human data set with real RNA-Seq reads are shown in Figure 3.9 and Table 3.6. On exon as well as transcript level AUGUSTUS shows the highest prediction sensitivity, while the IPred combinations (without including potentially novel genes) show the highest specificity. However, on the exon level the overall accuracy of AUGUSTUS predictions (79.9) is slightly higher than the accuracy of combinations by IPred based on Cufflinks or Cufflinks and GIIRA (79.6 and 77.2, respectively). On the transcript level the difference in sensitivity is not as pronounced as on the exon level. Hence, here the overall accuracy of IPred predictions (without potentially novel genes) is higher than for AUGUSTUS, due to the improved specificity of IPred.
Also on this data set including potentially novel genes resulted in higher sensitivity (on the exon level more pronounced than on the transcript level). However, at the cost of reduced specificity and overall reduced accuracy. In comparison with Cuffmerge and EVidenceModeler, IPred shows improved prediction accuracy, in particular on the transcript level. On the exon level, combinations by EVidenceModeler are comparable to IPred. Cuffmerge shows the highest exon level sensitivity of all combination methods, but at the cost of the lowest specificity.

Human real data set

			Exon			Transcript		
method	missed	novel	Sn	Sp	F	Sn	Sp	F
AUGUSTUS	13,212	44,753	**86.3**	74.3	**79.9**	**29.0**	25.0	26.9
GIIRA	52,160	78,092	43.6	38.1	40.7	0.0	0.0	0.0
Cufflinks	36,061	49,702	63.4	58.7	61.0	0.0	0.0	0.0
IPred_Cufflinks	38,053	15,255	74.3	**85.7**	79.6	25.6	**40.2**	**31.3**
IPred_Cufflinks+nov	33,159	53,719	75.9	65.3	70.2	25.6	11.9	16.2
IPred_GIIRA	69,753	**14,800**	58.4	83.7	68.8	22.4	38.4	28.3
IPred_GIIRA+nov	35,019	86,664	68.7	47.8	56.4	22.4	4.4	7.4
IPred_all	41,384	18,308	72.1	83.0	77.2	24.5	33.8	28.4
IPred_all+nov	39,733	29,112	72.5	77.1	74.7	24.5	21.1	22.7
EVidenceModeler	36,134	37,725	74.2	75.1	74.6	17.8	16.4	17.1
Cuffmerge	**10,896**	114,470	78.9	45.2	57.5	1.2	0.3	0.5

Table 3.6.: Cuffcompare evaluation of the exon and transcript level for the human real data set. Note that only missed and novel exons are reported by Cuffcompare, but not the numbers for the transcript level. "IPred_all" denotes the combination of AUGUSTUS with both Cufflinks and GIIRA. Combinations indicated with "+nov" include genes exclusively predicted by the evidence-based methods. The best values for each category are marked in bold. Abbreviations: Sn = sensitivity, Sp = specificity, F = F-measure.

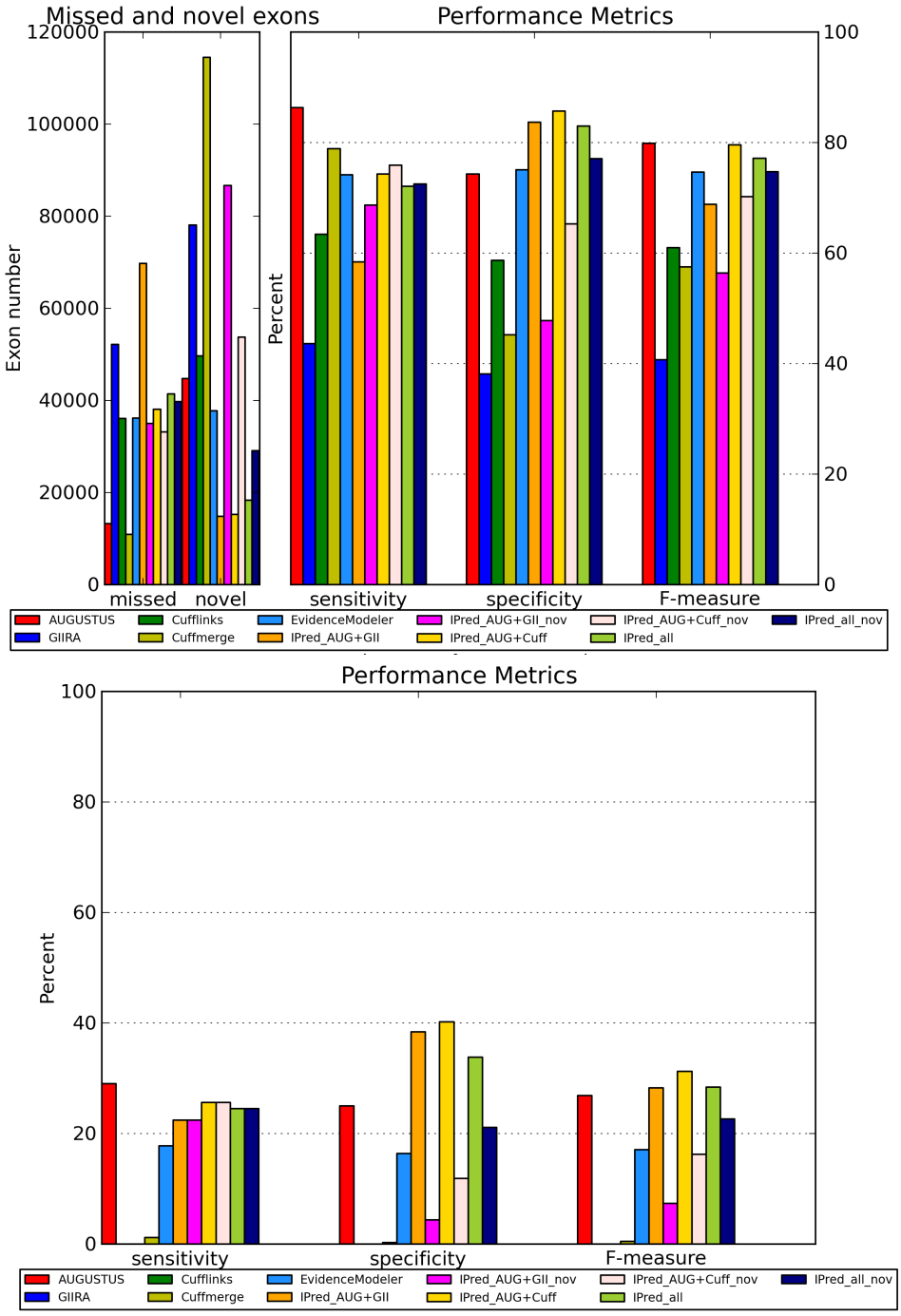

Figure 3.9.: Overview of Cuffcompare metrics for predictions on the human real data set. The upper figure shows the exon level evaluation, the lower figure the comparison on the transcript level. Note that "+nov" reports overall supported genes as well as genes missed by AUGUSTUS, but supported by the evidence-based methods.

3.6.5. Robustness to different overlap thresholds

Based on the simulated experiments with known ground truth, we analyzed the effect of different overlap thresholds on the performance of IPred. Tables 3.7 and 3.8 show the comparison between the default overlap threshold (80%) and an overlap threshold of 50%. Particularly for the *E. coli* data set we see that the differences between results obtained with the two overlap thresholds are only small and that the overall prediction accuracies of combinations are similar. As expected, with a smaller threshold the sensitivity for the combinations is slightly improved, while the specificity is slightly reduced. Combined, this results in very similar F-measures for both thresholds.
For the human simulation, the influence of the overlap threshold is more pronounced. Again, we observe an increase in sensitivity and a decrease in specificity when reducing the threshold. On the exon level the impact on the sensitivity increase is significantly more pronounced than in the prokaryotic simulation. We see considerable increases in sensitivity of up to 20% on the exon level. However, this effect does not carry on to the transcript level, where the increase in sensitivity is much smaller and also coupled with a significant loss in specificity (in range of 6.5% to 9.3%).

(1) *E. coli* simulation - GeneMark-based

threshold	method	missed	novel	sensitivity	specificity	F-measure
50	IPred_Cufflinks	**105**	11	**94.8**	97.9	**96.3**
80		113	**8**	94.6	**98.1**	**96.3**
50	IPred_GIIRA	**78**	31	**95.2**	96.9	**96.0**
80		133	**21**	92.9	**97.2**	95.0
50	IPred_all	**55**	30	**96.2**	97.0	96.6
80		58	**20**	96.1	**97.3**	**96.7**

(2) *E. coli* simulation - GLIMMER3-based

threshold	method	missed	novel	sensitivity	specificity	F-measure
50	IPred_Cufflinks	**136**	21	**92.3**	96.2	94.2
80		142	**12**	92.2	**96.5**	**94.3**
50	IPred_GIIRA	**107**	37	**93.1**	95.4	**94.2**
80		163	**20**	90.8	**96.0**	93.3
50	IPred_all	**85**	38	**94.0**	95.4	94.7
80		88	**20**	93.9	**96.1**	**94.9**

Table 3.7.: Comparison between an overlap threshold of 80% and 50% for the *E. coli* simulation. "IPred_all" denotes combinations with both Cufflinks and GIIRA. The best value for each category is marked in bold.

Human simulation

threshold	method	Exon missed	novel	Sn	Sp	F	Transcript Sn	Sp	F
50	IPred_Cufflinks	**5,619**	1,428	**80.7**	92.0	**86.0**	**36.3**	52.8	**43.0**
80		13,921	**600**	60.5	**95.1**	74.0	33.1	**61.0**	42.9
50	IPred_GIIRA	**5,256**	1,641	**81.4**	91.4	**86.1**	**36.7**	51.6	42.9
80		13,977	**640**	60.3	**94.9**	73.7	33.4	**61.1**	**43.2**
50	IPred_all	**9,963**	542	**70.4**	95.2	**80.9**	**40.6**	58.1	**47.8**
80		13,792	**275**	61.3	**96.9**	75.1	37.8	**64.6**	47.7

Table 3.8.: Comparison between an overlap threshold of 80% and 50% for the human simulation. All IPred combinations are based on AUGUSTUS predictions. "IPred_all" denotes the combination with both Cufflinks and GIIRA. The best value for each category is marked in bold. Abbreviations: Sn = sensitivity, Sp = specificity, F = F-measure.

3.6.6. EVidenceModeler - evaluation of different weight settings

On each simulated data set we performed two runs with EVidenceModeler: One with equal weights for all methods, and one with higher weights assigned to methods based on evidence, as recommended on the EVidenceModeler webpage. Tables 3.9 and 3.10 present the Cuffcompare metrics for the two runs of each experiment, compared against the known ground truth.
As shown in the tables, the EVidenceModeler predictions using equal weights have a slightly better accuracy than using unequal weights. For all data sets, sensitivity and specificity are improved with equal weights, and the number of missed and novel exons is reduced. Thus, the configuration based on equal weights is used for comparison with IPred combinations.

(1) *E. coli* simulation - GeneMark-based

weights	missed	novel	sensitivity	specificity	F-measure
equal	**118**	**1,039**	**81.7**	**65.3**	**72.6**
unequal	155	1,088	81.0	64.7	71.9

(2) *E. coli* simulation - GLIMMER3-based

weights	missed	novel	sensitivity	specificity	F-measure
equal	**127**	**944**	**81.4**	**66.9**	**73.4**
unequal	156	994	81.0	66.4	72.9

Table 3.9.: Absolute numbers and percentages of the Cuffcompare evaluation of the exon level for the different weight settings of EVidenceModeler on the simulated *E. coli* data sets. Best values for each category are marked in bold.

Human simulation

Exon					
weights	missed	novel	sensitivity	specificity	F-measure
equal	**6,274**	**8,753**	**79.5**	**76.0**	**77.7**
unequal	6,498	8,923	78.6	75.1	76.8
Transcript					
weights	missed	novel	sensitivity	specificity	F-measure
equal	-	-	**33.3**	**27.2**	**29.9**
unequal	-	-	30.0	25.0	27.3

Table 3.10.: Cuffcompare evaluation of the exon and transcript level for the different weight settings of EVidenceModeler on the human data set. Best values for each category are marked in bold. Note that only missed and novel exons are reported by Cuffcompare, but not the numbers for the transcript level.

3.7. Discussion

Despite the availability of sophisticated gene prediction methods, they all have different biases. Thus, we developed IPred to combine results of different prediction strategies and thereby improve the accuracy of single method predictions. This makes IPred a valuable addition to proteogenomic workflows because it can be used as a post-processing method to provide more accurate gene prediction-based databases. We stress that IPred is not intended as a novel gene finder but rather as an easy-to-use post-processing software to verify predictions and filter out false positives. Therefore, it strongly depends on the quality and performance of the input gene finders, but is independent of the underlying data sets or the nature of the information used for evidence-based prediction. Thus, IPred in general facilitates the detection of rare or hard-to-predict events, for instance genes following a non-standard coding scheme, as long as at least some of the input gene finders predict those events.

