

Visual Analysis of RNAseq Data

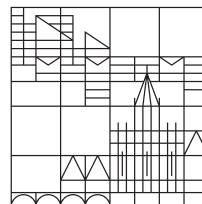
Discovering Genes in Bacteria

Dissertation zur Erlangung des
akademischen Grades eines Doktors der
Naturwissenschaften

vorgelegt von
Svenja Simon

an der

Universität
Konstanz



Mathematisch-Naturwissenschaftliche Sektion
Informatik und Informationswissenschaft

Tag der mündlichen Prüfung: 17. Juli 2015

1. Referent: Prof. Dr. Daniel A. Keim
2. Referent: Prof. Dr. Siegfried Scherer

Abstract

RNA sequencing (RNAseq) using next-generation-sequencing (NGS) technologies allows, nowadays, to produce transcriptomic data in a high throughput fashion. However, the analysis of these large and complex biological data sets remains a great challenge. This analysis is highly of explanatory nature and requires to constantly connect observations with implicit domain knowledge. This requires interactive visual analysis systems and an expert user in the analysis loop. The challenge of designing interactive visual analysis systems for the analysis of RNAseq data demands interdisciplinary research at the interface between molecular biology and visual data analysis. However, the epistemic distance between both fields is typically very high and, therefore, knowledge gaps and interdisciplinary communication issues hamper effective collaboration. In order to bridge the knowledge gap between domain and visualization experts, I introduce the *Liaison* role for problem-driven research in the visualization domain which fosters a better and richer interdisciplinary communication. In this thesis, I contribute a problem characterization and task descriptions to discover and describe genes using RNAseq data. Based on the problem characterization, I identify two research gaps: First, assessing the trustworthiness of RNAseq data in the analysis and, second, discovering and relating genes to identify their functions. With the systems *NGS Overlap Searcher* and *VisExpress*, I present two visual analysis solutions that address these research gaps. Furthermore, I evaluate and apply both systems on real data sets with real experts leading to important insights for the biological domain as well as for problem-driven visualization research.

Zusammenfassung

Die Anwendung von Sequenzierungstechnologien der nächsten Generation (next-generation-sequencing (NGS)) erlaubt es heute Transkriptomdaten mit hoher Durchsatzgeschwindigkeit zu produzieren (RNAseq). Die Analyse dieser großen und komplexen biologischen Datensätze bleibt allerdings eine große Herausforderung, da hier die Exploration der Daten im Vordergrund steht und Beobachtungen immer mit implizitem Expertenwissen in Zusammenhang gebracht werden müssen. Daher werden interaktive visuelle Analysesysteme benötigt, die Experten in den Analysezyklus miteinbeziehen. Um der Herausforderung zu begegnen, interaktive Visualisierungssysteme für die Analyse von RNAseq Daten zu entwickeln, wird eine interdisziplinäre Forschung an der Schnittstelle zwischen Molekularbiologie und visueller Datenanalyse benötigt. Wissenslücken und Probleme der interdisziplinären Kommunikation, die durch die hohe epistemische Distanz zwischen beiden Forschungsgebieten gehäuft vorkommen, behindern allerdings eine effektive Kollaboration. Um diese Wissenslücke zwischen Domänen- und Visualisierungsexperten zu überbrücken und eine bessere und reichere Kommunikation zu fördern, führe ich die *Liaison* Rolle für problemorientierte Forschung im Bereich Visualisierung ein, die zwischen beiden Fachgebieten vermittelt. Mit dieser Dissertation trage ich eine Problemcharakterisierung sowie eine Beschreibung von Aufgaben bei, um Gene mit Hilfe von RNAseq Daten zu entdecken und zu beschreiben. Basierend auf dieser Problemcharakterisierung identifiziere ich zwei Forschungslücken: Erstens, die Vertrauenswürdigkeit von RNAseq Daten in der Analyse zu bewerten und zweitens, Gene zu entdecken und miteinander in Verbindung zu bringen, um ihre Funktionen aufzuklären. Mit den Systemen *NGS Overlap Searcher* und *VisExpress* stelle ich zwei visuelle Analyse Systeme vor, welche die genannten Forschungslücken behandeln. Beide Systeme wurden mit echten Daten und von echten Experten angewandt und evaluiert, was zu wichtigen neuen Einblicken in der Biologie, als auch im Gebiet der problemorientierten Visualisierungsforschung geführt hat.

Danksagung

An erster Stelle möchte ich meinem Doktorvater Prof. Daniel A. Keim danken. Mit seiner Unterstützung, aber auch durch die Freiheit meine Forschungsthemen selbst zu wählen, hat Prof. Keim meinen Weg zu einer eigenständigen und erfolgreichen Forscherin unterstützt.

Ferner gilt mein Dank meinen Kooperationspartnern im FOG-Project (s.u.). Des Weiteren möchte ich Klaus Neuhaus, Richard Landstorfer, Lea Fellner und Prof. Siegfried Scherer aus Freising sowie Katharina Mir, Steffen Schober und Prof. Martin Bossert aus Ulm für die vielen interessanten und lehrreichen Gespräche und die gute Zusammenarbeit danken.

Mein Dank gilt ebenfalls meinen Kollegen am Lehrstuhl Datenanalyse und Visualisierung für die stets gute Zusammenarbeit und Arbeitsatmosphäre. Insbesondere möchte ich meinen lieb gewonnenen Bürokollegen Andrada Tatu, Miloš Krstajić und Hansi Senaratne, sowie Dr. Andreas Stoffel und Dr. Daniela Oelke für ihre Unterstützung und ihre Ratschläge danken.

Letztlich gilt mein Dank meiner Familie, die mich immer moralisch unterstützt hat und in besonderem Maße meinem Freund Sebastian Mittelsädt für seine persönliche wie fachliche Unterstützung. Meiner guten Freundin Angela Gilles danke ich für das Korrekturlesen dieser Arbeit.

Nicht zu vergessen ist auch die Deutsche Forschungsgemeinschaft, die meine Dissertation im Rahmen des folgenden Projekts finanziert hat:

FOG-Project: “Finding new overlapping genes and their theory (FOG-Theory)”, part of the priority programme “Information and Communication Theory in Molecular Biology” (InKoMBio SPP 1395) of the German Research Foundation (DFG), 2010-2015.

Contents

I	Introduction	1
I-1	Information Visualization and Visual Analytics	1
I-2	Biological Data Visualization	2
I-3	Structure and Contributions of this Thesis	3
I-4	Challenges in Visual Analytics and Biological Visualizations	4
I-5	Citation Conventions	8
I-6	Publications Utilized for this Thesis	9
I-7	Further Publications	13
II	Bridging the Gap to Domain Experts - The <i>Liaison</i> Role	15
II-1	Introduction	15
II-2	Related Work	16
II-3	The Interdisciplinary Communication Issue	17
II-4	The <i>Liaison</i> Role	19
II-4.1	How to Become a <i>Liaison</i>	19
II-4.2	Instantiations of the <i>Liaison</i> Role and the VIS Team	20
II-5	Benefits and Tasks of the <i>Liaison</i> and the VIS Team	21
II-5.1	Example Project with a <i>Liaison</i> - The <i>VisExpress</i> -Project	21
II-5.2	Tasks & Benefits	23
II-6	Discussion and Limitations	24
III	Requirement Analysis and Problem Abstraction	27
III-1	Introduction	28
III-2	Biological Background	28
III-2.1	Open Reading Frames and Genes	29
III-2.2	RNAseq Using Next-Generation-Sequencing	33
III-3	Data	40

CONTENTS

III-3.1	Genomic Data	40
III-3.2	RNAseq Data	42
III-4	Problem Abstraction	44
III-4.1	Task Taxonomy	45
III-4.2	Assessment of the Trustworthiness of RNAseq Measurements	47
III-4.3	Comparison of Gene Activity Levels between Different Conditions	52
IV	Visual Analysis for the Trustworthiness Assessment of RNAseq Measurements	57
IV-1	Introduction	58
IV-2	Requirements	59
IV-3	State of the Art and Related Work	60
IV-4	Pixel-based Representation of RNAseq Reads Coverage	63
IV-5	The <i>NGS Overlap Searcher</i> - An Enhanced Genome Browser	68
IV-5.1	System Architecture	68
IV-5.2	Visualization of RNAseq Read Coverage	69
IV-5.3	Providing an Interestingness Function	71
IV-5.4	Evaluation	74
IV-6	Discussion & Lessons Learned	75
IV-7	Limitations & Future Work	76
V	Visual Analysis of Differential Gene Expression	79
V-1	Introduction	79
V-2	Requirements	82
V-3	State of the Art and Related Work	83
V-4	The <i>VisExpress</i> System	85
V-4.1	Design Process	85
V-4.2	Architecture of <i>VisExpress</i>	87
V-4.3	Visualizing GAR Patterns	88
V-4.4	Components of <i>VisExpress</i>	94
V-4.5	Interaction Design of <i>VisExpress</i>	99
V-4.6	Evaluation	102
V-5	Discussion & Lessons Learned	110
V-6	Limitations & Future Work	112
VI	Concluding Remarks and Perspectives	113
VI-1	Experiences and Lessons Learned	114
VI-2	Were Are We Today?	116

CONTENTS

VI-3 Further Challenges in Biological Data Visualization	119
My own Publications	123
References	127

CONTENTS

Glossary

count

Number of reads overlapping an [open reading frame \(ORF\)](#). xi, xiii, 36, 42, 43, 71

differential gene expression

A gene or ORF is differentially expressed, if its [gene activity levels](#) differ significantly between two experiment conditions. Different statistical methods exist to determine differentially expressed genes (ORFs). In this thesis, all pair-wise comparisons (n:n) are considered, however, in many studies only a (1:n) comparison is considered. Thus, several conditions are compared against one reference condition. See also Section [RNAseq Data](#) (p. 42) and [Differential Gene Expression Data](#) (p. 43). 8, 42, 43, 52, 53, 80–82, 115

FOG-Project

FOG-Project: “Finding new overlapping genes and their theory (FOG-Theory)”, part of the priority programme “Information and Communication Theory in Molecular Biology” (InKoMBio SPP 1395) of the German Research Foundation (DFG), 2010-2015. 3, 32, 34, 44, 70, 80, 86, 102, 114, 117

fold-change

Fold change denotes the ratio between the [gene activity levels \(counts\)](#) of two experiment conditions. See Section [RNAseq Data](#) (p. 42). xii, 42, 43, 93

gene activity level

Value describing the strength of transcription of an ORF. For instance, the normalized [Reads Per Kilobase per Million mapped reads \(RPKM\)](#) value or the raw [counts](#). xi, xii, xiv, 42–45, 47, 52, 70–72, 82, 99, 114

gene activity ratio (GAR)

The gene activity ratio is the ratio of the [gene activity levels](#) of a gene (or ORF) of two experiment conditions. The gene activity ratio is also denoted as [fold-change](#). [xii](#), [43](#), [82](#), [91](#), [98](#), [115](#)

gene activity ratio (GAR) pattern

The GAR pattern of a gene (or ORF) comprises the [gene activity ratios \(GARs\)](#) of all pairs of condition comparisons. For instance for four conditions, the GAR of condition 1vs2, 1vs3, 1v4, 2vs1, 2vs2, 2vs3, 2vs4, 3vs1, 3vs2, 3vs3, 3vs4, 4vs1, 4vs2, 4vs3, 4vs4. See [V.4](#) (p. [89](#)) for an illustration. [43](#), [53](#), [55](#), [82](#), [88](#), [89](#), [94](#), [99](#), [100](#), [117](#)

gene activity ratios (GARs)

See [gene activity ratio \(GAR\)](#). [xii](#)

next-generation-sequencing (NGS)

Sequencing technologies of the next generation which sequences DNA in a high throughput fashion by synthesis. See Section [RNAseq Using Next-Generation-Sequencing](#) (p. [33](#)). [2](#), [33](#), [58](#), [79](#), [114](#)

nucleotide

The DNA is composed of the **nucleotides** adenine (A), cytosine (C), guanine (G) and thymine (T). See also Section [Open Reading Frames and Genes](#) (p. [29](#)). [29](#), [33](#), [42](#), [65–67](#)

open reading frame (ORF)

An Open Reading Frame is defined by a start and a stop codon on the same reading frame. ORFs which encode for a protein are denoted genes, i.e., the term ORFs comprises genes as well as ORF not known to be coding. See Section [Open Reading Frames and Genes](#) (p. [29](#)). [xi](#), [xii](#), [xiv](#), [29](#), [42](#), [47](#), [51](#), [52](#), [58](#), [69](#), [70](#)

open reading frames (ORFs)

See [open reading frame \(ORF\)](#). [40](#), [44](#), [50](#), [56](#), [58](#), [59](#), [63](#), [72](#), [74](#), [117](#)

operon

An operon describes several adjacent genes which are transcribed together. They build one long mRNA. In general, the genes of an operon are functionally related. [50](#)

overlapping genes (OLGs)

An overlapping gene pair is defined by two genes whose reading frames overlap at the same genome location. See Section [Overlapping Genes](#) (p. 32). [xiv](#), [32](#), [40](#), [44](#), [52](#), [70](#), [74](#), [77](#), [111](#)

pathogenicity

Pathogenicity is the ability to cause a disease. [2](#), [44](#), [79](#)

plasmid

Plasmids are small DNA molecules which exist separately in many bacteria cells. [42](#), [48](#)

Polymerase Chain Reaction (PCR)

The Polymerase Chain Reaction allows an exponential multiplication of the DNA fragments. See Section [Amplification by Polymerase Chain Reaction](#) (p. 36). [36](#)

read

Sequenced fragments of DNA are named **reads**. See Section [RNAseq Using Next-Generation-Sequencing](#) (p. 33). [xiv](#), [33](#), [40](#), [42](#), [60](#), [63](#)

read coverage

Read coverage describes the number of reads mapped to each genome position. The reads overlapping an ORF are denoted as the read coverage of the ORF, meant is a vector with the numbers of overlapping reads per ORF position which can be visualized as a line chart (Figure [III.5](#) (p. 35)). Due to uncertainties in the RNA sequencing, experts need to assess the trustworthiness of the read coverage to verify a gene (or ORF) as active. An active ORF is most likely a gene which has not been detected yet. See also Section [RNAseq Using Next-Generation-Sequencing](#) (p. 33). [xiv](#), [34](#), [36](#), [42](#), [44](#), [50](#), [58–60](#), [63](#), [65–67](#), [69–72](#), [74](#), [82](#), [115](#)

Reads Per Kilobase per Million mapped reads (RPKM)

Normalized value to describe the strength of transcription of an ORF. **Counts** are normalized for the length of the ORF and the total number of reads mapped to the genome in the respective sequencing run. See Section [RNAseq Data](#) (p. 42). [xi](#), [42](#), [99](#)

RNA Sequencing (RNAseq)

RNAseq describes the use of NGS to indirectly sequence and, therefore, quantify RNA from a genome at a given condition and moment in time. RNA is transcribed to DNA

Glossary

(so-called copy DNA), since NGS can only sequence DNA. See Section [RNAseq Using Next-Generation-Sequencing](#) (p. 33), as well as [RNAseq Data](#) (p. 42). 2, 33, 44, 58, 79

RNAseq measurement

RNAseq measurements are the [reads](#) mapped to the genome. The RNAseq measurement of a gene (or ORF) are the reads mapped to the gene (or ORF) and can be described by the [read coverage](#) or the [gene activity level](#). If a gene (or ORF) has a trustworthy RNAseq measurement, the gene is active. Due to uncertainties in the RNAseq measurement, experts need to assess the trustworthiness of the RNAseq measurement to verify a gene (or ORF) as active. An active ORF is most likely a gene which has not been detected yet. See also Section [RNAseq Using Next-Generation-Sequencing](#) (p. 33) and [Assessment of the Trustworthiness of RNAseq Measurements](#) (p. 47). 44, 47, 48, 50–52, 58, 59, 63, 69, 74, 80, 116

shadow ORF (sORF)

An [open reading frame \(ORF\)](#) which overlaps with a gene is denoted shadow ORF (sORF). sORFs are potential [overlapping genes \(OLGs\)](#). See Section [Overlapping Genes](#) (p. 32). 40

untranslated region (UTR)

Untranslated region (UTR): The transcription of a gene starts before the start codon and ends after the stop codon. The regions not belonging to the gene are called untranslated regions (UTRs), as they are not translated into the protein. The UTR of one gene might start or end within an adjacent gene. [xiv](#), 49, 59, 71

UTR

See [untranslated region \(UTR\)](#). 50, 71

Chapter I

Introduction

I-1 Information Visualization and Visual Analytics

As vision is one of our most important human senses, it is not surprising that visualizations have been used early on in human history. Starting with rock engravings as a symbolic communication in prehistory, of which some could be as old as 40,000 years, symbolic representations have nowadays become an important medium of communication. So called *Infographics* are frequently used in print and online media as well as on television to present information, for entertainment, or both.

However, advancements in computer graphics and computer science in general have opened up the possibility to combine graphics with interactions, enabling to use visualization for interactive exploration in knowledge generation. This offers a great advantage over pure automatic method, for which tasks or data patterns have to be well defined since computers can only provide results if the human asks the right questions in the right way. However, tasks are often ill-defined when researchers want to advance the state-of-the art in their domain. Often they can just state that they want to gain new insights from their data [[van Wijk, 2006](#)]. The advantage of visualization is to incorporate the expert into the analysis process. Experts can steer the analysis to match the current analysis task and help to answer or even to identify new questions. A further advantage of visualizations is the efficiency of our vision system to identify patterns intuitively that may be hard to verbalize or describe in a form that a computer would understand. Studies of visual representations to reinforce such human cognition processes build the research field of visualizations (Vis).