The dependency on input gene predictions is particularly shown in the human simulation experiment (refer to Section 3.6.2.). Here the evaluation shows considerable differences in sensitivity between combinations including and excluding "novel" predictions. The reason for the observed differences is that AUGUSTUS often reports a transcript with an incorrect first or last exon (i.e. it reports an additional exon, data not shown). This is also reflected in the high number of novel exons predicted by AUGUSTUS and in its low specificity. Though a detailed analysis of this phenomenon is beyond the scope of this work, a likely explanation is that the additional exons might be an artifact of the *ab initio*-based prediction employed by AUGUSTUS (that also predicts genes that are not expressed in our simulation). Hence, in combinations with Cufflinks and GIIRA the exon chains of the compared

methods disagree and none of the predictions appears to be sufficiently supported. Although including novel genes significantly increases the sensitivity, IPred is still affected by the discrepancies between AUGUSTUS and evidence-based predictions because it shows a sensitivity only comparable to Cufflinks and GIIRA and therefore is only comparable in the overall accuracy.
However, as shown for all compared data sets, IPred succeeds in improving single method predictions with its combination approach. Although every improvement is eventually limited by the performance of the input gene finders, the overall accuracy is almost always increased. Also the number of false predictions (in the simulations true and false positives are known) is reduced by IPred, as for instance reflected in the *E. coli* simulation in the reduced numbers of "novel" predictions of IPred combinations compared to the results of GIIRA, GLIMMER3, and GeneMark. This indicates that erroneous predictions are filtered out during the merging process because an erroneous prediction by one of the methods is almost always not present in the other method.
Naturally, IPred combinations that include predictions of evidence-based methods that received no *ab initio* support do not benefit from this filtering process, as reflected in their reduced specificity (compared to the combination not including novel genes). However, when including more than one evidence-based method, this effect is outweighed: The increase in sensitivity is still accompanied by a decrease in specificity, but the reduction is not as pronounced as in the combinations including novel predictions predicted by only one evidence-based method. This indicates that combining two or more evidence-based methods is a suitable strategy to further verify predictions and to avoid a loss in specificity that accompanies a simple integration of all novel predictions. An exception is the two-methods combination based on Cufflinks in the *E. coli* simulation. Here the loss in specificity is only minor for the combined predictions. This reason for the effect is that for this data set Cufflinks predicted no completely novel exons, and thus no genes that are regarded as false positives are added in the "IPred_Cufflinks+nov" combination. Additional erroneous predictions only arise if Cufflinks predicts an exon that does not perfectly match the reference annotation.
Another example for the benefit of IPred is the performance of Cufflinks and the different Cufflinks combinations in the human simulation. Cufflinks as a single method is the most accurate method on the exon level, but the least accurate method on the transcript level. This is due to the fact that Cufflinks is very accurate at predicting intermediate exons but does not predict start and stop codons. Thus, beginning and end of a transcript almost never match the reference annotation, leading to reduced performance in the evaluation. Here, IPred is very useful because it complements the overall exon accuracy of Cufflinks with the start and stop prediction accuracy of other methods. This is reflected in the considerable increase in transcript level sensitivity and specificity of Cufflinks-based combinations compared

to Cufflinks alone (together with only slightly decreased exon level accuracy).
Also in the prokaryotic experiments IPred significantly improves the accuracy of Cufflinks predictions. As shown in Section 3.6.1, IPred combinations achieve over 30% higher F-measures than Cufflinks alone. Cufflinks does not predict structural genes but only the expressed transcript, which is insufficient for the operon organization in prokaryotes. Hence, its original sensitivity and specificity are comparably low, but are significantly increased when combined with other methods predicting structural genes. On this data set we also see different effects on prediction accuracy depending on the evidence-based method combined with *ab initio* predictions. For instance, Cufflinks appears to provide a set of predictions that is more complementary to GeneMark than GIIRA predictions because the combination of GeneMark and Cufflinks shows a higher sensitivity and fewer missed exons than GeneMark combined with GIIRA.
On the *E. coli* real data set, interestingly the *ab initio* method GLIMMER3 shows a significantly higher sensitivity compared to all other methods, including the combinations (refer to Section 3.6.3). Since none of the approaches that include RNA-Seq evidence show a comparable sensitivity, this is likely due to the choice of the ground truth annotation set that might still contain genes that are not expressed but are rather mapping artifacts. Here, including other evidence, such as protein alignments, might further increase the accuracy of combined predictions. In addition, on this data set Cufflinks and also Cuffmerge show very low accuracy, which indicates that they are more suitable for application on eukaryotes than on prokaryotes.
In the experiment based on human real data, interestingly the combination with both Cufflinks and GIIRA (including novel genes predicted by both methods) resulted in similar levels of specificity reduction as the combinations with only one evidence-based method (including novel evidence-based genes). This indicates that although Cufflinks and GIIRA agree on certain expressed regions, these predictions still require further analysis to ensure that they are not mapping artifacts. However, these regions might also hint to novel genes, but additional evidence, for instance from ESTs or protein libraries, would be necessary for further verification.
We also show that IPred is robust regarding the choice of the overlap threshold (see Section 3.6.5). The overall accuracy of IPred combinations remained similar regardless of overlap choice, in particular for the prokaryotic simulation. In the human simulation the influence is more pronounced, although only on the exon level. The stronger effect on the exon level can be explained by the additional exons predicted by AUGUSTUS. Here, the exon chains of Cufflinks and GIIRA predictions do not match the prediction of AUGUSTUS. Reducing the overlap threshold results in more matches since unequal exon chains are more readily accepted.

4. Integrative proteogenomics beyond six-frames and single nucleotide polymorphisms

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. (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 (Castellana et al., 2008; Fanayan et al., 2013; Ahn 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 (Li et al., 2010; Ahn et al., 2013). Six-frame translation has the advantage of being independent from any *a priori* annotation of the nucleotide sequence. However, it introduces an artificial six-fold increase of the (unknown) target database, which can result in a bias in peptide identification (Reiter et al., 2009; Blakeley et al., 2012; Jeong et al., 2012; Branca et al., 2014).

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. 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 (Ning and Nesvizhskii, 2010; Wang and Zhang, 2014; Krug et al., 2014; Safavi-Hemami et al., 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 integrating variations or splicing information to the database (Wang et al., 2011; Woo et al., 2013; Wang and Zhang, 2013). These approaches are either only suitable for eukaryotes (having splicing events) or are still only seen as

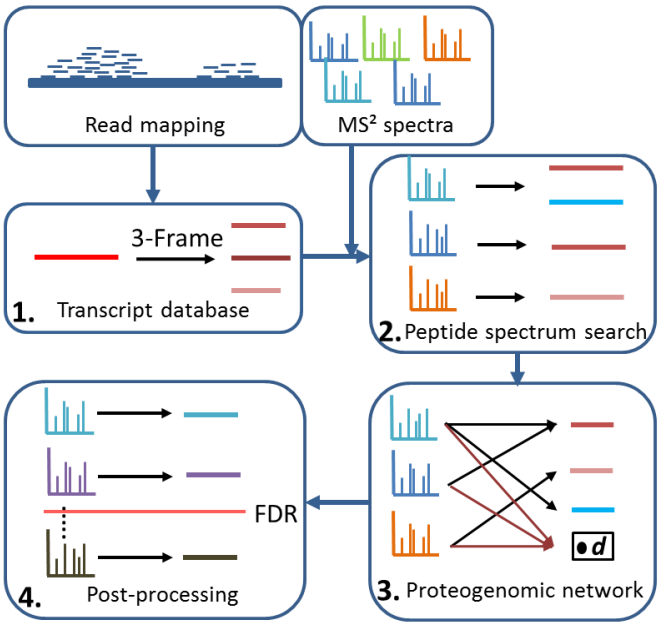

Figure 4.1.: The overall workflow of MSProGene. (1.): An RNA-Seq read mapping is analyzed to infer transcript sequences, which (2.): provide the database for spectra search. (3.): The resulting peptide spectrum matches are represented by a network, which is analyzed to resolve protein inference and to select the correct frame per transcript. (4.): Finally, peptide identifications are controlled with regard to their false discovery rate (FDR).

an extension or refinement of the standard approach that uses 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 (Serang et al., 2010; Claassen, 2012; Huang et al., 2012). 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.
We overcome current limitations by introducing 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. These sample-

specific databases avoid unnecessary enlargement by six-frame translations and increase 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. Figure 4.1 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 reassigning 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.).

4.1. Transcript database and spectra search

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).
MSProGene analyzes the read mapping and refines the transcript sequence according to mutations present in the RNA-Seq reads (refer to Figure 4.2). 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, threshold can be specified by the user) and (ii) it is supported by the majority

Figure 4.2.: An example for the introduction of SNPs present in an RNA-Seq read mapping to a transcript sequence (which is the region between the vertical black lines). Only the orange SNP is integrated in the transcript sequence, the green ones are either only present in one read or do not have the majority support.

of the reads. If the user intends to use an *a priori* defined database for the peptide spectrum search, MSProGene can also be provided with custom sequences in fasta format, without the need for RNA-Seq evidence. 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 is for instance 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. The user can specify a minimum peptide length for spectra search (per default 5 amino acids), and thus 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 s^g (or score from other prediction methods) 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:

$$c^t = \frac{n^t \cdot l^r}{l^t}.$$

The initial transcript score s^t is normalized over the 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:

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

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

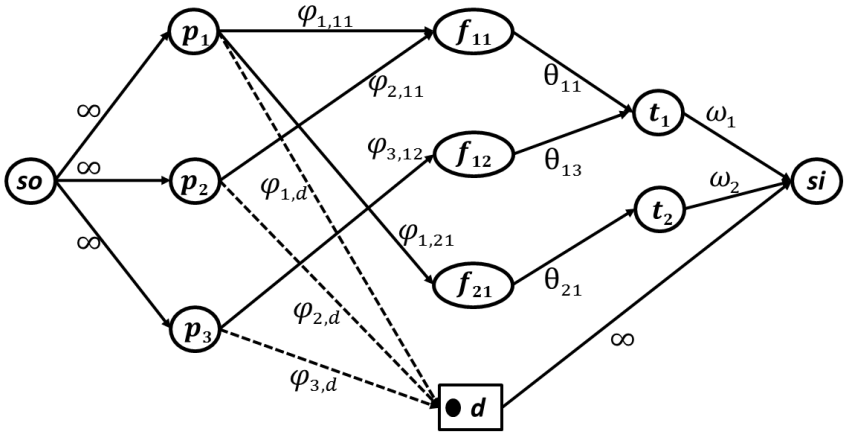

Figure 4.3.: Simplified example of a proteogenomic network: peptide nodes p_i are connected to the frames f_j they map to, and all sister frames are connected to their corresponding transcript node t_k. A so called *dummy* node d ensures that incorrect peptide identifications can be reassigned. All edges are labeled according to their capacity indicating the support from experimental data for a connection between 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.

adapted to also work with other search methods. After the search, the resulting peptide spectrum matches 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 peptide spectrum match score provided by the search engine is extracted, and normalized to the $[0, 1]$ interval.

4.2. 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 peptide spectrum matches are correct. Thus, (iii) we have to detect and remove incorrect identifications.

To meet these three objectives we first represent all peptide spectrum matches a network, which is optimized in order to solve the inference. The network $G = \{N, E\}$ (depicted in Figure 4.3) 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 that connects 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 4.3: 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.
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. In addition, the capacity 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 ($y_{p_i,f_j} = 0$):

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

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 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:

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

Two additional constraints ensure that only one match per peptide (Eq. 1) and only one frame per transcript (Eq. 2) is selected, respectively:

$$(1) \quad \sum_{j} y_{p_i,f_j} = 1 \;\; \forall\, i | p_i \in P,$$

$$(2) \quad \sum_{j} m_{f_j,t_k} = 1 \;\; \forall\, k | t_k \in T.$$

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:

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

where $\lambda_{p_i,d}$ corresponds to a penalty term equivalent 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), which equals the negative of the confidence score of the peptide spectrum match, 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:

$$\sum_{i | e_{p_i,f_j} \in E} \varphi_{p_i,f_j} \leq \theta_{f_j,t_k} \;\; \forall\, e_{f_j,t_k} \in E.$$

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 introduced penalty this connection is only chosen if the penalty is outweighed 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.

4.3. Post-processing

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 are expected to 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 cumu-

lative density functions $F_T(s_x^p)$ for N_T and $F_D(s_x^p)$ for N_D:

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

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

4.4. 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 post-processing and visualization, the coordinates as well as the confidence score and the 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 peptide spectrum matches (set denoted as P^t):

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

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).

4.5. Implementation

MSProGene is implemented in Java (http://www.java.com) and uses helper scripts that are written in Python (http://www.python.org/) and Gnu R (http://www.r-project.org/). MSProGene is open source and the software can be downloaded from http://sourceforge.net/projects/msprogene/. Not only the source code, but also a precompiled package is available.
To speed up the spectra search and analysis of peptide spectrum matches, MSProGene is designed to run parallel on several processing units, if available.

4.6. Experiments

4.6.1. 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 data sets). 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) specifying 10 tandem mass spectra per retention time bin, a gradient of 3,000s, an instrument resolution of 200,000, 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.

4.6.2. *Bartonella henseale*

MSProGene was also 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: 0.7.0a). 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., 2009a) 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*) that is 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 "genomeguided" mode (using the BFAST mapping as a guide). The resulting assembly contained 1,907 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 1,488 annotated *B. henselae* proteins available at NCBI. In addition, we performed a standard search on a database including SNPs indicated by a samtools *mpileup* (Li et al., 2009a) 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 the 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. 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. 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 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.
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.