In the field, three directions are distinguished: Information Visualization (InfoVis), Visual Analytics (VA) and Scientific Visualization (SciVis). InfoVis combines visualization techniques with interactions to build systems which support users in analyzing their data interactively. VA is closely related to InfoVis but has the focus to tightly integrate visualizations with automatic models, as visualizations can help to understand and steer algorithms. One advantage of the integration of visualizations and automatic methods is, for instance, that meanings behind automatic method parameters can be conveyed, leading to the possibility of an informed parameter adjustment within the system. Furthermore, resulting uncertainties can be incorporated in the data visualizations. While in InfoVis and VA the spatial representation can be chosen to

I. INTRODUCTION

represent a data attribute, the spatial representation is given in SciVis. SciVis can be defined by visualizing data with an inherent structure, in which the continuous spatial dimensions express (natural) structural information. Often 3D phenomena are considered, for instance, computer tomography measurements or 3D structures of molecules.

Visualization research is mainly driven by real-world problems. Either directly by problem-driven research which deals with real users, real data and relevant domain problems or indirectly by technique-driven research. Technique-driven research develops new techniques for general (abstract) tasks and/or data set which are applicable in several domains. Other directions address evaluations and meta-research categories like methodologies. The challenges in visualization, especially visual analytics research, have been discussed by Keim *et al.* [Keim and Zhang, 2011, Keim *et al.*, 2009, 8]. See Section I-4 for more information.

The books of Colin Ware [Ware, 2004], Ward *et al.* [Ward *et al.*, 2010] and Tamara Munzner [Munzner, 2014] provide an overview on the Vis field, from perception to design, techniques and applications. “Mastering the information age - solving problems with visual analytics” of Keim *et al.* [Keim *et al.*, 2010], introduces and discusses visual analytics in more detail.

I-2 Biological Data Visualization

Advances in molecular biology can lead to new knowledge about diseases and development of new medical treatments (medicines). However, the complex relations and dependencies in biology necessitate a human in the analysis to connect implicit domain knowledge with measured data. Furthermore, high-throughput technologies have lead to the need of exploration to generate new hypotheses from the immense data volumes.

Many genes are, for instance, still not discovered, even in well researched organisms like *Escherichia coli*. Furthermore, the function of many genes remains unknown. The exploration of the functions of bacteria genes would open up many lines of research. An improved understanding of human pathogenicity would help, for example, to develop new medical treatments and a better comprehension of bacteria utilized in biotechnology would contribute to the production of new substances. RNA Sequencing (RNAseq) by next-generation-sequencing (NGS) is a technology which allows to make advancements in this direction. RNAseq enables measurement of genes in a high throughput fashion. The large and complex data sets necessitate new scalable and interactive data analysis approaches which support directed verification of hypothesis, as well as data exploration. In this context, I address an interesting and fascinating molecular biology topic in my thesis - the visual analysis of RNAseq data to discover and describe genes in bacteria.

Furthermore, many general visualization challenges need to be addressed to design and develop interactive visualization systems for molecular biology applications. The most important ones are: *scalability*, *uncertainty*, *evaluation* and *interestingness*, which I discuss in Section I-4 in detail. I see further challenges which are especially relevant in the biological domain. First, the challenge to bridge the gap between domain and visualization experts and, second, to abstract data and tasks in an appropriate way to address scalability, uncertainty and interactions. See Section I-4 for a discussion of these points.

I-3 Structure and Contributions of this Thesis

My aim was to orient my work on real problems in RNAseq analysis, therefore, I performed problem-driven visualization research in the course of the [FOG-Project](#)¹. Due to the interdisciplinary nature of this project, I had the opportunity to collaborate closely with domain experts and to analyze real data.

The reader can learn from this thesis the specifics of visualization challenges in problem-driven biological research (next Section I-4). In chapter II I introduce the *Liaison* role to tackle the general problem of interdisciplinary research which is the knowledge gap between domain and visualization experts and the interdisciplinary communication issue leading to misunderstandings in communication. The definition of this role and its tasks description allows readers to utilize this role in their own problem-driven visualization research to overcome the general *Bridging the gap* challenge between domain and visualization experts.

Chapter III provides the reader with an introduction to the biological topic of genes and RNAseq data. In Section III-4 of this Chapter I identify the two main research gaps for the (visual) analysis of RNAseq data to discover and describe genes in bacteria: first to assess the trustworthiness of RNAseq measurements and second to discover and relate genes to identify their functions. For these research gaps I contribute two analysis systems that are described in Chapter IV and Chapter V. Based on the problem characterization and a set of abstracted tasks, readers can develop alternative systems. Definitions of *interestingness* and *uncertainty* are given to bypass, respectively understand, these challenges for the stated tasks in RNAseq analysis.

Chapter IV introduces the *NGS Overlap Searcher* system which allows to assess the trustworthiness of RNAseq measurements. Thereby, the *NGS Overlap Searcher* provides a solution to address the *scalability* and the *uncertainty* challenge for the described tasks in RNAseq analysis.

¹FOG-Project: “Finding new overlapping genes and their theory (FOG-Theory)”, part of the priority programme “Information and Communication Theory in Molecular Biology” (InKoMBio SPP 1395) of the German Research Foundation (DFG), 2010-2015

I. INTRODUCTION

Chapter V introduces the *VisExpress* system which supports data exploration to discover and detect new genes as well as to relate genes with functions. Thereby, the *VisExpress* system provides a solution to address the *scalability* and the *uncertainty* challenge for the described tasks in RNAseq analysis. The design of *VisExpress* is validated with a pair analytics study [Arias-Hernandez et al., 2011], showing the applicability of this approach to address the *evaluation* challenge.

Chapter VI will conclude the thesis, summarizing the contributions and outlining a number of interesting open issues for future research.

I-4 Challenges in Visual Analytics and Biological Visualizations

Thomas and Cook, and Daniel A. Keim *et al.* have introduced researcher directions and challenges of visual analytics in the books “Illuminating the Path: Research and Development Agenda for Visual Analytics” [Thomas and Cook, 2005] and “Mastering the information age - solving problems with visual analytics” [Keim et al., 2010]. Keim *et al.* have also further discussed visual analytics challenges in [Keim et al., 2009, Keim and Zhang, 2011, 8]¹: *scalability, uncertainty, hardware, interaction, evaluation, infrastructure, interestingness* and *text data stream*.

I add a description of two further challenges which seem to be relevant, in my opinion, especially in the biological domain: *bridging the gap* and *abstraction*. All challenges are covered by the main aim of visual analytics which is to generate new knowledge with visual analytics systems (see Sacha *et al.* [Sacha et al., 2014]).

Hereinafter I briefly explain the challenges most relevant for biological data visualization and name the specific biological characteristics. Challenges not included in [Keim et al., 2009, Keim and Zhang, 2011, 8] are marked with *.

Bridging the gap*

If visual analytics addresses complex real world problems, the first challenge is to characterize the domain problem. This is often hampered by a knowledge gap between domain and visualization experts. This is also described as the interdisciplinary communication issue (see Section II-3 and [13]). A missing mutual knowledge and different domain languages, often lead to misunderstandings and sub-optimal designs. Beside the knowledge gap between domain and visualization experts, an interest gap exists. Domain experts need a tool to accomplish their

¹I co-authored the publication Keim *et al.* [8]. However, I have not contributed to the challenge definition. See for the work distribution [Solving Problems with Visual Analytics: Challenges and Applications](#) (p. 12).

aims which might be a simple or automatic solution. Visualization experts are interested in visualization research and do not want work as a toolsmith. Van Wijk has described these gaps and discussed how to bridge them in [van Wijk, 2006].

Specific biological Bridging the gap characteristics. Molecular biology is an especially complex domain, with its own domain language and expectations for many rules. Thus, bridging the knowledge gap is hard and often needs much time. See Chapter II for an approach to address this issue with a *Liaison*.

Defining Interestingness

The human visual system is powerful in perceiving patterns. However, complex and big data necessitates to abstract data. To do so, user tasks need to be considered since they define which data aspects are of interest. Since tasks are often ill-defined, this necessitates understanding interestingness in the domain. This understanding allows subsequently to match the mental model of domain experts with the visual and interaction design to optimally support users. Additionally, automatic methods need to be defined which can capture the interesting parts in the data. Feedback mechanisms could be used in this context to learn individual interestingness functions, based on user behavior.

Specific biological interestingness characteristics. See last subsection [Specific biological Bridging the gap characteristics](#). Chapter III provides an abstraction for RNAseq data and the tasks to discover and describe genes, leading to the definition of interestingness for the aim to discover and describe genes.