4.6.3. *Litomosoides sigmodontis*

We also compared MSProGene to a six-frame based analysis on a *L. sigmodontis* data set (assembly nLS.2.1 from www.nematodes.org). *L. 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). The analyzed 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 data sets. Hence, since at the time of this study only low coverage 454 transcriptome data was available for *L. sigmodontis*, we chose Illumina 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 200bp.
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, 1,042 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 BLAST bit score cutoff of 50. Note that although often the BLAST E value is used for evaluation, we did not use the E value in our analysis to allow a fair comparison. An E value threshold may have favored the evaluation towards MSProGene because it has a smaller query database size than the six-frame translation.

4.6.4. Peptide search parameters

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

4.7. Results

4.7.1. Algorithm evaluation

We analyzed the peptide spectrum matches before and after the network optimization of MSProGene. Details are shown in Figure 4.4. Of 21,715 spectra that MSGF+ matched to the original protein (sometimes among multiple proteins), 21,617 were assigned correctly by MSProGene (99.5%). Overall, the algorithm correctly reassigned over 90% of the spectra that had multiple protein hits (933 of 1,031).

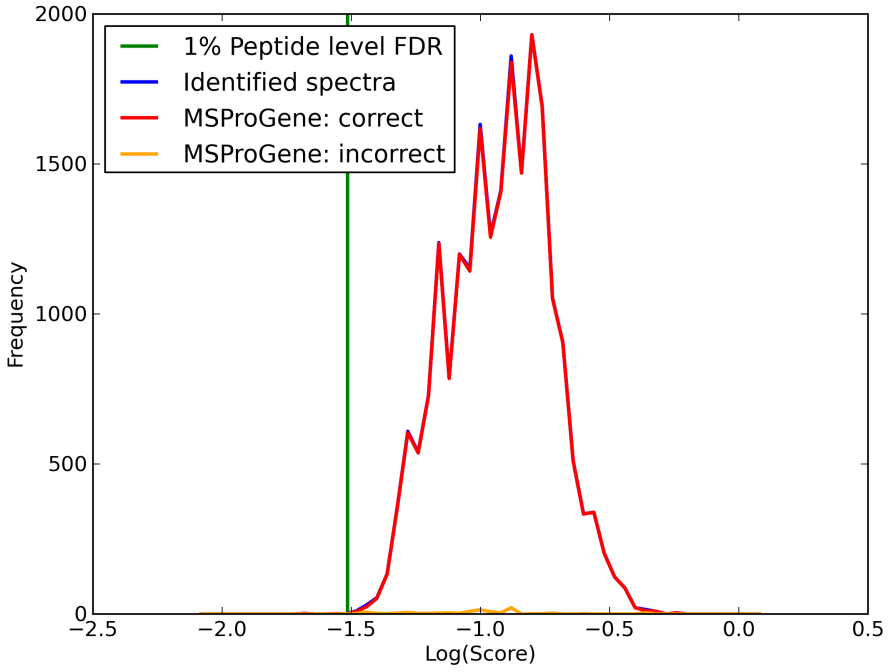

Figure 4.4.: Figure illustrating the distribution of peptides correctly and incorrectly reassigned by MSProGene. 99.5% of the peptides were assigned to their original ground truth protein.

4.7.2. *B. henselae* data

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 4.1, the transcript database constructed for spectra search contains 1,568 sequences. This number is reduced to 1,397 when taking spectra support into account. This leads to a decrease in recall from 78.2% to 76.5%. In contrast, the precision strongly increases from 79.0% to 85.1% when spectra support is taken into account.

	without spectra	with spectra	with spectra, without single hits
Predicted	1,568	1,397	1,286
#matches to annotation	1,238	1,189	1,143
#identified annotations	**1,164**	1,139	1,109
Recall (%)	**78.2**	76.5	74.5
Precision (%)	79.0	85.1	**88.9**

Table 4.1.: Prediction results of MSProGene, exclusively based on RNA-Seq, verified by spectra support, and in addition excluding proteins with only one spectrum hit. Evaluation on the *B. henselae* data set, compared to the reference annotation comprising 1,488 genes. The best value for each accuracy measure is marked in bold.

Comparison to reference-free methods:

For the three compared methods 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 4.2.
The transcript database constructed for spectra search by MSProGene contains 1,568 sequences. This is significantly smaller than the number of sequences searched in the Six-frame analysis and Assembly, which shows the suitability of RNA-Seq data to provide smaller and more tailored search databases.
Overall the Six-frame approach predicts the highest number of spectra-supported genes and also achieves the highest recall given the peptide level FDR. However, this is at the cost of specificity: Six-frame has 3.7% higher recall but 8.5% less precision than MSProGene (also refer to Figure 4.5). 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.

	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

Table 4.2.: Prediction of reference-free methods on the *B. henselae* data set, compared to the reference annotation with 1,488 genes. The row indicated as "1%-AnnotationFDR" shows results for an additional 1% annotation-based FDR on the protein level. The best value for each accuracy measure is marked in bold.

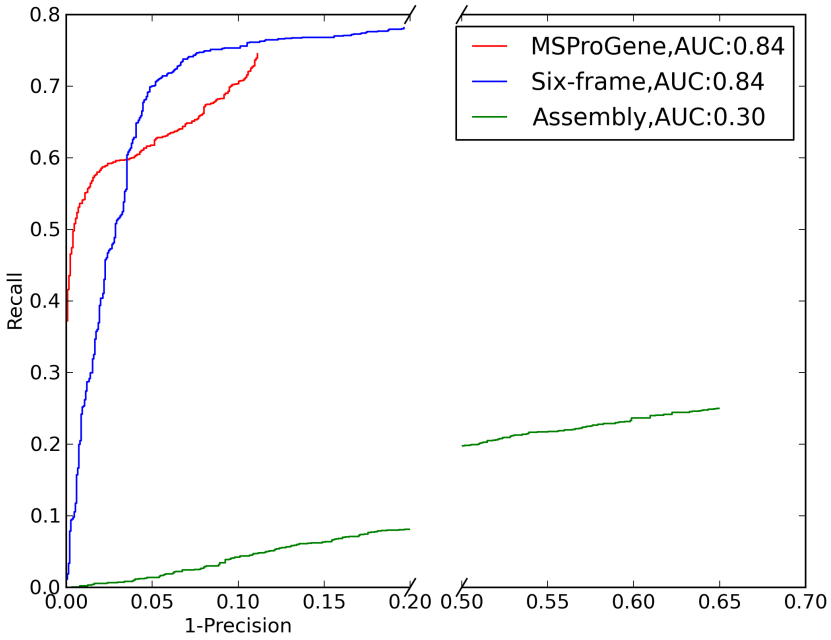

Figure 4.5.: Receiver operating curve illustrating recall and precision of MSProGene, Six-frame and Assembly for the *B. henselae* data set. MSProGene shows the highest precision of all three methods. Particularly for highly scored predictions it achieves better sensitivity at the same precision level.

As illustrated in Figure 4.6, MSProGene and the Six-frame approach coincide in 941 of the 1,488 annotations. In contrast, Assembly only shared 304 and 317 annotations with MSProGene and Six-frame, respectively. Taken together, the three methods identified 1,340 of the 1,488 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.
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.

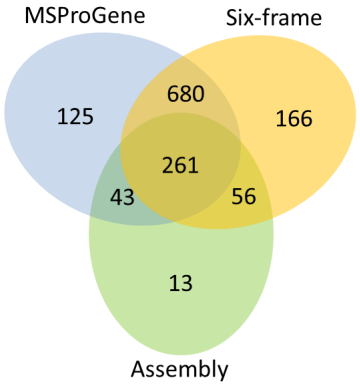

Figure 4.6.: Venn diagram illustrating the number of identified annotated proteins of the *B. henselae* data set for MSProGene, Six-frame, and Assembly. Together, 1,340 of the annotated proteins were identified, although no method shows a complete prediction by itself.

Comparison to reference-based methods:

To generate the mutated database 2,592 variants were called with samtools on the RNA-Seq read mapping and included in the reference protein sequences. Both Standard and Mutated method identified 1,274 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. With 1,109 identified proteins, MSProGene has a lower recall than both Standard and Mutated method. However, as shown in Figure 4.7, it identifies 84 proteins not detected by the standard searches.

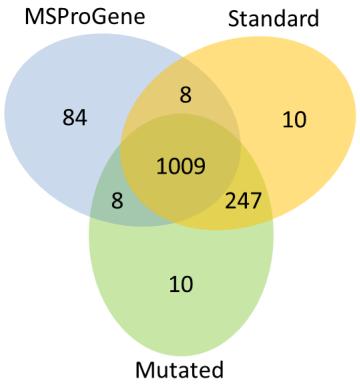

Figure 4.7.: Venn diagram illustrating the number of identified annotated proteins of the *B. henselae* data set for MSProGene, Standard, and Mutated. Together, 1,376 of the annotated proteins were identified, although no method shows a complete prediction by itself.

When compared to the annotated reference database, 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 1,180,672, 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, for instance present in strain BM1374165 (BLAST E value: 1e-178, identity: 99%). The second one shows high similarity to a hemin binding protein E, for instance present in strain BM1374163 (BLAST E value: 5e-145, identity: 100%). Thus, both genes are likely candidates for novel genes in the Houston-1 reference strain of the *B. henselae* taxonomy.

4.7.3. *L. sigmodontis* data

The results of the evaluation on the *L. sigmodontis* data set are shown in Table 4.3. Also for this data set 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:

	MSProGene	Six-frame
Database size	28,009	189,512
Predicted	2,146	**4,297**
Median spectra count	**8**	6
BLAST hits all	1,462	**1,804**
Median bit score all	**54.5**	25.8
BLAST hits above threshold	**779**	42
Median bit score	**89.7**	70.1

Table 4.3.: Evaluation for the *L. sigmodontis* data set, with best values for each category marked in bold. 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.

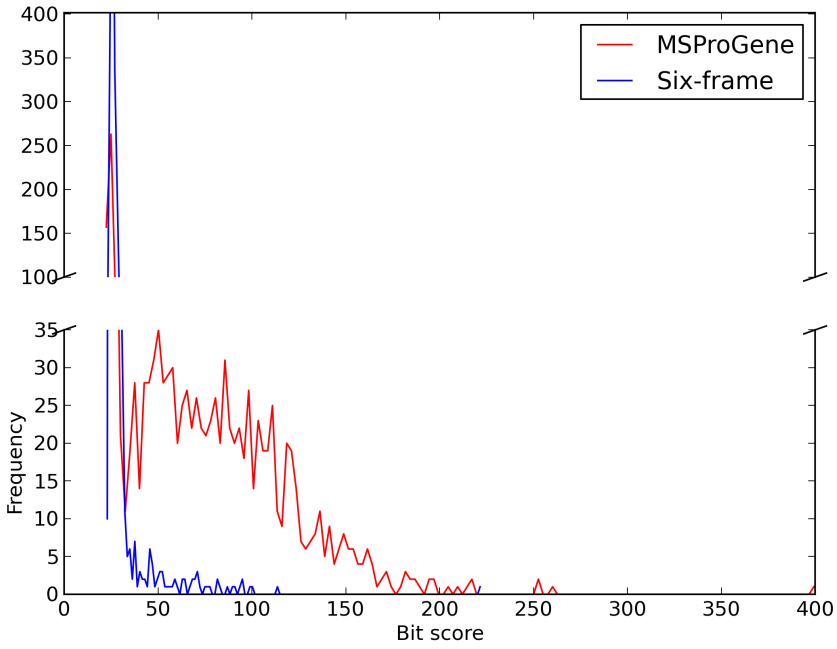

Figure 4.8.: The frequency of bit scores for MSProGene and the six-frame approach for the BLAST search of predicted sequences against a *B. malayi* reference. The confidence of MSProGene sequence alignments significantly exceeds the confidence of six-frame sequence alignments.

As illustrated in Figure 4.8, the confidence of BLAST hits is considerably improved for MSProGene sequences. Further, only 42% of the six-frame sequences receive a BLAST hit at all, while in contrast 68% of MSProGene predictions can be mapped. With 1,804 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 (779 hits for MSProGene vs. 42 hits 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.

4.7.4. System requirements

The computational performance of MSProGene is evaluated using the transcripts predicted by GIIRA (for *B. henselae*) or Cufflinks (for *L. sigmodontis*). We tested MSProGene on a linux system (with 256GB of available memory), using 20 threads. The main contributors to the run time are the two spectra searches performed by MSGF+: The search of 1.16 million spectra on the *B. henselae* data set required 35.7h. The search of 856,380 spectra on the *L. sigmodontis* data set required 40.8h. Overall, MSProGene used 30GB RAM and 36.5h to analyze the *B. henselae* data set, and 30GB RAM and 41.6h to analyze the *L. sigmodontis* data set.