Achieving an Abstraction*

Achieving a meaningful abstraction for data and tasks is challenging as a grounded domain knowledge is needed, as well as a grounded knowledge of task analysis and visualization techniques. Furthermore, algorithms might need to be abstracted or replaced by heuristics to allow a subsequent scalability and seamless interactions. Thus, also knowledge on the algorithmic side is needed. A good abstraction matches the mental model of the domain experts to support the generation of insights. See also “Data Representations and Transformations (Chapter 4)” in [Thomas and Cook, 2005].

I. INTRODUCTION

Specific biological abstraction characteristics. See subsection [Specific biological Bridging the gap characteristics](#) (p. 5). Chapter III provides an abstraction for RNAseq data and the tasks to discover and describe genes.

Conveying Uncertainty

Uncertainty can occur in visualization on different levels. First on the raw data level, second on the pre-processing data level and, third, on the perception level. Raw data can already be erroneous, for instance, due to inaccurate measured or missing values. Sometimes the strength of bias can be stated and visualized if a measuring instrument has a known margin of error. In other cases the bias cannot be stated. Also data pre-processing can introduce uncertainties, for instance, by simple data pre-processing steps like binning or complex ones like data models or prediction which might be inaccurate. However, if uncertainty can be measured, for instance, as the confidence of an analysis algorithm, visualizations can incorporate and represent these to raise the awareness of users for data quality. The third aspect is the human perception of visual representations and color. Some visual representations are more accurate than others, and some visual designs might be misleading and ambiguous which depends predominantly on the analysis task and on faithfully representing data [[Mittelstädt et al., 2015a](#)]. Beside the type of visual representations, color is an important visual variable but color vision is also influenced by contrast effects [[Mittelstädt et al., 2014](#)]. Humans perceive colors differently, depending on the surrounding color. Therefore, designers of visualization systems have to consider accepted design guidelines, as well as human perception and cognition principles. Furthermore, all measurable and relevant uncertainties need to be incorporated in the design to enable the users to make informed decisions.

Specific biological uncertainty characteristics. Biological data can contain many unmeasurable uncertainty sources, due to many consecutive error-prone data preparation steps. This is the case, for instance, with RNAseq data (see Section [III-2.2](#)). Furthermore, the awareness of quality and the assessment of trust is very important for biologists, as subsequent validation experiments are time and cost intensive (see Section [III-4.2](#) and [III-4.3](#)). See chapters [IV](#) and [V](#) for systems which address the *Uncertainty* challenge.

Reaching Scalability

Problems addressed with visualizations often deal with complex, heterogeneous and large data sets (Big Data). Following, a visual representation of all data and/or all different data aspects is not possible due to limited screen space. Furthermore, automatic analysis slows

I-4 Challenges in Visual Analytics and Biological Visualizations

down with data size which can only be partially compensated by modern computer hardware, as especially visual analytics requires real-time interactions. Therefore, appropriate data aggregations are needed to analyze data according to the visual information seeking mantra of Shneiderman [Shneiderman, 1996] “Overview first, zoom and filter, then details-on-demand”. Automatic analysis needs to be replaced or combined with heuristic approaches to provide users with estimated solutions. Furthermore, visualizations of preliminary analysis results are an interesting direction as these allow users to steer and influence the algorithms by adjusting parameters during runtime.

Specific biological scalability characteristics. For visualization of large and complex biological data sets, data aggregations are needed which are effective and intuitive to read. A limitation might be here that biologist are often not trained in visualizations, hampering the use of too complex visualizations. Specific trainings might be a solution here. However, how to effectively teach visualizations still needs more research. See Chapters IV and V for systems which address the *Scalability* challenge.

Learning form Evaluation

Visual Analytics solutions address complex real world problems which often aim to advance the state of the art in the application domain by generating new insights from data. This fact hampers an evaluation in form of control lab studies - insight is not directly measurable. New evaluation methodologies need to be developed, which account for this fact, and address all steps of the design process. See papers of Munzner and Meyer *et al.* [Munzner, 2009, Meyer *et al.*, 2013, McKenna *et al.*, 2014] going in the same direction. Additionally, evaluation methods are needed which help to achieve a better understanding how visualizations support human cognition and decision processes. See Arias-Hernandez *et al.* [Arias-Hernandez *et al.*, 2011] for an approach to address this issue.

Specific biological evaluation characteristics. Even if real data is analyzed by real users in an evaluation, the complexity of the biological domain can hamper to capture and/or understand gained insights for the visualization researcher. This hampers to extend or improve the visualization solution in the right way. See Chapter II for an approach to address this issue with a *Liaison* and chapter V for a system which has a strong evaluation.

Further Specific Biology Characteristics

Biology is a very fast advancing and developing field. On the one side, new technologies and falling cost bring up new interesting lines of research which could not be addressed before.

I. INTRODUCTION

However, often accuracy and biases are, at first, not known for new technologies. On the other side, new research results bring up new questions for old data sets which could be re-analyzed in this respect. Public data sets can be analyzed, e.g., for overlapping genes (see [Overlapping Genes](#) (p. 32)). However, such problems are often very specific. Necessitating to develop many different and specially tailored systems. In this connection it is challenging to identify commonalities between different specific tasks to increase the applicability of systems. Additionally, systems should be decoupled from data sources, as these can change over time. For instance, many systems have been devolved to analyze [differential gene expression](#) from DNA mircoarrays but nowadays RNAseq has become the standard for gene expression data.

I-5 Citation Conventions

This thesis is based on published papers I authored or co-authored (see also [Publications Utilized for this Thesis](#) (p.9)). A different reference style is used in order to distinguish these publications from references. My publications are numbered with arabic numbers, for instance, [14]. References are cited with aberrations, for instance, [[Sedlmair et al., 2012b](#)].

Most chapters and sections comprise some content of my publications. Parts of these chapters appeared verbatim in my publications ¹. Other parts are based on my publications, but the text is paraphrased and extended. At the beginning of each chapter or section I state the publication it is based on. For instance:

Note

This chapter is based on the following publication and parts of this chapter appeared in this publication [12]:

Svenja Simon, Sebastian Mittelstädt, Daniel A. Keim, and Michael Sedlmair. “*Bridging the Gap of Domain and Visualization Experts with a Liaison.*” Eurographics Conference on Visualization (EuroVis) - Short Papers, Cagliari, Italy, 25 - 29 May 2015, 127-133, The Eurographics Association, [10.2312/eurovisshort.20151137](#), 2015.

Paragraphs that are based on the contributions (and text) of co-authors are quoted and italicized. Related work and state-of-the-art is cited according to the common reference style in the computer science community, for instance: Sedlmair *et al.* introduced the design study methodology framework [[Sedlmair et al., 2012b](#)].

¹All parts which are copied from publications are written by myself or quoted. See also [Publications Utilized for this Thesis](#) (p.9) for a listing of the work distribution among the co-authors.

I-6 Publications Utilized for this Thesis

This section lists all publications utilized for this thesis: [First Author Publications](#) and [Co-Authored Publications](#). For each publication the contributions are stated and assigned to the corresponding author. Furthermore, the division of responsibilities and work is stated.

First Author Publications

Bridging the Gap of Domain and Visualization Experts with a Liaison

Svenja Simon, Sebastian Mittelstädt, Daniel A. Keim, and Michael Sedlmair. “*Bridging the Gap of Domain and Visualization Experts with a Liaison.*” Eurographics Conference on Visualization (EuroVis) - Short Papers, Cagliari, Italy, 25 - 29 May 2015, 127-133, The Eurographics Association, DOI: [10.2312/eurovisshort.20151137](https://doi.org/10.2312/eurovisshort.20151137), 2015. [12]

The main research problem, how to deal with the knowledge gap between domain and visualization experts, was identified in a discussion by myself, S. Mittelstädt, A. Stoffel, BC Kwon and D.A. Keim, during the paper project *VisExpress*. **The contributions** of this paper are:

1. Description of the *Liaison* role and its variations to address the interdisciplinary communication issue (ICI).

1a. A simple model, based on a metaphor of spaces to illustrate the ICI.

2. Guidelines how to utilize and integrate the *Liaison* in the design process.

3. A discussion of benefits and pitfalls of the *Liaison* role based on experiences in the *VisExpress* design study.

Identification and Development of Contributions:

Contribution 1 and 2 were identified in a discussion with all authors and developed by myself.

Contribution 1a was identified by M. Sedlmair and developed by M. Sedlmair and myself.

Contribution 3 was identified and developed by myself.

Implementation: Does not apply.

Authorship: The paper is written by myself. All authors reviewed the paper.

Supervision: M. Sedlmair and D.A. Keim supervised the paper project and commented on paper drafts and contributions.