4.8. Discussion

MSProGene facilitates automated and reference independent spectra search by constructing customized transcript databases. We demonstrated the accuracy of MSProGene identifications in several experiments on prokaryotic and eukaryotic organisms. Unlike methods focused on using RNA-Seq information for extension with splice sites (Woo et al., 2013) or SNPs (Krug et al., 2014), we exploit the full information of RNA-Seq experiments.
The tailored construction of RNA-Seq based databases leads to significantly reduced database size, as we show for the *B. henselae* and *L. sigmodontis* data sets. This size reduction has a positive effect on identification accuracy: The results of MSProGene are more precise than the results of a standard six-frame translation (refer to Section 4.7.2). We briefly evaluated the effect of combining RNA-Seq information with tandem mass spectrometry and show that both measures correlate

well and lead to significantly increased precision in protein identifications.
The direct comparison of MSProGene to other database-independent approaches shows that the six-frame translation has a high recall, but at the cost of precision. On the *B. henselae* data set MSProGene identifies slightly fewer proteins, but provides higher confidence in the resulting predictions. The approach based on *de novo* transcriptome assembly showed an overall low identification accuracy, also compared to the six-frame translation. 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.
The standard six-frame translation is sensitive, but lacks confidence and precision. MSProGene is specific, but currently it shows reduced overall recall in comparison to the six-frame approach. This is due to the dependency of MSProGene on the quality of predicted transcript sequences. Here, gene identifications exclusively based on RNA-Seq might not identify all possible transcripts and a more comprehensive RNA-Seq based prediction might be more sensitive. As shown in Section 4.7.2, *de novo* assembly can recover some of the missing transcripts; however, this approach overall has the least accuracy. Hence, in regard to precision, customized transcript databases as employed by MSProGene should be preferred.
We also compared MSProGene to reference-based approaches, either using reference databases or databases extended with sample-specific SNPs. We see that on the one hand MSProGene has a lower recall than the standard approaches, but on the other hand it identifies novel unannotated proteins, confirmed by BLAST search. Further, it also detects annotated proteins not identified by the standard methods. Interestingly, including sample-specific mutations observed in the RNA-Seq experiment did not improve the recall of the standard database search. 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.
When comparing MSProGene and the approach based on a mutated reference protein database (i.e. Mutated), 92 proteins are unique to MSProGene, and 257 proteins are unique to the other approach. 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 the constructed transcript sequences. Since RNA-Seq is challenging as the exclusive source for gene prediction, 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 published 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 they are supported by peptides that have scores below the FDR threshold in the Mutated approach. For instance, 51 of the missing 92 proteins of the approach searching an extended database can be identified with an FDR threshold of 5% (instead of the original 1%). This illustrates the precision of MSProGene peptide assignments since it identifies these proteins under a more conservative FDR.
As shown in the *B. henselae* experiment, reference-dependent approaches fail to detect novel genes and additionally 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 that go beyond the identification of annotated genes.
On the *L. sigmodontis* data set, the benefits of using RNA-Seq based transcript databases instead of six-frame translations are even more pronounced: Here, MSProGene identified an order of magnitude more sequences verified by BLAST search than the standard translation. In addition, it identified significantly fewer overall proteins than the six-frame translation, which appears disadvantageous at first glance, but is rather a strength of our method: The proportion of meaningful identifications is considerably higher for MSProGene than for the standard search.
Further, we also introduced MSProGene as a new method for shared peptide protein inference. We represent peptide spectrum matches in a network and resolve shared peptide connections using RNA-Seq evidence. This eliminates the need for protein grouping and allows a more specific protein identification. As shown in the simulated *E. coli* experiment, MSProGene accurately resolves shared peptides and detects incorrect PSMs, which further stresses the suitability of MSProGene for accurate proteogenomic analysis.

5. Summary and outlook

Proteogenomics is established as a key research area that integrates data from different high-throughput technologies to facilitate accurate annotation. Typically, studies are focused on the annotation and revision of genes and the identification of proteins. Despite varying research efforts and sophisticated analysis methods, proteogenomics remains a challenging field. The choice and construction of databases tailored to sample-specific tandem mass spectra search is one of the key questions that still requires further attention. In addition, the quality control of peptide identifications and the inference of proteins including shared peptides remain difficult problems.

This thesis describes three new computational methods that facilitate the integration of genomic, transcriptomic, and proteomic data to approach the current challenges in proteogenomic analysis. We introduce new methods for improved evidence-based gene model prediction that serve as the basis for customized spectra search databases beyond six-frame translations and independent from *a priori* annotations. We develop graphical models to approach the correct assignment of ambiguous reads and ambiguous peptides. Thereby, we perform an RNA-Seq evidence assisted protein inference including shared peptides and estimated the proportion of false identifications with a decoy-free FDR calculation.

In Chapter 2 we describe the new method GIIRA for RNA-Seq-based gene prediction to obtain sample-specific gene models as a basis for database construction. RNA-Seq is a particularly suitable source to assist gene model prediction as it reflects currently expressed genes and also captures complex structures, such as alternative splicing. GIIRA aims at explicitly including ambiguous read mappings for better identification sensitivity, rather than excluding these reads from the analysis as is performed by most RNA-Seq analysis pipelines. Particularly the detection of homologous gene regions or genes present in multiple copies can be impeded by excluding ambiguity. Thus, we integrate ambiguous mappings in a network that represents candidate gene sequences and their read support. Based on the reliability of candidate genes and mappings, reads can be reassigned to their most likely origin using a maximum-flow network optimization. In several experiments we show that including ambiguous reads indeed improves the prediction accuracy and leads to more sensitive results. Further, GIIRA performs favorably in comparison to other gene finders on prokaryotic as well as eukaryotic organisms. It demonstrates superior prediction accuracy for exons and introns on analyzed human and

S. cerevisiae data sets. Although all compared methods have different strengths and weaknesses and often show comparable performance on the eukaryotic data sets, GIIRA overall provides a sensitive prediction associated with high precision. Further, unlike other RNA-Seq-based methods, GIIRA particularly resolves structural genes contained in continuously expressed transcript sequences in prokaryotes. It uses an iterative alignment-scoring-based optimization that provides an accurate prediction also for prokaryotic organisms. Thus, GIIRA shows a superior performance in sample-specific gene identification in the *E. coli* experiments.
In Chapter 3 we describe the method IPred, which builds on the output of GIIRA and post-processes the results of gene finders to combine their strengths for improved prediction accuracy. *Ab initio* methods are not sample-specific, but in return sensitive in predicting genes following standard coding schemes and structures. In contrast, evidence-based prediction methods are sample-specific, but limited in their accuracy by noisy or incomplete evidence. However, particularly with regard to condition-specific experiments, a tailored gene identification beyond *ab initio* predictions is necessary. IPred explicitly combines predictions by *ab initio* and evidence-based strategies to benefit from their respective advantages while filtering false positive predictions and ensuring an experiment-specific prediction. The method analyzes the input predictions and compares the overlap support of evidence and *ab initio*-based identifications, thereby controlling variations between predicted sequences with a specific overlap threshold. Particularly, IPred respects RNA-Seq-specific challenges, such as varying coverage levels throughout genes, and balances variation with the introduced overlap threshold. In several experiments on prokaryotes and eukaryotes we demonstrate the superior accuracy of IPred combinations in comparison to single method predictions and to other combination approaches. We show that IPred detects and filters false positive identifications. Further, it adapts *ab initio* predictions based on the input evidence.
The last project, introduced in Chapter 4, uses the previously described methods to construct customized proteogenomic databases for spectra search. The method MSProGene goes beyond six-frame translations and reference databases and provides an RNA-Seq-based transcript database tailored to specific experiments. We show that MSProGene achieves significantly smaller database sizes than six-frame or *de novo* assembly strategies, thereby reducing peptide identification biases associated with large databases. On a *B. henselae* data set we show that this leads to an increased precision in peptide identification. Further, MSProGene approaches shared peptide protein inference by constructing a proteogenomic network based on the observed peptide spectrum matches. Shared peptides are resolved by optimizing the network with a maximum-flow approach based on the peptide hit quality and RNA-Seq information. We demonstrate that this optimization not only facilitates the assignment of shared peptides, but also identifies false peptide hits. On a *L. sigmodontis* data set, where a reference-based search is not possible due

to lack of comprehensive annotation, MSProGene identifies twenty times as many proteins verified by BLAST search as the standard six-frame translation. This illustrates the strength of MSProGene to accurately identify proteins and novel genes, also on organisms that lack annotation.
Together, the three described methods provide an accurate and sample-specific proteogenomic analysis. They overcome current limitations, such as database increase by six-frame translations, the challenging detection of novel genes, and the assignment of ambiguous RNA-Seq reads and peptide hits. Thereby, they contribute to more sophisticated and integrative proteogenomic analysis.

5.1. Outlook

The three methods described in this thesis facilitate automated and accurate proteogenomic analysis. However, naturally the methods can be improved or extended in various aspects, which we detail in the following and which can be categorized as computational improvements, conceptual extensions, and additional applications.
The first aspect is the computational improvement: Currently the described approaches are individual methods, which are applied independently from each other. Thus, although the availability of stand-alone software is desirable in general, in the context of performing experiment-specific proteogenomic analysis a workflow that automatically combines all three methods would even further increase the usability. To this regard, the GIIRA software is already integrated in MSProGene for default gene model prediction. However, a proteogenomic pipeline that allows the integration of other prediction methods using IPred without the need to call each program separately would be beneficial for users that are not experienced in the use of command-line software. In addition, this would lead to more flexibility in the choice of evidence used for database construction. The network framework of MSProGene is in general independent of the evidence information that assists the peptide assignment. Hence, it would be desirable to allow the automated integration of evidence beyond or in addition to RNA-Seq. Further, currently GIIRA may require gigabytes of memory and hours of run time, depending on the size of the input data set. This is mainly due to the ambiguous read reassignment: With millions of RNA-Seq reads and high proportions of ambiguity, the network constructed to resolve ambiguous connections can become very large and computationally expensive. Although the linear program optimization already provides a fast solution algorithm, further strategies to pre-process the network or construct the network in a less memory-demanding way would be desirable. For instance, a possible approach could perform a pre-processing of read mappings to construct multiple connected components, i.e. sub-networks that include groups of candi-

date genes that only share reads among another, but not with genes from other sub-networks. A sequential optimization of the connected components could then reduce the memory requirements.
Second, also conceptual improvements are possible for the three described methods: In our work we show that RNA-Seq is a valuable source of evidence to assist database construction. However, the RNA-Seq-based prediction of genes is challenging and methods are currently far from perfect. In particular, GIIRA sometimes shows difficulties in predicting complex alternative isoforms. One strategy to overcome this problem can be the integration of splice graph approaches for better intron-chain prediction. As performed in methods such as Cufflinks (Trapnell et al., 2012) or Scripture (Guttman et al., 2010), paths through a splice graph, which represents intron exon connections, help to define the correct series of exons. Another possibility is the usage of read coverage information to infer alternative isoforms based on their expression level. Exons can be associated to their respective isoform based on coverage similarities or dissimilarities to other exons, as for instance performed in (Drewe et al., 2013).
Also the post-processing of predictions with IPred can be further improved. Currently, due to non-standard outputs of gene finders, IPred exclusively reports transcript and exon features for each gene. However, to facilitate in-depth analysis with comprehensive gene models also the annotation of untranslated regions and coding sequences is of interest. When more standardized formats become available, IPred has to be adapted to provide more detailed annotations. Further, currently IPred only aims at combining prediction outputs, in contrast to performing a prediction itself given the information from other gene finders. The overlap-threshold-based comparison of positional similarities is a fast combination approach and showed superior performance to existing methods. However, the method could be extended, for instance by searching for new start codons in case of dissimilarities between evidence-based prediction results. Another example could be the combination of unsupported alternative isoforms. If each gene finder predicted a different isoform, it is likely that none of them is correct (what IPred currently assumes), but still the evidence points to at least one present isoform. Thus, a comparison and combination of different alternatives might benefit the accuracy of alternative isoform identification.
With better gene prediction strategies and post-processing also the database used for spectra search can be improved, which benefits the third method MSProGene. In our experiments we show that MSProGene is a very precise method, but that it currently lacks sensitivity compared to other proteogenomic analysis methods. Since MSProGene strongly depends on the quality of the genes underlying the database construction, improved gene models predicted by GIIRA and IPred can help to approach this problem. Furthermore, the use of other evidence in combination with RNA-Seq, for instance ESTs or protein alignments, should be considered

to further improve gene predictions and thus the constructed database. This will enhance the sensitivity of MSProGene identifications and facilitate a more comprehensive and yet precise proteogenomic identification. In addition, for better evaluation of identification accuracy it would be desirable to estimate the protein level FDR of MSProGene, instead of using an annotation-based FDR. This could be performed by taking decoy protein hits into account. Currently, the graphical model of MSProGene only resolves target database hits. However, in theory also an application to decoy identifications is possible. For instance, characteristics and scores of target sequences could be transferred to their corresponding reversed decoy sequence. Given these transferred qualities, in theory also ambiguous decoy hits can be reassigned. Then, the resulting supported decoy proteins can then be used for the calculation of a protein level FDR. However, a careful evaluation of the applicability of this approach is necessary since the RNA-Seq-based information might not be directly applicable to the artificial decoy sequences, which could bias the reassignment.

As a third aspect, not only improvements of the described methods are possible, but also their application to questions beyond their current use cases. For instance, the described proteogenomic workflow is focused on the qualitative identification of genes and proteins. Read and spectra coverages are provided and also used in subsequent analysis, but are not the focus of the software. However, an application to quantitative analyses, such as differential expression or protein quantification, is a worthwhile topic for future research. To do so, the correlation of read and spectra coverages needs to be further investigated. Then possible applications of a direct association between read coverage and spectra coverage are two-fold: (i) Information on expression levels of genes and proteins could be applied in the graphical models of GIIRA and MSProGene to infer the correct origins of ambiguous reads and shared peptides. (ii) The network approaches can be used not only to identify genes and proteins, but to analyze and compare expression levels.