I. INTRODUCTION

Visual Analysis of Next-Generation Sequencing Data to Detect Overlapping Genes in Bacterial Genomes

Svenja Simon, Daniela Oelke, Richard Landstorfer, Klaus Neuhaus, and Daniel A. Keim. “*Visual Analysis of Next-Generation Sequencing Data to Detect Overlapping Genes in Bacterial Genomes.*” 2011 IEEE Symposium on Biological Data Visualization, October 23 - 24, Providence, Rhode Island, USA, 47-54, IEEE, DOI: [10.1109/BioVis.2011.6094047](https://doi.org/10.1109/BioVis.2011.6094047), 2011. [14]

The main research challenge, expressively visualizing RNAseq data and guiding the search in large RNAseq data sets, was identified in discussion with all authors. **The contributions** of this paper are:

1. A representation of the RNAseq measurements without introducing artifacts.
2. A visualization of RNAseq measurements in the open reading frame (ORF) representation allowing to determine how well the transcript fits to the ORF location.
3. A filter functionality to focus on interesting ORFs to handle the large data volumes.
4. An overview representation to adapt filter parameters based on visual feedback, as well as a navigation possibility to ORFs of interest.

Identification and Development of Contributions:

Contribution 1-4 were identified in discussions with D. Oelke and developed by myself.

Additionally: R. Landstorfer and K. Neuhaus contributed biological background information.

Implementation:

The executable system prototype was implemented in Java by myself, D. Oelke and Daniel Seebacher (student assistant). Data processing was performed by myself with R.

Authorship:

Introduction: R. Landstorfer and K. Neuhaus.

All other sections: The rest of the paper is written by myself. All authors reviewed the paper.

Supervision:

D. Oelke and D.A. Keim supervised the paper project and commented on paper drafts and contributions.

VisExpress - Visual Exploration of Differential Gene Expression Data

Svenja Simon, S. Mittelstädt, BC Kwon, A. Stoffel, R. Landstorfer, K. Neuhaus, A. Mühlig, S. Scherer, and D.A. Keim. “*VisExpress - Visual Exploration of Differential Gene Expression Data.*” Information Visualization, 1-26, DOI: [10.1177/1473871615612883](https://doi.org/10.1177/1473871615612883), 2015. [13]

The main research challenge, allowing a quality aware visual exploration of differential gene expression data for expert users, was identified by myself. **The contributions** of this paper are:

1. Problem characterization and abstraction of tasks & data for the topic “visual exploration of differential gene expression data”.
2. The validated visualization design of *VisExpress*, based on an overview to detail visualization approach and *gene fingerprints* to explore differential gene expression data.
 - 2a. Final design and validation of *VisExpress*.
 - 2b. Colormap design for the *gene fingerprint* designs “Stacked” and “2D colormap” matrix.
 - 2c. Optimization of the recursive pattern layout.
3. A pair analytics study to validate the design of *VisExpress*.
4. A discussion of the resulting biological findings.

Identification and Development of Contributions:

Contributions 1 and 4 were identified and developed by myself. The co-authors R. Landstorfer, K. Neuhaus and A. Mühlig commented on the corresponding paper parts from a biological view. Contribution 2 was identified by myself and developed by myself and the co-authors S. Mittelstädt, BC Kwon and A. Stoffel as a VIS team (see [Design Process](#) (p.85) for further information).

Contribution 2a was developed by myself.

Contributions 2b and 2c were identified and developed by S. Mittelstädt.

Contribution 3. BC Kwon had the idea to validate the system design with a pair analytics study [[Arias-Hernandez et al., 2011](#)] and commented on the study design. I designed the study myself and performed the study with R. Landstorfer, K. Neuhaus and A. Mühlig.

Implementation:

The executable system prototype was implemented in Java by S. Mittelstädt. Data processioning was performed by myself with R.

Authorship:

Colormap design in [Stacked matrix](#) (p. 91) and [2D colormap matrix](#) (p. 92): S. Mittelstädt.

[Optimization details of the recursive pattern layout](#) (p. 97): S. Mittelstädt.

All other sections: The rest of the paper is written by myself. All authors reviewed the paper.

Supervision:

BC Kwon, A. Stoffel, D.A. Keim and S. Scherer supervised the paper project and commented on paper drafts and contributions.

I. INTRODUCTION

Co-Authored Publications

Visual Boosting in Pixel-based Visualizations

Daniela Oelke, Halldór Janetzko, Svenja Simon, Klaus Neuhaus, and Daniel A. Keim. “*Visual Boosting in Pixel-based Visualizations*.” Computer Graphics Forum, 30(3):871-880, DOI: [10.1111/j.1467-8659.2011.01936.x](https://doi.org/10.1111/j.1467-8659.2011.01936.x), 2011. [10]

The main research idea, addressing the question how to boost interesting and important information in pixel-based visualizations and providing a guideline, was identified by Daniela Oelke. I co-authored this paper and contributed the following:

- a) Discussions about influencing factors for the effectiveness of boosting techniques.
- b) The idea for the distinction between image- and data-driven boosting as an influencing factor.
- c) Discussions about the effectiveness of boosting techniques based on influencing factors, leading to a guideline.
- d) Biological application scenario.
- e) Text for the biological application scenario and review of the paper.
- f) The executable prototype was implemented in Java. The prototype was based on a pixel-based visualization implementation of Daniela Oelke. Halldór Janetzko and myself took over the responsibility to implement a few of the suggested boosting techniques within the implementation of Daniela Oelke.

Solving Problems with Visual Analytics: Challenges and Applications

Daniel A. Keim, Leishi Zhang, Miloš Krstajić, and Svenja Simon. “*Solving Problems with Visual Analytics: Challenges and Applications*.” Journal of Multimedia Processing and Technologies, Special Issue on Theory and Application of Visual Analytics, 3(1):1-11, 2012. [8]

The main research idea, addressing the challenges and applications in visual analytics, was stated by Daniel A. Keim. I co-authored this paper and contributed the following parts:

- a) Application example in the area of Next-Generation-Sequencing data analysis.
- b) Text for the biological application.

I-7 Further Publications

During my PhD I authored or co-authored the following publications which are not part of this thesis. See also also Chapter [VI Applications](#) (p. 118).

Peer-reviewed Publications

[9]: R. Landstorfer, [Svenja Simon](#), S. Schober, D. A. Keim, S. Scherer and K. Neuhaus. “*Comparison of strand-specific transcriptomes of enterohemorrhagic Escherichia coli O157:H7 EDL933 (EHEC) under eleven different environmental conditions including radish sprouts and cattle feces.*” BMC Genomics, 15(1):353, DOI: [10.1186/1471-2164-15-353](#), 2014.

[7]: L. Fellner, N. Bechtel, M. A. Witting, [Svenja Simon](#), P. Schmitt-Kopplin, D. A. Keim, S. Scherer and K. Neuhaus. “*Phenotype of htgA (mbiA), a recently evolved orphan gene of Escherichia coli and Shigella, completely overlapping in antisense to yaaW.*” FEMS Microbiology Letters, 350(1):57–64, DOI: [10.1111/1574-6968.12288](#), 2014.

[4]: F. Benites, [Svenja Simon](#) and E. Sapozhnikova. “*Mining Rare Associations between Biological Ontologies.*” PLoS ONE, Public Library of Science, 9(1):e84475, DOI: [10.1371/journal.pone.0084475](#), 2014.

[3]: M. Behrisch, J. Davey, [Svenja Simon](#), T. Schreck, D. A. Keim and J. Kohlhammer. “*Visual Comparison of Orderings and Rankings.*” EuroVis Workshop on Visual Analytics, The Eurographics Association, DOI: [10.2312/PE.EuroVAST.EuroVA13.007-011](#), 2013.

[11]: [Svenja Simon](#), R. Guthke, T. Kamradt and O. Frey. “*Multivariate analysis of flow cytometric data using decision trees.*” Frontiers in Microbiology, 3(00114), DOI: [10.3389/fmicb.2012.00114](#), 2012.

I. INTRODUCTION

Poster and Other Publications

[6]: M. El Assady, D. Hafner, M. Hund, A. Jäger, W. Jentner, C. Rohrdantz, F. Fischer, Svenja Simon, T. Schreck and D. A. Keim. “*Visual Analytics for the Prediction of Movie Rating and Box Office Performance.*” VAST Challenge 2013 - Award for Effective Analytics, 2013.

[2]: F. Al-Masoudi, D. Seebacher, M. Schreiner, M. Stein, C. Rohrdantz, F. Fischer, Svenja Simon, T. Schreck and D. A. Keim. “*Similarity-Driven Visual-Interactive Prediction of Movie Ratings and Box Office Results.*” VAST Challenge 2013 - Award for Effective Visualization, 2013.

[5]: M. Chen, J. Heinrich, J. Kennedy, A. Kerren, F. Schreiber, Svenja Simon, C. Stolte, C. Vehlow, M. Westenberg and B. Wong. “*Uncertainty Visualization.*” Chapter in Biological Data Visualization (Dagstuhl Seminar 12372). Dagstuhl Reports, Volume 2, Issue 9, Chapter 4.6, pages 154-155. Editors: Carsten Görg and Lawrence Hunter and Jessie Kennedy and Sean O’Donoghue and Jarke J. van Wijk, DOI: [10.4230/DagRep.2.9.131](https://doi.org/10.4230/DagRep.2.9.131), 2013.