An additional possible use case is the application to metaproteomics: In general it makes no difference for the introduced methods if one genome is analyzed or several ones (metagenomes can be treated as different chromosomes or contigs of a single organism). The reassignment of shared peptides and ambiguous reads could be extended to different genomes in a metagenome, for instance by integrating a new "metagenome"-layer in the proteogenomic network. Genes and proteins that are conserved in various organisms are a challenge for the functional annotation of metagenomic samples. Thus, an approach such as MSProGene, which resolves shared connections and is independent of existing reference databases, can provide a framework for a tailored and sample-specific metagenomic and metaproteomic analysis.

Another possible use case beyond gene and protein identification or quantification is the application of the described workflow to SNP calling. Currently, as a side

product, MSProGene outputs a VCF file with variants observed in the RNA-Seq mappings that are supported by spectra hits. However, there is significant potential in a combined spectra and transcriptome (or also genome) based SNP calling. Of course, this approach could only be applied to coding regions. But this limitation does not impede the potential and use cases of the approach: e.g., for questions related to disease-associated protein mutations particularly the changes in expressed regions are of interest. A careful investigation of the correlation of SNPs on the genome and single amino acid variations would be necessary to design a proteogenomic SNP caller that includes proteomic support in variation verification. Here, MSProGene can serve as a starting point since it provides information on RNA-Seq-based SNPs supported by spectra.
Further, MSProGene could be used for the analysis of diploid (or polyploid) SNPs and their effect on gene expression. Given a diploid SNP in the RNA-Seq mapping, the corresponding transcript used for spectra search could be duplicated. MSProGene can then simultaneously analyze both variants and measure the spectra support for each alternative. Since shared peptides can be distributed between both variants, the support for each alternative could be directly measured. This provides new possibilities for the analysis of allele-specific mutations with regard to their influence on gene expression.

A. Appendix

GIIRA evaluation on the simulated human data set

Method	Base	Exon	Intron	Intron-Chain	Transcript	Locus
Sensitivity						
GIIRA_w/_ambiguous	**97.2**	85.7	91.0	44.6	38.5	59.1
GIIRA_w/o_ambiguous	93.5	80.1	85.6	43.6	37.9	57.2
Cufflinks	93.0	71.6	86.7	**48.8**	0.6	56.2
AUGUSTUS	93.4	**88.6**	**91.9**	45.4	**39.3**	**59.7**
Specificity						
GIIRA_w/_ambiguous	98.0	**89.1**	96.7	43.3	34.9	43.9
GIIRA_w/o_ambiguous	**98.4**	88.4	**98.4**	44.3	34.1	39.8
Cufflinks	97.8	78.2	97.3	**51.7**	0.5	44
AUGUSTUS	82.3	81.4	85.3	49.1	**38.1**	**44.8**
F-measure - exact						
GIIRA_w/_ambiguous	**97.6**	**87.4**	**93.8**	43.9	36.6	50.4
GIIRA_w/o_ambiguous	95.9	84.0	91.6	43.9	35.9	46.9
Cufflinks	95.4	74.8	91.7	**50.2**	0.5	49.4
AUGUSTUS	87.5	84.8	88.5	47.2	**38.7**	**51.2**
Fuzzy Sensitivity						
GIIRA_w/_ambiguous	-	**89.8**	91.7	58.0	**44.9**	63.5
GIIRA_w/o_ambiguous	-	84.2	86.1	53.7	43.0	60.0
Cufflinks	-	85.2	87.2	63.2	36.0	60.3
AUGUSTUS	-	89.4	**92.3**	**70.2**	40.6	**74.3**
Fuzzy Specificity						
GIIRA_w/_ambiguous	-	**93.4**	97.4	56.3	**40.6**	47.1
GIIRA_w/o_ambiguous	-	92.9	**99.0**	54.5	38.7	41.7
Cufflinks	-	93.0	97.8	67.0	35.5	47.2
AUGUSTUS	-	82.1	85.7	**75.9**	39.4	**54.9**
F-measure - fuzzy						
GIIRA_w/_ambiguous	-	**91.6**	**94.5**	57.1	**42.6**	54.1
GIIRA_w/o_ambiguous	-	88.3	92.1	54.1	40.7	49.2
Cufflinks	-	88.9	92.2	65.0	35.7	53.0
AUGUSTUS	-	85.6	88.9	**72.9**	40.0	**63.1**

Table A.1.: Cuffcompare analysis for the simulated human data. The highlighted numbers indicate the best results for each category for sensitivity and specificity for GIIRA including ambiguous reads (GIIRA_w/_ambiguous), GIIRA excluding ambiguous reads (GIIRA_w/o_ambiguous), Cufflinks, and AUGUSTUS.

Sampled GIIRA evaluation on the simulated human data set

Method	Base	Exon	Intron	Intron-Chain	Transcript	Locus
Sensitivity						
GIIRA_w/_ambiguous	87.2	76.3	80.7	42.5	36.6	55.7
GIIRA_w/o_ambiguous	84.6	73.6	77.9	41.9	36.3	54.1
Cufflinks	60.5	47.2	58.4	38.4	0.3	44.9
AUGUSTUS	**91.0**	**88.0**	**91.6**	**45.4**	**37.5**	**56.9**
Specificity						
GIIRA_w/_ambiguous	**98.7**	91.5	97.5	48.9	41.2	53.1
GIIRA_w/o_ambiguous	**98.7**	**91.7**	**98.3**	50.8	42.6	52.1
Cufflinks	98.4	75.3	98.0	52.4	0.4	43.2
AUGUSTUS	89.8	85.8	88.6	**54.3**	**44.6**	**54.5**
F-measure - exact						
GIIRA_w/_ambiguous	**92.6**	83.2	88.3	45.5	38.8	54.4
GIIRA_w/o_ambiguous	91.1	81.7	86.9	45.9	39.2	53.1
Cufflinks	74.9	58.0	73.2	44.3	0.3	44.0
AUGUSTUS	90.4	**86.9**	**90.1**	**49.5**	**40.7**	**55.7**
Fuzzy Sensitivity						
GIIRA_w/_ambiguous	-	79.7	81.4	55.5	**42.9**	59.5
GIIRA_w/o_ambiguous	-	76.7	78.4	51.9	41.2	56.9
Cufflinks	-	58.3	58.6	49.1	27.9	48.3
AUGUSTUS	-	**88.8**	**92.0**	**70.2**	38.8	**71.5**
Fuzzy Specificity						
GIIRA_w/_ambiguous	-	**95.5**	98.3	63.9	48.2	56.8
GIIRA_w/o_ambiguous	-	**95.5**	**98.9**	62.9	**48.4**	54.7
Cufflinks	-	93	98.3	67.0	34.8	46.5
AUGUSTUS	-	86.6	89.0	**84.0**	46.2	**67.4**
F-measure - fuzzy						
GIIRA_w/_ambiguous	-	86.9	89.1	59.4	**45.4**	58.1
GIIRA_w/o_ambiguous	-	85.1	87.5	56.9	44.5	55.8
Cufflinks	-	71.7	73.4	56.7	31.0	47.4
AUGUSTUS	-	**87.7**	**90.5**	**76.5**	42.2	**69.4**

Table A.2.: Cuffcompare analysis for the simulated human data on a sample of 600 predictions for each compared method, evaluated against 992 reference transcripts. The highlighted numbers indicate the best results for each category for sensitivity and specificity for GIIRA including ambiguous reads (GIIRA_w/_ambiguous), GIIRA excluding ambiguous reads (GIIRA_w/o_ambiguous), Cufflinks, and AUGUSTUS.

GIIRA evaluation on the simulated yeast data set

Methods	Base	Exon	Intron	Intron-Chain	Transcript	Locus
Sensitivity						
GIIRA_w/_ambiguous	**99.2**	**85.9**	66.7	65.6	**86.8**	**87.4**
GIIRA_w/o_ambiguous	97.4	85.3	66.7	65.6	86.1	86.7
Cufflinks	93.3	1.8	**69.7**	**68.8**	0.5	3.5
Specificity						
GIIRA_w/_ambiguous	99.1	**85.5**	**88.0**	**84.0**	**85.4**	**86.0**
GIIRA_w/o_ambiguous	**99.3**	84.5	**88.0**	**84.0**	84.4	85.2
Cufflinks	98.0	1.9	82.1	81.5	0.6	3.6
F-measure - exact						
GIIRA_w/_ambiguous	**99.2**	**85.7**	**75.9**	73.7	**86.1**	**86.7**
GIIRA_w/o_ambiguous	98.3	84.9	**75.9**	73.7	85.2	85.9
Cufflinks	95.6	1.9	75.4	**74.6**	0.6	3.6
Fuzzy Sensitivity						
GIIRA_w/_ambiguous	-	**86.8**	66.7	65.6	**87.6**	**87.8**
GIIRA_w/o_ambiguous	-	86.3	66.7	65.6	87.0	87.3
Cufflinks	-	73.1	**69.7**	**68.8**	72.5	73.4
Fuzzy Specificity						
GIIRA_w/_ambiguous	-	**86.4**	**88.0**	**84.0**	**86.2**	**86.4**
GIIRA_w/o_ambiguous	-	85.5	**88.0**	**84.0**	85.3	85.7
Cufflinks	-	76.1	82.1	81.5	75.1	76.0
F-measure - fuzzy						
GIIRA_w/_ambiguous	-	**86.6**	**75.9**	73.7	**86.9**	**87.1**
GIIRA_w/o_ambiguous	-	85.9	**75.9**	73.7	86.1	86.5
Cufflinks	-	74.6	75.4	**74.6**	73.8	74.7

Table A.3.: Cuffcompare analysis for the simulated yeast data set. The highlighted numbers indicate the best results for each category for GIIRA including ambiguous reads (GIIRA_w/_ambiguous), GIIRA excluding ambiguous reads (GIIRA_w/o_ambiguous), and Cufflinks.

(1) *E. coli* real - complete reference

	exact measure									fuzzy measure					
	Base			Exon			Locus			Exon			Loci		
	Sn	Sp	F	Sn	Sp	F	Sn	Sp	F	Sn	Sp	F	Sn	Sp	F
GIIRA	61.4	93.3	74.1	42.3	42.3	42.3	44.7	50.2	47.3	42.9	42.9	42.9	45.3	50.8	47.9
Cufflinks	40.7	72.2	52.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.1	0.1
GeneMark	56.1	47.9	51.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLIMMER3	**96.7**	**94.6**	**95.6**	**71.9**	**67.8**	**69.8**	**72.3**	**69.4**	**70.8**	**73.1**	**68.9**	**70.9**	**73.5**	**70.5**	**72.0**

(2) *E. coli* real - reference subset

	exact measure									fuzzy measure					
	Base			Exon			Locus			Exon			Loci		
	Sn	Sp	F	Sn	Sp	F	Sn	Sp	F	Sn	Sp	F	Sn	Sp	F
GIIRA	95.6	**65.3**	**77.6**	66.4	32.1	43.3	67.6	**38.4**	**49.0**	67.4	32.6	43.9	68.6	**38.9**	**49.6**
Cufflinks	70.0	55.8	62.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.1	0.1
GeneMark	60.3	23.2	33.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLIMMER3	**97.4**	42.8	59.5	**71.2**	**32.4**	**44.5**	**71.9**	35.0	47.1	**72.4**	**33.0**	**45.3**	**73.0**	35.5	47.8

(3) *E. coli* real - alternative

	exact measure									fuzzy measure					
	Base			Exon			Locus			Exon			Loci		
	Sn	Sp	F	Sn	Sp	F	Sn	Sp	F	Sn	Sp	F	Sn	Sp	F
GIIRA	17.2	94.0	29.1	**12.6**	50.0	**20.1**	**13.2**	57.2	**21.5**	**12.9**	51.1	**20.6**	**13.5**	58.5	**21.9**
Cufflinks	15.1	65.1	27.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.2	0.0
GeneMark	**48.5**	46.9	**47.7**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GLIMMER3	15.3	**98.3**	26.5	10.0	**83.0**	17.8	11.0	**83.1**	19.4	10.0	**83.2**	17.9	11.0	**83.3**	19.4

Table A.4.: Cuffcompare analysis for the real *E. coli* data set compared against the complete annotated reference of 4,146 genes (1) and a subset of reference genes (2). The third table shows the comparison against the complete reference, based on a selected sample of 500 predictions for each method. The highlighted numbers indicate the best results for each category. Abbreviations: Sn = Sensitivity, Sp = Specificity, F = F-measure.