[1]: J. Aerts, J.-F. Fontaine, M. Lappe, R. Machiraju, C. Nielsen, A. Schafferhans, Svenja Simon, M. O. Ward and J. J. van Wijk. “*Sequence Data Visualization.*” Chapter in Biological Data Visualization (Dagstuhl Seminar 12372). Dagstuhl Reports, Volume 2, Issue 9, Chapter 4.2, pages 143-148. Editors: Carsten Görg and Lawrence Hunter and Jessie Kennedy and Sean O’Donoghue and Jarke J. van Wijk, DOI: [10.4230/DagRep.2.9.131](https://doi.org/10.4230/DagRep.2.9.131), 2013.

[16]: Svenja Simon, D. Oelke, K. Neuhaus and D. A. Keim. “*Visualization of the sensitivity of BLAST to changes in the parameter settings.*” Poster at GCB 2012 - German Conference on Bioinformatics 2012, Jena, Germany (Poster), 2012.

[15]: Svenja Simon, D. Oelke, R. Landstorfer, K. Neuhaus and D. A. Keim. “*Visual Analysis of RNAseq Data to Detect Overlapping Genes in Bacterial Genomes.*” Poster, VIZBI 2012, Heidelberg, Germany, 2012

Chapter II

Bridging the Gap to Domain Experts: The *Liaison* Role for Problem-Driven Visualization Research

Note

This chapter is based on the following publication and parts of this chapter appeared in the following publication [12]¹:

[12]: [Svenja Simon](#), [Sebastian Mittelstädt](#), [Daniel A. Keim](#), and [Michael Sedlmair](#). “*Bridging the Gap of Domain and Visualization Experts with a Liaison*.” Eurographics Conference on Visualization (EuroVis) - Short Papers, Cagliari, Italy, 25 - 29 May 2015, 127-133, The Eurographics Association, DOI: [10.2312/eurovisshort.20151137](https://doi.org/10.2312/eurovisshort.20151137), 2015.²

Please note that I will use “we” throughout this chapter instead of “I”, as this chapter is based on a publication¹. “I” will only be used to refer to my role as a *Liaison*.

¹For the division of responsibilities and work, as well as a statement of contributions in this publication, see [Bridging the Gap of Domain and Visualization Experts with a Liaison](#) (p. 9).

²I own (with the co-authors) the copyright of this publication. EUROGRAPHICS holds the exclusive license for publishing ([12]). The definitive version is available at <http://diglib.org/>
Direct link to the published article: <http://diglib.org/handle/10.2312/eurovisshort.20151137.127-131>

II-1 Introduction

In the last chapter I introduced challenges and opportunities of problem-driven research in the application area of molecular biology. One issue is the collaboration with domain experts which is essential for a design study [[Sedlmair et al., 2012b](#)]. Effective collaboration is heavily based on communication. However, often a large knowledge gap between domain and visualization experts exist and, thus, a missing common language and understanding often hampers an effective communication ([Bridging the gap*](#) challenge).

This *knowledge gap* is especially high in exploratory data analysis and visualization projects. First, tackled problems in visualization research are often *ill-defined* and even domain ex-

II. BRIDGING THE GAP TO DOMAIN EXPERTS - THE *LIAISON* ROLE

perts cannot clearly define their tasks, as they 'just' want to generate new insight and to advance the state of the art [van Wijk, 2006]. Secondly, problems are *inherently complex* and need a human in the loop to integrate implicit domain knowledge in the analysis process. In application domains, such as genomics [Meyer et al., 2009, Meyer et al., 2010b], security applications [Mittelstädt et al., 2015b], or automotive engineering [Sedlmair et al., 2011, Piringer et al., 2010] the *knowledge gap* to visualization researchers is especially high and additionally patterns of thinking and strategies for solving problems differ significantly. This might lead to difficulties and impede the work of visualization researchers identifying the needs and understanding domain experts. This *knowledge gap* hampers an effective communication, leading to an interdisciplinary communication issue.

Due to the specifics in exploratory data analysis and visualization projects, methods from Software Engineering (e.g., Requirement Analysis [Grady, 2013] and Human-Computer Interaction (e.g. in User-centered Design [Vredenburg et al., 2002]) do not sufficiently address the interdisciplinary communication issue for visualization research. Despite the issue for problem-driven research, visualization literature has focused little on communication processes so far.

In this chapter,

- we describe the concept of a *Liaison* role as one approach to foster a better and richer interdisciplinary communication.
- we provide a simple model that can be used to reason and understand the interdisciplinary communication issue.
- we characterize the *Liaison* and how different variations of this role can be utilized in problem-driven visualization research.

The idea for the *Liaison* is based on our own experience from several different design studies where we implicitly used this role. For illustration of benefits, characteristics, and potential limitations of the *Liaison*, we will refer to the *VisExpress* project [13], in which we have first explicitly utilized this role (see also Chapter V).

II-2 Related Work

The HCI community has spent a considerable amount of work on better understanding how to include users into design processes (e.g., User-Centered Design [Vredenburg et al., 2002]). Participatory Design [Spinuzzi, 2005] goes even further as users actively participate in the design process. For participatory design and co-design [Albinsson et al., 2007] also the term *liaison* is used. However, a clear definition is missing. A *liaison* in these areas usually refers to domain experts involved in the design process or to a person who gives technical support to

target users. In contrast, we characterize the *Liaison* for problem-driven visualization projects as a role that abstracts domain problems for visualization experts but do not involve domain experts actively in the design process. In the visualization community, Sedlmair *et al.* specified roles in their Design Study Methodology framework [Sedlmair *et al.*, 2012b]. Their *translator* is similar to our *Liaison* but has been merely mentioned and not been characterized. We decided to use the term “*Liaison*” to strengthen the cooperation and mediation aspect.

Independent of the kind of – broadly speaking – software design a common understanding is needed. The higher the *knowledge gap* to the problem domain, the more common understanding is needed. Bratteteig discussed mutual learning [Bratteteig, 1997] in this respect. For visualization projects, Lloyd & Dykes proposed to use lectures to introduce visualizations to domain experts and domain presentations for the visualization expert [Lloyd and Dykes, 2011]. Kirby & Meyer give recommendations for successful visualization collaborations [Kirby and Meyer, 2013] and suggest learning the domain expert language. The use of the domain language and the associated domain understanding supports to capture the mental model and thereby to build intuitive visualization systems [Kirby and Meyer, 2013]. Gaining domain knowledge and learning the domain language is one way to become a *Liaison* (see Section II-4).

In the visualization literature guidance for the visualization design and evaluation process is given by a number of frameworks, models and methods. Sedlmair *et al.* provided a nine stage framework for design studies in order to structure the visualization process. Furthermore, they identified common pitfalls not only in the design process itself but also in the precondition phase of a design study [Sedlmair *et al.*, 2012b]. Meyer *et al.* proposed the nested blocks and guidelines model for design and validation of visualization systems [Meyer *et al.*, 2013]. McKenna *et al.* provided a design activity framework to break down each activity of design & evaluation in *motivation*, *outcome* and *methods* [McKenna *et al.*, 2014]. In order to capture reasoning processes Arias-Hernandez *et al.* introduced Pair Analytics [Arias-Hernandez *et al.*, 2011].

II-3 The Interdisciplinary Communication Issue

For illustration of the issues of interdisciplinary communication we propose a simple model based on a metaphor of spaces (see Fig. II.1)¹. The domain expert/s span a *Problem Space* which comprise domain problems composed of *facets* such as domain goal, tasks, data, and constraints. The visualization expert/s (short VIS team), on the other hand, span a *Design Space* of visual solutions composed of visual analysis tasks & data abstractions, visual encoding & interaction techniques, and algorithms. Addressing a domain problem implies that all its facets

¹Michael Sedlmair had the idea to illustrate the interdisciplinary communication with a simple model, based on a metaphor of spaces. I developed this idea and designed the graphic in Figure II.1.

II. BRIDGING THE GAP TO DOMAIN EXPERTS - THE *LIAISON* ROLE

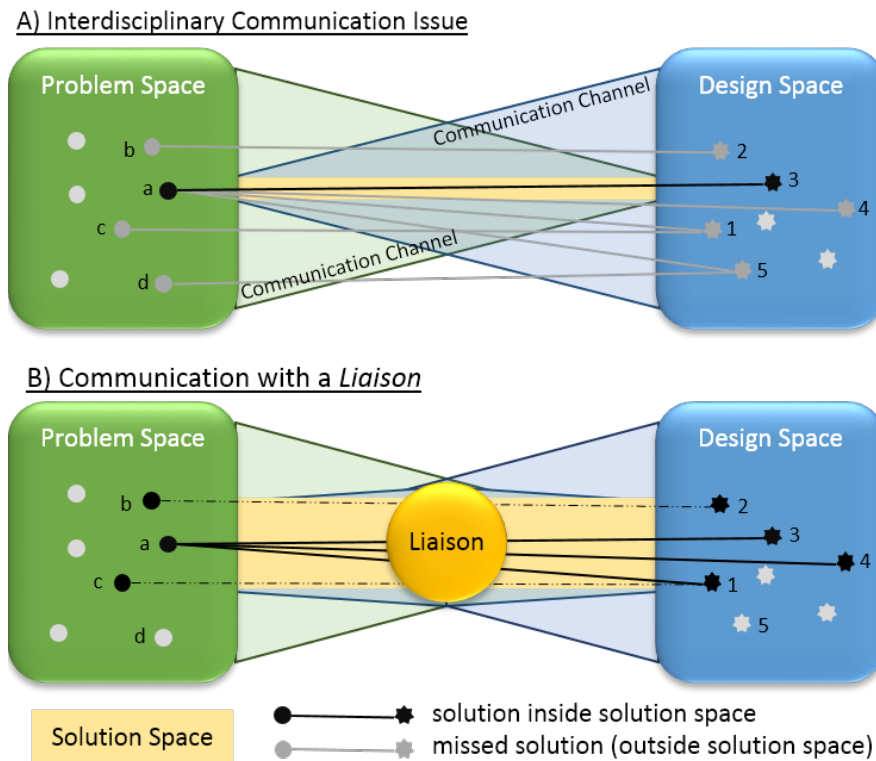


Figure II.1: The *Problem Space* comprises all domain problems and the *Design Space* all visual solutions. (A) Without a common language the domain and visualization experts communication builds a cone, leading to a small *Solution Space*. Thus, many possible solutions are missed (gray lines). (B) A *Liaison* mediates between domain and visualization experts to widen the *Solution Space*, which covers more possible solutions (1,3,4) for (a) and allows the identification of additional interesting domain problems (b, c). This graphic appeared in [12].

need to be understood at first which requires large domain knowledge. The design of a visual solution (indicated by lines in our model) requires that different design choices need to be considered that match problem abstractions and techniques to domain problems and tasks. Thus, a good solution requires both, a large domain and a large visualization knowledge. Otherwise, solutions can be composed of bad design choices and do not solve the domain problem.

Ideally one person covers both knowledge spaces but the issue of problem driven research is that rarely one person has a grounded knowledge in two domains. Thus, typically a domain and a VIS team work together and communicate to connect the knowledge of both spaces with the aim to capture all design alternatives (solution lines) for a domain problem. Without a common understanding both communication endeavors build a cone resulting in a restricted overlap and common understanding (see Figure II.1 A). Thus, just a small part of the solution lines are contained in the *Solution Space* leading to potentially sub-optimal solutions. We denote this

issue as the interdisciplinary communication issue and suggest the *Liaison* role as a solution to broaden the communication channel and *Solution Space* (see Figure II.1 B).

II-4 The *Liaison* Role

The goal of the *Liaison* is to overcome the interdisciplinary communication issue. A *Liaison* shares knowledge and language with both domains for mediating between domain and visualization experts. This establishes a common understanding and greater coverage of the *Problem* and *Design Space* resulting in a larger *Solution Space* and, thus, a better yield of good solutions (see Figure II.1 B). The *Liaison* grasps information of the domain experts and interprets, selects and processes these for the VIS team. Therefore, the *Liaison* needs knowledge from both domains. In particular, the *Liaison* needs the domain language to allow a free speech and collaborative analysis with domain experts (see benefits, Section II-5). Even though a grounded visualization knowledge and language is beneficial, a basic understanding is sufficient. The VIS team can compensate this missing knowledge, whereas a certain domain knowledge is essential to bridge the knowledge gap.

II-4.1 How to Become a *Liaison*.

There are three general ways to become a *Liaison* (see Fig. II.2) which have been used implicitly but not been reported explicitly yet. First, starting as a domain expert interested in visualization, e.g. in [Mittelstädt et al., 2015b] (domain *Liaison*); second, starting as a visualization expert who gathered much knowledge in an application domain during a design study, e.g. in [Sedlmair et al., 2011, Sedlmair et al., 2012a] (visualization *Liaison*) and, third, inherently starting from an interdisciplinary subject, such as, bio-, geo-, or business-informatics, e.g. in [12] (interdisciplinary *Liaison*). All three types have different advantages and disadvantages.

The domain knowledge of a **visualization *Liaison*** might not be sufficient to master the problem complexity, as gaining domain knowledge requires much time. Staying in one application domain is, therefore, advisable. The benefit of this *Liaison* is that the grounded visualization knowledge might allow a smaller VIS team. In order to broaden the *Problem Space* and to ensure that solutions match the domain problem, joint meetings with domain experts and the VIS team are recommended. Such meetings also address the issue of focusing just on a research contribution and not on solving the domain problem.

The other extreme is the **domain *Liaison*** who might have problems to identify an interesting visualization problem, due to a small visualization knowledge. However, this *Liaison* is effective in capturing the problem complexity and in validating design alternatives of the VIS team since

II. BRIDGING THE GAP TO DOMAIN EXPERTS - THE *LIAISON* ROLE

she focuses on a practical solution. A close collaboration with a strong VIS team is advisable who can focus on technical novelty.

The **interdisciplinary *Liaison*** has grounded knowledge in both domains, which makes her more effective in problem and task abstractions than the other *Liaison* types. The prevalence of further advantages and disadvantages depends on the current focus of the interdisciplinary *Liaison*. The interdisciplinary background is a strong advantage since the interdisciplinary *Liaison* can contribute interdisciplinary methods to improve data and analytical grounding for visualizations. Even though, an interdisciplinary *Liaison* might rarely be at hand, interdisciplinary researchers might be interested to join a project as *Liaison* and would be willing to learn more about visualizations.

II-4.2 Instantiations of the *Liaison* Role and the VIS Team

Both *Liaison* and VIS team are roles and can be instantiated in different ways. The minimal team would be a two-man-show; the *Liaison* and one visualization colleague. However, with this team instantiation the *Design Space* will be small and suboptimal-solutions are probable. A senior visualization supervisor (as VIS team) might compensate for this issue and span a “broad-enough” *Design Space*. Even though we recommend a VIS team (several visualization experts) to ensure a broad *Design Space* and to design a visual solution. Prototyping, tool-building and paper writing can be done by one or more members of the VIS team. In any instantiation the *Liaison* works closely with the VIS team. Figure II.2 defines the tasks both roles have to perform in each design study step.

For completion of the design study team, domain experts are essential. How the work is distributed and organized can differ between projects. However, as the *Liaison* is proposed to address the interdisciplinary communication issue, we assume a knowledge gap between domain and visualization experts. Therefore, joint meetings are often only effective for high level discussions. The *Liaison* can help here to avoid misunderstandings due to different usage of terms or wrong presumptions on both sides.

An engagement of visualization and domain experts in mutual learning to establish a common understanding is sometimes performed and has advantages and disadvantages. First this needs a lot of time on both sides and visualization experts might run in the same pitfall as a *Liaison*, that the **Awareness of the problem complexity contradicts with a practical solution**, as the VIS team is not independent. Second, domain experts with grounded visualization knowledge might mistake visualization researchers as tool smiths, for instance, by stating explicit requests leaving no room for design alternatives.

II-5 Benefits and Tasks of the *Liaison* and the VIS Team

On the other hand side, advantages are that misunderstandings can be resolved in a direct communication (no [Lost in translation](#)) and that the appropriateness of ideas can be judged by all visualization experts (avoiding [A Liaison may suppress ideas](#)). Additionally, a grounded visualization knowledge of domain experts can also have the potential to build highly tailored and well adapted systems. Design study projects with a close collaboration and a mutual learning between domains are close to participatory design. Due to a reduced or closed knowledge gap in such studies a *Liaison* is not necessary.

However, a further point to consider is the possible difference between a common language and a domain language. A common language is less rich and limited in expressiveness hampering the capturing of the mental model. Kirby and Meyer argue, therefore, that visualization experts should learn the domain language instead of establishing a common language [[Kirby and Meyer, 2013](#)].

II-5 Benefits and Tasks of the *Liaison* and the VIS Team

Hereinafter we will present the *VisExpress*-project to exemplify the application of the *Liaison* role. Further on, we will discuss the benefits of the *Liaison* for the design study process according to concrete tasks (see Fig. [II.2](#)).

II-5.1 Example Project with a *Liaison*- The *VisExpress*-Project

The *VisExpress*-project is a design study with the goal to identify “interesting genes” in a vast amount of biological data (see Chapter [V](#)). More precisely this is a high level aim with ill-defined tasks. Biologists first requested to inspect genes with potential quality issues. The VIS team abstracted tasks & data and came to the conclusion that the problem is related to time series analysis with interactive filters (exclude genes without potential quality issues). This allows to efficiently handle quality issues and to reduce the amount of data for the analysis. A standard visualization solution with small multiple line charts was sufficient for this problem and task abstraction (see Figure [III.16 I](#)). When the solution was deployed, the VIS team identified that the design was intuitive to the domain experts and quality aware analysis could be performed, however, it seemed that the solution did not meet their expectation.