GIIRA evaluation on the real yeast data set

(1) Accuracy

Methods	Base	Exon	Intron	Intron-Chain	Transcript	Locus
Sensitivity						
GIIRA_w/_ambiguous	**82**	**0.4**	**56.9**	**56.3**	**0.2**	**2.9**
GIIRA_w/o_ambiguous	77.8	0.3	56.6	56	**0.2**	2.8
Cufflinks	75.1	0.0	55.2	54.9	0.0	2.6
Specificity						
GIIRA_w/_ambiguous	80.1	**0.2**	61.3	59.5	**0.1**	1.5
GIIRA_w/o_ambiguous	**80.5**	**0.2**	**69.5**	**67.4**	**0.1**	1.5
Cufflinks	77.3	0.0	61.8	61.0	0.0	**2.4**
F-measure - exact						
GIIRA_w/_ambiguous	**81.0**	**0.3**	59.0	57.9	**0.1**	2.0
GIIRA_w/o_ambiguous	79.1	0.2	**62.4**	**61.2**	**0.1**	2.0
Cufflinks	76.2	0.0	58.3	57.8	0.0	**2.5**
Fuzzy Sensitivity						
GIIRA_w/_ambiguous	-	**0.4**	**57.6**	**57.4**	**0.2**	**3.0**
GIIRA_w/o_ambiguous	-	**0.4**	57.2	57.0	**0.2**	2.9
Cufflinks	-	0.0	56.2	56.7	0.0	2.7
Fuzzy Specificity						
GIIRA_w/_ambiguous	-	**0.2**	62.1	60.7	**0.1**	1.5
GIIRA_w/o_ambiguous	-	**0.2**	**70.3**	**68.7**	**0.1**	1.6
Cufflinks	-	0.0	62.9	63.1	0.0	**2.5**
F-measure - fuzzy						
GIIRA_w/_ambiguous	-	**0.3**	59.8	59.0	**0.1**	2.0
GIIRA_w/o_ambiguous	-	**0.3**	**63.1**	**62.3**	**0.1**	**2.1**
Cufflinks	-	0.0	59.6	0.0	0.0	0.0

(2) Missed and novel exons (in percent)

Methods	Missed exons	Novel exons	Missed intron	Novel intron	Missed loci	Novel loci
GIIRA_w/_ambiguous	**10.6**	11.5	**41.7**	35.7	**9.7**	11.3
GIIRA_w/o_ambiguous	12.6	10.5	42.1	**27.1**	11.5	10.4
Cufflinks	20	**5.5**	43.1	35.5	18.7	**4.6**

Table A.5.: Table (1) shows the Cuffcompare analysis for the real yeast data set (evaluated on the complete genome). Table (2) shows the proportions of completely missed and completely novel predictions. GIIRA was applied in two configurations: including ambiguous reads (GIIRA_w/_ambiguous), and excluding ambiguous reads (GIIRA_w/o_ambiguous). The best values for each category are marked in bold.

Sampled GIIRA evaluation on the real yeast data set

Methods	Base	Exon	Intron	Intron-Chain	Transcript	Locus
Sensitivity						
GIIRA_w/_ambiguous	**55**	**0.2**	**54.8**	**54.2**	**0.1**	**2.7**
GIIRA_w/o_ambiguous	53.9	**0.2**	54.1	53.4	**0.1**	2.6
Cufflinks	54.5	0.0	53.8	53.4	0.0	2.5
Specificity						
GIIRA_w/_ambiguous	78.7	**0.3**	64.1	62.2	**0.1**	**3.7**
GIIRA_w/o_ambiguous	**78.8**	**0.3**	**69.8**	**67.6**	**0.1**	**3.7**
Cufflinks	74.5	0.1	65.3	64.1	0.0	3.5
F-measure - exact						
GIIRA_w/_ambiguous	**64.8**	**0.2**	59.1	57.9	**0.1**	**3.1**
GIIRA_w/o_ambiguous	64.0	**0.2**	**61.0**	**59.7**	**0.1**	**3.1**
Cufflinks	63.0	0.0	59.0	58.3	0.0	2.9
Fuzzy Sensitivity						
GIIRA_w/_ambiguous	-	**0.3**	**55.5**	**55.2**	**0.1**	**2.7**
GIIRA_w/o_ambiguous	-	**0.3**	54.8	54.5	**0.1**	**2.7**
Cufflinks	-	0.0	54.8	**55.2**	0.0	2.6
Fuzzy Specificity						
GIIRA_w/_ambiguous	-	**0.4**	64.9	63.5	**0.2**	**3.8**
GIIRA_w/o_ambiguous	-	**0.4**	**70.7**	**68.9**	**0.2**	**3.8**
Cufflinks	-	0.1	66.5	66.2	0.0	3.6
F-measure - fuzzy						
GIIRA_w/_ambiguous	-	**0.3**	59.8	59.1	**0.1**	**3.2**
GIIRA_w/o_ambiguous	-	**0.3**	**61.7**	**60.9**	**0.1**	**3.2**
Cufflinks	-	0.0	60.1	60.2	0.0	3.0

Table A.6.: Cuffcompare analysis for the real yeast data set on a sample of 4,200 predictions for each compared method, evaluated against 5,905 reference transcripts. The highlighted numbers indicate the best results for each criterion for sensitivity and specificity for GIIRA including ambiguous reads (GIIRA_w/_ambiguous), GIIRA excluding ambiguous reads (GIIRA_w/o_ambiguous), and Cufflinks.

Bibliography

J.-M. Ahn, M.-S. Kim, Y.-I. Kim, S.-K. Jeong, H.-J. Lee, S. H. Lee, Y.-K. Paik, A. Pandey, and J.-Y. Cho. Proteogenomic analysis of human chromosome 9-encoded genes from human samples and lung cancer tissues. *Journal of proteome research*, 13(1):137–146, 2013.

J. E. Allen and S. L. Salzberg. JIGSAW: integration of multiple sources of evidence for gene prediction. *Bioinformatics*, 21(18): 3596–3603, 2005.

J. E. Allen, M. Pertea, and S. L. Salzberg. Computational gene prediction using multiple sources of evidence. *Genome research*, 14 (1):142–148, 2004.

S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402, 1997.

C. Ansong, S. O. Purvine, J. N. Adkins, M. S. Lipton, and R. D. Smith. Proteogenomics: needs and roles to be filled by proteomics in genome annotation. *Briefings in functional genomics&proteomics*, 7(1):50–62, 2008.

S. D. Armstrong, S. A. Babayan, N. Lhermitte-Vallarino, N. Gray, D. Xia, C. Martin, S. Kumar, D. W. Taylor, M. L. Blaxter, J. M. Wastling, et al. Comparative analysis of the secretome from a model filarial nematode (Litomosoides sigmodontis) reveals maximal diversity in gravid female parasites. *Molecular & cellular proteomics*, 13(10):2527–2544, 2014.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series b (methodological)*, 57(1):289–300, 1995.

J. Besemer, A. Lomsadze, and M. Borodovsky. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic acids research*, 29(12):2607–2618, 2001.

C. Bielow, S. Aiche, S. Andreotti, and K. Reinert. MSSimulator: Simulation of mass spectrometry data. *Journal of proteome research*, 10(7):2922–2929, 2011.

P. Blakeley, I. M. Overton, and S. J. Hubbard. Addressing statistical biases in nucleotide-derived protein databases for proteogenomic search strategies. *Journal of proteome research*, 11(11):5221–5234, 2012.

T. Bonfert, G. Csaba, R. Zimmer, and C. Friedel. A context-based approach to identify the most likely mapping for RNA-seq experiments. *BMC bioinformatics*, 13(Suppl 6):S9, 2012.

E. Bonzon-Kulichenko, F. Garcia-Marques, M. Trevisan-Herraz, and J. Vázquez. Revisiting peptide identification by high-accuracy mass spectrometry: problems associated to the use of narrow mass precursor windows. *Journal of proteome research*, 14(2):700–710, 2014.

R. A. Bradshaw, A. L. Burlingame, S. Carr, and R. Aebersold. Reporting protein identification data the next generation of guidelines. *Molecular & cellular proteomics*, 5(5):787–788, 2006.

R. M. Branca, L. M. Orre, H. J. Johansson, V. Granholm, M. Huss, Å. Pérez-Bercoff, J. Forshed, L. Käll, and J. Lehtiö. HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics. *Nature methods*, 11(1):59–62, 2014.

M. R. Brent. How does eukaryotic gene prediction work? *Nature biotechnology*, 25(8):883, 2007.

M. Burset and R. Guigó. Evaluation of gene structure prediction programs. *Genomics*, 34(3):353–367, 1996.

N. Castellana and V. Bafna. Proteogenomics to discover the full coding content of genomes: a computational perspective. *Journal of proteomics*, 73(11):2124–2135, 2010.

N. E. Castellana, S. H. Payne, Z. Shen, M. Stanke, V. Bafna, and S. P. Briggs. Discovery and revision of arabidopsis genes by proteogenomics. *Proceedings of the national academy of sciences*, 105(52):21034–21038, 2008.

J. S. Choudhary, W. P. Blackstock, D. M. Creasy, and J. S. Cottrell. Interrogating the human genome using uninterpreted mass spectrometry data. *Proteomics*, 1(5):651–667, 2001.

D. Chung, P. F. Kuan, B. Li, R. Sanalkumar, K. Liang, E. H. Bresnick, C. Dewey, and S. Keles. Discovering Transcription Factor Binding Sites in Highly Repetitive Regions of Genomes with Multi-Read Analysis of ChIP-Seq Data. *PLoS computational biology*, 7(7):e1002111, 07 2011.

M. Claassen. Inference and validation of protein identifications. *Molecular & cellular proteomics*, 11(11):1097–1104, 2012.

J.-M. Claverie. Computational methods for the identification of genes in vertebrate genomic sequences. *Human molecular genetics*, 6(10):1735–1744, 1997.

E. P. Consortium. The encode (encyclopedia of dna elements) project. *Science*, 306(5696):636–640, 2004.

B. Cooper. The problem with peptide presumption and the downfall of target–decoy false discovery rates. *Analytical chemistry*, 84 (22):9663–9667, 2012.

CPLEX. International Business Machines Corporation. v12.4: User's manual for CPLEX. *IBM ILOG CPLEX*, 2011. URL `http://www-01.ibm.com/software/integration/optimization/cplex-optimizer/`.

R. Craig and R. C. Beavis. Tandem: matching proteins with tandem mass spectra. *Bioinformatics*, 20(9):1466–1467, 2004.

R. Craig, J. Cortens, D. Fenyo, and R. C. Beavis. Using annotated peptide mass spectrum libraries for protein identification. *Journal of proteome research*, 5(8):1843–1849, 2006.

F. Crick et al. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.

V. Dancik, T. A. Addona, K. R. Clauser, J. E. Vath, and P. A. Pevzner. De novo peptide sequencing via tandem mass spectrometry. *Journal of computational biology*, 6(3-4):327–342, 1999.

A. L. Delcher, K. A. Bratke, E. C. Powers, and S. L. Salzberg. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics*, 23(6):673–679, 2007.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series b (methodological)*, 39(1):1–38, 1977.

F. Denoeud, J.-M. Aury, C. D. Silva, B. Noel, O. Rogier, M. Delledonne, M. Morgante, G. Valle, P. Wincker, C. Scarpelli, O. Jaillon, and F. Artiguenave. Annotating genomes with massive-scale RNA sequencing. *Genome biology*, 9(12):R175, 2008.

E. W. Deutsch, L. Mendoza, D. Shteynberg, T. Farrah, H. Lam, N. Tasman, Z. Sun, E. Nilsson, B. Pratt, B. Prazen, et al. A guided tour of the Trans-Proteomic Pipeline. *Proteomics*, 10(6):1150–1159, 2010.

P. Drewe, O. Stegle, L. Hartmann, A. Kahles, R. Bohnert, A. Wachter, K. Borgwardt, and G. Rätsch. Accurate detection of differential rna processing. *Nucleic acids research*, 41(10):5189–5198, 2013.

T. H. A. Ederveen, L. Overmars, and S. A. F. T. van Hijum. Reduce manual curation by combining gene predictions from multiple annotation engines, a case study of start codon prediction. *PLoS one*, 8(5):e63523, 05 2013.

C. G. Elsik, A. J. Mackey, J. T. Reese, N. V. Milshina, D. S. Roos, and G. M. Weinstock. Creating a honey bee consensus gene set. *Genome biology*, 8(1):R13, 2007.

J. K. Eng, A. L. McCormack, and J. R. Yates. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the american society for mass spectrometry*, 5(11):976–989, 1994.

V. C. Evans, G. Barker, K. J. Heesom, J. Fan, C. Bessant, and D. A. Matthews. De novo derivation of proteomes from transcriptomes for transcript and protein identification. *Nature methods*, 9(12):1207–1211, 2012.

S. Fanayan, J. T. Smith, L. Y. Lee, F. Yan, M. Snyder, W. S. Hancock, and E. Nice. Proteogenomic analysis of human colon carcinoma cell lines lim1215, lim1899, and lim2405. *Journal of proteome research*, 12(4):1732–1742, 2013.

N. Fawal, Q. Li, C. Mathé, and C. Dunand. Automatic multigenic family annotation: risks and solutions. *Trends in genetics*, 30(8): 323–325, 2014.

D. Fermin, B. B. Allen, T. W. Blackwell, R. Menon, M. Adamski, Y. Xu, P. Ulintz, G. S. Omenn, et al. Novel gene and gene model detection using a whole genome open reading frame analysis in proteomics. *Genome biology*, 7(4):R35, 2006.

L. R. Ford and D. R. Fulkerson. Maximal flow through a network. *Canadian journal of mathematics*, 8(3):399–404, 1956.

A. Frank and P. Pevzner. Pepnovo: de novo peptide sequencing via probabilistic network modeling. *Analytical chemistry*, 77(4): 964–973, 2005.

B. E. Frewen, G. E. Merrihew, C. C. Wu, W. S. Noble, and M. J. MacCoss. Analysis of peptide ms/ms spectra from large-scale proteomics experiments using spectrum libraries. *Analytical chemistry*, 78(16):5678–5684, 2006.

M. Garber, M. G. Grabherr, M. Guttman, and C. Trapnell. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature methods*, 8:469–477, 06 2011.

S. Gerster, E. Qeli, C. H. Ahrens, and P. Bühlmann. Protein and gene model inference based on statistical modeling in k-partite graphs. *Proceedings of the national academy of sciences*, 107(27):12101–12106, 2010.

S. H. Giese, F. Zickmann, and B. Y. Renard. Specificity control for read alignments using an artificial reference genome-guided false discovery rate. *Bioinformatics*, 30(1):9–16, 2014.

GLPK. GNU Linear Programming Kit, v4.47. *GLPK*, 2006. URL `http://www.gnu.org/software/glpk/glpk.html`.

S. J. Goodswen, P. J. Kennedy, and J. T. Ellis. Evaluating high-throughput ab initio gene finders to discover proteins encoded in eukaryotic pathogen genomes missed by laboratory techniques. *PLoS one*, 7(11), 2012.

M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology*, 29(7):644–652, 2011.

R. Guigó, P. Flicek, J. Abril, A. Reymond, J. Lagarde, F. Denoeud, S. Antonarakis, M. Ashburner, V. Bajic, E. Birney, R. Castelo, E. Eyras, C. Ucla, T. Gingeras, J. Harrow, T. Hubbard, S. Lewis, and M. Reese. EGASP: the human ENCODE genome annotation assessment project. *Genome biology*, 7(Suppl 1):S2, 2006.

N. Gupta, S. Tanner, N. Jaitly, J. N. Adkins, M. Lipton, R. Edwards, M. Romine, A. Osterman, V. Bafna, R. D. Smith, et al. Whole proteome analysis of post-translational modifications: applications of mass-spectrometry for proteogenomic annotation. *Genome research*, 17(9):1362–1377, 2007.

N. Gupta, J. Benhamida, V. Bhargava, D. Goodman, E. Kain, I. Kerman, N. Nguyen, N. Ollikainen, J. Rodriguez, J. Wang, et al. Comparative proteogenomics: combining mass spectrometry and comparative genomics to analyze multiple genomes. *Genome research*, 18(7):1133–1142, 2008.

M. Guttman, M. Garber, J. Z. Levin, J. Donaghey, J. Robinson, X. Adiconis, L. Fan, M. J. Koziol, A. Gnirke, C. Nusbaum, et al. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincrnas. *Nature biotechnology*, 28(5):503–510, 2010.

B. J. Haas, S. L. Salzberg, W. Zhu, M. Pertea, J. E. Allen, J. Orvis, O. White, C. R. Buell, and J. R. Wortman. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome biology*, 9(1):R7, 2008.

C. Holt and M. Yandell. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC bioinformatics*, 12(1):491, 2011.

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

N. Homer, B. Merriman, and S. F. Nelson. BFAST: an alignment tool for large scale genome resequencing. *PloS one*, 4(11):e7767, 2009.

T. Huang, J. Wang, W. Yu, and Z. He. Protein inference: a review. *Briefings in bioinformatics*, 13(5):586–614, 2012.

N. N. Ivanova, P. Schwientek, H. J. Tripp, C. Rinke, A. Pati, M. Huntemann, A. Visel, T. Woyke, N. C. Kyrpides, and E. M. Rubin. Stop codon reassignments in the wild. *Science*, 344(6186):909–913, 2014.

J. D. Jaffe, H. C. Berg, and G. M. Church. Proteogenomic mapping as a complementary method to perform genome annotation. *Proteomics*, 4(1):59–77, 2004.

P. D. Jagtap, J. E. Johnson, G. Onsongo, F. W. Sadler, K. Murray, Y. Wang, G. M. Shenykman, S. Bandhakavi, L. M. Smith, and T. J. Griffin. Flexible and accessible workflows for improved proteogenomic analysis using the galaxy framework. *Journal of proteome research*, 13(12):5898–5908, 2014.

K. Jeong, S. Kim, and N. Bandeira. False discovery rates in spectral identification. *BMC bioinformatics*, 13(Suppl 16):S2, 2012.

L. Käll, J. D. Canterbury, J. Weston, W. S. Noble, and M. J. MacCoss. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature methods*, 4(11):923–925, 2007.

D. E. Kalume, S. Peri, R. Reddy, J. Zhong, M. Okulate, N. Kumar, and A. Pandey. Genome annotation of anopheles gambiae using mass spectrometry-derived data. *BMC genomics*, 6(1):128, 2005.

M. Kearse, R. Moir, A. Wilson, S. Stones-Havas, M. Cheung, S. Sturrock, S. Buxton, A. Cooper, S. Markowitz, C. Duran, T. Thierer, B. Ashton, P. Meintjes, and A. Drummond. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28(12):1647–1649, 2012.

D. S. Kelkar, D. Kumar, P. Kumar, L. Balakrishnan, B. Muthusamy, A. K. Yadav, P. Shrivastava, A. Marimuthu, S. Anand, H. Sundaram, et al. Proteogenomic analysis of mycobacterium tuberculosis by high resolution mass spectrometry. *Molecular & cellular proteomics*, 10(12):M111–011627, 2011.

D. S. Kelkar, E. Provost, R. Chaerkady, B. Muthusamy, S. S. Manda, T. Subbannayya, L. D. N. Selvan, C.-H. Wang, K. K. Datta, S. Woo, et al. Annotation of the zebrafish genome through an integrated transcriptomic and proteomic analysis. *Molecular & cellular proteomics*, 13(11):3184–3198, 2014.

A. Keller, A. I. Nesvizhskii, E. Kolker, and R. Aebersold. Empirical statistical model to estimate the accuracy of peptide identifications made by ms/ms and database search. *Analytical chemistry*, 74(20):5383–5392, 2002.

D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, and S. Salzberg. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology*, 14(4):R36, 2013.

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

I. Korf. Gene finding in novel genomes. *BMC bioinformatics*, 5(1):59, 2004.

K. Krug, A. Carpy, G. Behrends, K. Matic, N. C. Soares, and B. Macek. Deep coverage of the escherichia coli proteome enables the assessment of false discovery rates in simple proteogenomic experiments. *Molecular & cellular proteomics*, 12(11):3420–3430, 2013.

K. Krug, S. Popic, A. Carpy, C. Taumer, and B. Macek. Construction and assessment of individualized proteogenomic databases for large-scale analysis of nonsynonymous single nucleotide variants. *Proteomics*, 14(23-24):2699–2708, 2014.

D. Kumar, A. K. Yadav, P. K. Kadimi, S. H. Nagaraj, S. M. Grimmond, and D. Dash. Proteogenomic analysis of bradyrhizobium japonicum usda110 using genosuite, an automated multi-algorithmic pipeline. *Molecular & cellular proteomics*, 12(11):3388–3397, 2013.

B. Küster, P. Mortensen, J. S. Andersen, and M. Mann. Mass spectrometry allows direct identification of proteins in large genomes. *Proteomics*, 1(5):641–650, 2001.

H. Lam, E. W. Deutsch, J. S. Eddes, J. K. Eng, N. King, S. E. Stein, and R. Aebersold. Development and validation of a spectral library searching method for peptide identification from ms/ms. *Proteomics*, 7(5):655–667, 2007.

H. Lam, E. W. Deutsch, J. S. Eddes, J. K. Eng, S. E. Stein, and R. Aebersold. Building consensus spectral libraries for peptide identification in proteomics. *Nature methods*, 5(10):873–875, 2008.

B. Langmead and S. L. Salzberg. Fast gapped-read alignment with bowtie 2. *Nature methods*, 9(4):357–359, 2012.

H. Li and R. Durbin. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and . G. P. D. P. Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009a.

J. Li, L. J. Zimmerman, B.-H. Park, D. L. Tabb, D. C. Liebler, and B. Zhang. Network-assisted protein identification and data interpretation in shotgun proteomics. *Molecular systems biology*, 5(1), 2009b.

J. Li, D. T. Duncan, and B. Zhang. Canprovar: a human cancer proteome variation database. *Human mutation*, 31(3):219–228, 2010.

Y. F. Li and P. Radivojac. Computational approaches to protein inference in shotgun proteomics. *BMC bioinformatics*, 13(Suppl 16): S4, 2012.

A. J. Link, L. G. Hays, E. B. Carmack, and J. R. Yates. Identifying the major proteome components of haemophilus influenzae type-strain nctc 8143. *Electrophoresis*, 18(8):1314–1334, 1997.

Q. Liu, A. J. Mackey, D. S. Roos, and F. C. N. Pereira. Evigan: a hidden variable model for integrating gene evidence for eukaryotic gene prediction. *Bioinformatics*, 24(5):597–605, 2008.

A. Lomsadze, V. Ter-Hovhannisyan, Y. O. Chernoff, and M. Borodovsky. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic acids research*, 33(20):6494–6506, 2005.

A. V. Lukashin and M. Borodovsky. GeneMark.hmm: New solutions for gene finding. *Nucleic acids research*, 26(4):1107–1115, 1998.

B. Ma, K. Zhang, C. Hendrie, C. Liang, M. Li, A. Doherty-Kirby, and G. Lajoie. Peaks: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid communications in mass spectrometry*, 17(20):2337–2342, 2003.

W. H. Majoros, M. Pertea, and S. L. Salzberg. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics*, 20(16):2878–2879, 2004.

M. Mann. A shortcut to interesting human genes: peptide sequence tags, expressed-sequence tags and computers. *Trends in biochemical sciences*, 21(12):494–495, 1996.

J. Martin, W. Zhu, K. Passalacqua, N. Bergman, and M. Borodovsky. Bacillus anthracis genome organization in light of whole transcriptome sequencing. *BMC bioinformatics*, 11(Suppl 3):S10, 2010.

C. Mathé, M.-F. Sagot, T. Schiex, and P. Rouzé. Current methods of gene prediction, their strengths and weaknesses. *Nucleic acids research*, 30(19):4103–4117, 2002.

D. Meierhofer, C. Weidner, and S. Sauer. Integrative analysis of transcriptomics, proteomics, and metabolomics data of white adipose and liver tissue of high-fat diet and rosiglitazone-treated insulin-resistant mice identified pathway alterations and molecular hubs. *Journal of proteome research*, 13(12):5592–5602, 2014.

M. L. Metzker. Sequencing technologies—the next generation. *Nature reviews genetics*, 11(1):31–46, 2009.

A. M. Mezlini, E. J. Smith, M. Fiume, O. Buske, G. L. Savich, S. Shah, S. Aparicio, D. Y. Chiang, A. Goldenberg, and M. Brudno. iReckon: Simultaneous isoform discovery and abundance estimation from RNA-seq data. *Genome research*, 23(3):519–529, 2013.

C. U. Mohien, D. R. Colquhoun, D. K. Mathias, J. G. Gibbons, J. S. Armistead, M. C. Rodriguez, M. H. Rodriguez, N. J. Edwards, J. Hartler, G. G. Thallinger, et al. A bioinformatics approach for integrated transcriptomic and proteomic comparative analyses of model and non-sequenced anopheline vectors of human malaria parasites. *Molecular & cellular proteomics*, 12(1):120–131, 2013.

A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, 5(7):621–628, 2008.

K. Murakami and T. Takagi. Gene recognition by combination of several gene-finding programs. *Bioinformatics*, 14(8):665–675, 1998.

S. Nahnsen, A. Bertsch, J. Rahnenführer, A. Nordheim, and O. Kohlbacher. Probabilistic consensus scoring improves tandem mass spectrometry peptide identification. *Journal of proteome research*, 10(8):3332–3343, 2011.

A. I. Nesvizhskii. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *Journal of proteomics*, 73(11):2092–2123, 2010.

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

A. I. Nesvizhskii and R. Aebersold. Interpretation of shotgun proteomic data the protein inference problem. *Molecular & cellular proteomics*, 4(10):1419–1440, 2005.

A. I. Nesvizhskii, A. Keller, E. Kolker, and R. Aebersold. A statistical model for identifying proteins by tandem mass spectrometry. *Analytical chemistry*, 75(17):4646–4658, 2003.

A. I. Nesvizhskii, O. Vitek, and R. Aebersold. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nature methods*, 4(10):787–797, 2007.

T. Nilsson, M. Mann, R. Aebersold, J. R. Yates, A. Bairoch, and J. J. Bergeron. Mass spectrometry in high-throughput proteomics: ready for the big time. *Nature methods*, 7(9):681–685, 2010.

K. Ning and A. I. Nesvizhskii. 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(Suppl 11):S14, 2010.

U. Omasits, M. Quebatte, D. J. Stekhoven, C. Fortes, B. Roschitzki, M. D. Robinson, C. Dehio, and C. H. Ahrens. Directed shotgun proteomics guided by saturated rna-seq identifies a complete expressed prokaryotic proteome. *Genome research*, 23(11):1916–1927, 2013.

G. Oshiro, L. M. Wodicka, M. P. Washburn, J. R. Yates, D. J. Lockhart, and E. A. Winzeler. Parallel identification of new genes in saccharomyces cerevisiae. *Genome research*, 12(8):1210–1220, 2002.

N. Palmieri, V. Nolte, A. Suvorov, C. Kosiol, and C. Schlötterer. Evaluation of different reference based annotation strategies using RNA-Seq - a case study in Drosophila pseudoobscura. *PLoS one*, 7(10):e46415, 2012.

C. Pan, B. Park, W. McDonald, P. Carey, J. Banfield, N. VerBerkmoes, R. Hettich, and N. Samatova. A high-throughput de novo sequencing approach for shotgun proteomics using high-resolution tandem mass spectrometry. *BMC bioinformatics*, 11(1):118, 2010.

V. Pavlović, A. Garg, and S. Kasif. A bayesian framework for combining gene predictions. *Bioinformatics*, 18(1):19–27, 2002.

D. N. Perkins, D. J. C. Pappin, D. M. Creasy, and J. S. Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20(18):3551–3567, 1999.

J. K. Pickrell, J. C. Marioni, A. A. Pai, J. F. Degner, B. E. Engelhardt, E. Nkadori, J.-B. Veyrieras, M. Stephens, Y. Gilad, and J. K.Pritchard. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, 464 (7289):768–772, 2012.

S. R. Ramakrishnan, C. Vogel, T. Kwon, L. O. Penalva, E. M. Marcotte, and D. P. Miranker. Mining gene functional networks to improve mass-spectrometry-based protein identification. *Bioinformatics*, 25(22):2955–2961, 2009.

L. Reiter, M. Claassen, S. P. Schrimpf, M. Jovanovic, A. Schmidt, J. M. Buhmann, M. O. Hengartner, and R. Aebersold. Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. *Molecular & cellular proteomics*, 8(11):2405–2417, 2009.

B. Y. Renard, W. Timm, M. Kirchner, J. A. Steen, F. A. Hamprecht, and H. Steen. Estimating the confidence of peptide identifications without decoy databases. *Analytical chemistry*, 82(11):4314–4318, 2010.

B. Y. Renard, B. Xu, M. Kirchner, F. Zickmann, D. Winter, S. Korten, N. W. Brattig, A. Tzur, F. A. Hamprecht, and H. Steen. Overcoming species boundaries in peptide identification with bayesian information criterion-driven error-tolerant peptide search (BICEPS). *Molecular & cellular proteomics*, 11(7):M111–014167, 2012.

P. Rice, I. Longden, and A. Bleasby. EMBOSS: the European molecular biology open software suite. *Trends in genetics*, 16(6): 276–277, 2000.

H. Safavi-Hemami, H. Hu, D. G. Gorasia, P. K. Bandyopadhyay, P. D. Veith, N. D. Young, E. C. Reynolds, M. Yandell, B. M. Olivera, and A. W. Purcell. Combined proteomic and transcriptomic interrogation of the venom gland of conus geographus uncovers novel components and functional compartmentalization. *Molecular & cellular proteomics*, 13(4):938–953, 2014.

W. S. Sanders, S. M. Bridges, F. M. McCarthy, B. Nanduri, and S. C. Burgess. Prediction of peptides observable by mass spectrometry applied at the experimental set level. *BMC bioinformatics*, 8(Suppl 7):S23, 2007.

A. Savidor, R. S. Donahoo, O. Hurtado-Gonzales, N. C. Verberkmoes, M. B. Shah, K. H. Lamour, and W. H. McDonald. Expressed peptide tags: an additional layer of data for genome annotation. *Journal of proteome research*, 5(11):3048–3058, 2006.

S. Schliesky, U. Gowik, A. P. Weber, and A. Bräutigam. Rna-seq assembly–are we there yet? *Frontiers in plant science*, 3:220, 2012.

A. C. Schrimpe-Rutledge, M. B. Jones, S. Chauhan, S. O. Purvine, J. A. Sanford, M. E. Monroe, H. M. Brewer, S. H. Payne, C. Ansong, B. C. Frank, R. D. Smith, S. N. Peterson, V. L. Motin, and J. N. Adkins. Comparative omics-driven genome annotation refinement: Application across Yersiniae. *PLoS one*, 7(3):e33903, 03 2012.

N. J. Schurch, C. Cole, A. Sherstnev, J. Song, C. Duc, K. G. Storey, W. I. McLean, S. J. Brown, G. G. Simpson, and G. J. Barton. Improved annotation of 3' untranslated regions and complex loci by combination of strand-specific direct rna sequencing, rna-seq and ests. *PloS one*, 9(4):e94270, 2014.

J. Seidler, N. Zinn, M. E. Boehm, and W. D. Lehmann. De novo sequencing of peptides by ms/ms. *Proteomics*, 10(4):634–649, 2010.

O. Serang, M. J. MacCoss, and W. S. Noble. Efficient marginalization to compute protein posterior probabilities from shotgun mass spectrometry data. *Journal of proteome research*, 9(10):5346–5357, 2010.

S. P. Shah, G. P. McVicker, A. K. Mackworth, S. Rogic, and B. F. F. Ouellette. GeneComber: combining outputs of gene prediction programs for improved results. *Bioinformatics*, 19(10):1296–1297, 2003.

S. T. Sherry, M.-H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin. dbsnp: the ncbi database of genetic variation. *Nucleic acids research*, 29(1):308–311, 2001.

D. Shteynberg, A. I. Nesvizhskii, R. L. Moritz, and E. W. Deutsch. Combining results of multiple search engines in proteomics. *Molecular & cellular proteomics*, 12(9):2383–2393, 2013.

R. D. Sleator. An overview of the current status of eukaryote gene prediction strategies. *Gene*, 461(1):1–4, 2010.

R. Sorek and P. Cossart. Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity. *Nature reviews genetics*, 11(1):9–16, 2010.

M. Stanke, O. Schöffmann, B. Morgenstern, and S. Waack. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC bioinformatics*, 7:62, 2006.

M. Stanke, M. Diekhans, R. Baertsch, and D. Haussler. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics*, 24(5):637–644, 2008.

H. Steen and M. Mann. The abc's (and xyz's) of peptide sequencing. *Nature reviews molecular cell biology*, 5(9):699–711, 2004.

M. Sultan, M. H. Schulz, H. Richard, A. Magen, A. Klingenhoff, M. Scherf, M. Seifert, T. Borodina, A. Soldatov, D. Parkhomchuk, D. Schmidt, S. O'Keeffe, S. Haas, M. Vingron, H. Lehrach, and M.-L. Yaspo. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, 321(5891):956–960, 2008.

S. Tanner, Z. Shen, J. Ng, L. Florea, R. Guigó, S. P. Briggs, and V. Bafna. Improving gene annotation using peptide mass spectrometry. *Genome research*, 17(2):231–239, 2007.

C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5):511–515, 2010.

C. Trapnell, A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel, S. L. Salzberg, J. L. Rinn, and L. Pachter. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols*, 7(3):562–578, 2012.

Q. Tu, R. A. Cameron, K. C. Worley, R. A. Gibbs, and E. H. Davidson. Gene structure in the sea urchin Strongylocentrotus purpuratus based on transcriptome analysis. *Genome research*, 22(10):2079–2087, 2012.

M. J. van Baren, B. C. Koebbe, and M. R. Brent. Using n-scan or twinscan to predict gene structures in genomic dna sequences. *Current protocols in bioinformatics*, pages 4–8, 2007.

C. J. van Rijsbergen. Information retrieval. *London: butterworths*, 2nd ed., 1979.

M. Vingron and M. S. Waterman. Sequence alignment and penalty choice: Review of concepts, case studies and implications. *Journal of molecular biology*, 235(1):1–12, 1994.

Q. Wang, Y. Lei, X. Xu, G. Wang, and L.-L. Chen. Theoretical prediction and experimental verification of protein-coding genes in plant pathogen genome Agrobacterium tumefaciens strain C58. *PLoS one*, 7(9):e43176, 09 2012.

X. Wang and B. Zhang. customprodb: an r package to generate customized protein databases from rna-seq data for proteomics search. *Bioinformatics*, 29(24):3235–3237, 2013.

X. Wang and B. Zhang. Integrating genomic, transcriptomic and interactome data to improve peptide and protein identification in shotgun proteomics. *Journal of proteome research*, 13(6):2715–2723, 2014.

X. Wang, R. J. Slebos, D. Wang, P. J. Halvey, D. L. Tabb, D. C. Liebler, and B. Zhang. Protein identification using customized protein sequence databases derived from rna-seq data. *Journal of proteome research*, 11(2):1009–1017, 2011.

X. Wang, Q. Liu, and B. Zhang. Leveraging the complementary nature of rna-seq and shotgun proteomics data. *Proteomics*, 14 (23-24):2676–2687, 2014.

Z. Wang, M. Gerstein, and M. Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1):57–63, 2009.

C. Wei and M. Brent. Using ESTs to improve the accuracy of de novo gene prediction. *BMC bioinformatics*, 7(1):327, 2006.

D. L. Wheeler, D. M. Church, S. Federhen, A. E. Lash, T. L. Madden, J. U. Pontius, G. D. Schuler, L. M. Schriml, E. Sequeira, T. A. Tatusova, et al. Database resources of the national center for biotechnology. *Nucleic acids research*, 31(1):28–33, 2003.

E. Wijaya, M. C. Frith, P. Horton, and K. Asai. Finding protein-coding genes through human polymorphisms. *PLoS one*, 8(1): e54210, 2013.

M. Wilhelm, J. Schlegl, H. Hahne, A. M. Gholami, M. Lieberenz, M. M. Savitski, E. Ziegler, L. Butzmann, S. Gessulat, H. Marx, et al. Mass-spectrometry-based draft of the human proteome. *Nature*, 509(7502):582–587, 2014.

S. Woo, S. W. Cha, G. Merrihew, Y. He, N. Castellana, C. Guest, M. MacCoss, and V. Bafna. Proteogenomic database construction driven from large scale rna-seq data. *Journal of proteome research*, 13(1):21–28, 2013.

T. Woyke and E. M. Rubin. Searching for new branches on the tree of life. *Science*, 346(6210):698–699, 2014.

T. Yada, T. Takagi, Y. Totoki, Y. Sakaki, and Y. Takaeda. Digit: a novel gene finding program by combining gene-finders. In *Proceedings of the 8th Pacific Symposium on Biocomputing (PSB 2003), Lihue, Hawaii, USA*, pages 375–387, 2002.

J. R. Yates III, J. K. Eng, and A. L. McCormack. Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. *Analytical chemistry*, 67(18):3202–3210, 1995.

N. Yok and G. Rosen. Combining gene prediction methods to improve metagenomic gene annotation. *BMC bioinformatics*, 12(1): 20, 2011.

J.-F. Yu, Z.-Z. Guo, X. Sun, and J.-H. Wang. A review of the computational methods for identifying the over-annotated genes and missing genes in microbial genomes. *Current bioinformatics*, 9(2):147–154, 2014.

F. Zickmann, M. S. Lindner, and B. Y. Renard. GIIRA – RNA-Seq driven gene finding incorporating ambiguous reads. *Bioinformatics*, 30(5):606–613, 2014.

Eigenständigkeitserklärung

Ich versichere, dass ich die hier vorgelegte Dissertation selbstständig angefertigt habe und die benutzten Quellen und Hilfsmittel vollständig angegeben sind.
Ein Promotionsverfahren wurde zu keinem früheren Zeitpunkt an einer anderen in- oder ausländischen Hochschule oder bei einem anderen Fachbereich beantragt.
Die Bestimmungen der Promotionsordnung sind mir bekannt.

Franziska Zickmann, Berlin, Februar 2015

Lebenslauf

For reasons of data protection,
the curriculum vitae is not included in the online version

Publikationen

Zickmann F and Renard, BY : *MSProGene - Integrative proteogenomics beyond six-frames and single nucleotide polymorphisms*. Bioinformatics, 2015, 31(12), i106-i115.

Zickmann F and Renard, BY : *IPred - Integrating ab initio and evidence based predictions for better gene identification*. BMC Genomics. 2015, 16(1), 134.

Calvignac-Spencer S, Schulze JM, **Zickmann F**, and Renard, BY : *Clock Rooting Further Demonstrates that Guinea 2014 EBOV is a Member of the Zaire Lineage*. Plos current Biology, 2014, 10.1371/currents.outbreaks.c0e035c86d721668a6ad7353f7f6fe86

Zickmann F, Lindner, MS, and Renard, BY : *GIIRA - RNA-Seq driven gene finding incorporating ambiguous reads*. Bioinformatics, 2014, 30(5), 606-613.

Giese, SH, **Zickmann F**, and Renard, BY : *Specificity Control for Read Alignments Using an Artificial Reference Genome Guided False Discovery Rate*. Bioinformatics, 2014, 30(1), 6-16.

Lindner, MS, Kollock, M, **Zickmann F**, and Renard, BY : *Analyzing genome coverage profiles with applications to quality control in metagenomics*. Bioinformatics, 2013, 29(10), 1260-1267.

Renard BY, Xu B, Kirchner M, **Zickmann F**, Winter D, Korten S, Brattig NW, Tzur A, Hamprecht FA, and Steen H : *Overcoming species boundaries in peptide identification with Bayesian information criterion-driven error-tolerant peptide search (BICEPS)*. Mol Cell Proteomics, 2012, 11(7), M111.014167.

Scornavacca C, **Zickmann F**, and Huson DH : *Tanglegrams for rooted phylogenetic trees and networks*. Bioinformatics, 2011, 27(13), 248-256.