Due to the *interdisciplinary communication issue* it was hard for the VIS team to understand their problems. As a visualization Phd student with a major in bioinformatics, I identified the issues with the problem characterization based on the prototype. The full complexity of the problem was not captured in the first problem characterization. Indeed the domain experts

II. BRIDGING THE GAP TO DOMAIN EXPERTS - THE *LIAISON* ROLE

<p style="text-align: center;"><u>Am I a <i>Liaison</i>?</u></p> <ul style="list-style-type: none"> • Do you have a common understanding with your domain experts? • Can you speak with your domain experts in their language? • Can you abstract and canalize information from their domain? • Can you accomplish the following tasks? 		
General Steps	VIS team	<i>Liaison</i>
Domain Problem Characterization	<ul style="list-style-type: none"> • discuss promising domain problems 	<ul style="list-style-type: none"> • select promising domain problems • characterize domain problem
		<ul style="list-style-type: none"> • capture mental model of domain experts • abstract domain problem
Abstraction	<ul style="list-style-type: none"> • abstract data & tasks in visualization terms 	<ul style="list-style-type: none"> • ensure validity of abstraction with respect to the problem characterization
Design	<ul style="list-style-type: none"> • design visual encodings & interactions • span <i>Design Space</i> 	<ul style="list-style-type: none"> • map mental model with design • canalize <i>Design Space</i> • span <i>Solution Space</i>
Evaluation		<ul style="list-style-type: none"> • clarify domain tasks further • test the fit of the mental model • clarify feature extensions and usability • capture reasoning processes
Reflection	<ul style="list-style-type: none"> • formulate design guidelines 	<ul style="list-style-type: none"> • reflect human cognition, reasoning processes and knowledge generation

Figure II.2: Short test “Am I a *Liaison*?” and list of the *Liaison* and VIS team tasks in each design process step. This graphic appeared in [12].

needed a quality aware data exploration system to detect patterns in a vast amount of data. Handling data quality issues was just one aspect of this problem.

My experiences during my doctoral studies and especially in the course of the *VisExpress*-project led to the idea of the *Liaison* role. With a major in bioinformatics I acted as an interdisciplinary *Liaison* in the *VisExpress*-project. I was supported by a VIS team of three colleagues also working in the field of visualization, however, as a visualization PhD student I acted also as part of the VIS team. In this case the team has to be aware of role conflicts (see Section II-6). The revised problem characterization led to the complex visual exploration system *VisExpress* (Figure III.16 II and Chapter V). Here Gene-fingerprint matrices replaced the line charts, by representing all pair-wise time series ratios as well as their quality. Using the gene-fingerprints, a three levels architecture from overview (a) to data view (b) and detailed view (d) was designed to support data exploration and pattern detection. Hereinafter we will elaborate on lessons learned from utilizing the *Liaison* role in the *VisExpress*-project.

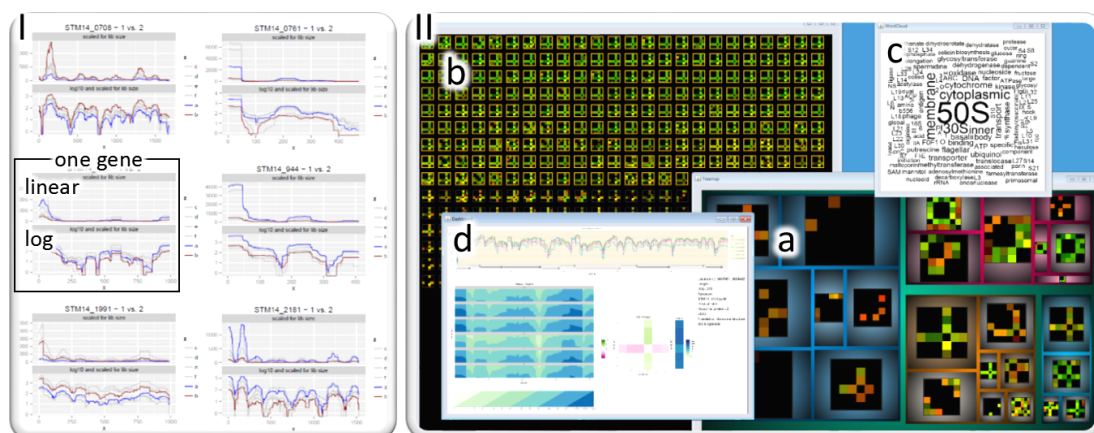


Figure II.3: Visualization approaches to visualize gene expression data. I) discarded prototype. II) final *VisExpress*-system. This graphic appeared in [12].

II-5.2 Tasks & Benefits

We will describe tasks and benefits of a *Liaison* and how this role can help to mitigate known pitfalls (PF) in the design process of problem-driven visualization projects [Sedlmair et al., 2012b] (ordered by their occurrence in Fig. II.2).

Capturing the problem complexity. Even though methods like, e.g., contextual inquiries [Beyer and Holtzblatt, 1997] work well, speaking the domain language and knowledge in the domain lead to a better problem understanding. Furthermore, also unspoken information can be captured and the risk to overlook things is minimized with a *Liaison*.

Capturing the mental model. In order to support insight generation, matching the mental model of the target users is one of the biggest challenges in visual design to allow the generation of insights [Yi et al., 2008]. However, capturing the mental model is challenging and requires a deep domain understanding. For the *Liaison* it is easier to capture the mental model since the *Liaison* can build on domain understanding and intensive discussions with domain experts in their language.

Faster and richer abstraction. A *Liaison* can avoid the pitfall to abstract too little (PF-19 in [Sedlmair et al., 2012b]) or erroneous. Despite the pitfall of capturing only parts of the problem, we observed in the *VisExpress*-project that the VIS team tended to concentrate on an interesting visualization problem, thereby changing the focus which did not match the domain problem. Thus, a *Liaison* is needed to ensure that task and data abstractions still meet the domain problem.

II. BRIDGING THE GAP TO DOMAIN EXPERTS - THE *LIAISON* ROLE

Design validation. Another common pitfall is observing a *Design Space* which is too small (PF-20). Here the independent VIS team ensures to span a broad *Design Space*. Without direct contact to domain experts the VIS team is independent and, thus, not biased by detailed domain issues that may hamper the development of ideas. Here, several persons are helpful to avoid a related pitfall which is to assume that the own latest visualization technique is a right match (PF-21). The *Liaison* canalizes the *Design Space* to balance design alternative against their fitting of the mental model.

Expressive and valuable evaluation. Evaluation issues are often artificial usage scenarios without real data & tasks (PF- 24) and little expressive statements like “The domain experts liked the tool.” (PF-26). The reasons are a missing grounded problem understanding and a layperson’s language. In contrast, the *Liaison* can speak the domain language and can act as a real analysis partner in a collaborative analysis with real data and tasks. Such an evaluation allows the *Liaison* to deeply discuss and assess findings during the study, leading to a clarification of tasks and usability issues. Feature requests can be captured between the lines in the domain language. In the *VisExpress*-project one statement was, e.g.: “I would like to order the genes of one cluster in synteny to look for operons”. The *Liaison* understood that the aim was to arrange genes sequentially to identify neighboring genes with the same pattern.

Furthermore, we see high potential for a *Liaison* in Pair Analytics where the goal is to capture users reasoning processes during collaborative analysis [[Arias-Hernandez et al., 2011](#)].

II-6 Discussion and Limitations

Awareness of the problem complexity contradicts with a practical solution. A deep understanding of the problem domain regularly brings up new issues which contradict with the current solution direction (PF-18 in [[Sedlmair et al., 2012b](#)] - learning too much). This can make it harder for the *Liaison* to narrow down to a self-contained but still meaningful and essential visualization problem. Therefore, a consultation of the VIS team for the selection of a promising domain problem is important in the problem characterization phase.

A *Liaison* may suppress ideas. There is a risk that the *Liaison* might over-criticize ideas of VIS team members, especially if the *Liaison* person is also part of the VIS team. In brainstorming the *Liaison* can, e.g., easily use the domain knowledge and language for supporting own ideas. Therefore, we suggest to first discuss the ideas of the VIS team. In this round the *Liaison* contributes no own ideas but objectively comments on the ideas of the VIS team

members. In the next step the *Liaison* contributes own ideas. All solutions are then presented, merged, refined or rejected in a discussion phase with the whole team.

Lost in translation. The *Liaison* reduces the direct communication between domain and visualization experts in a design study. Therefore, the success is highly dependent on the quality of the *Liaison*. Misinterpretations of domain problems, domain expert comments and study findings can lead to failed projects. In order to reduce these issues we recommend to discuss all interpretations with the domain experts to check their validity.

Domain Drift of the *Liaison* We argue that a more grounded knowledge is needed in the application domain compared to visualization knowledge. The critical point in problem-driven research is to really solve the addressed domain problem and to design a system which is adopted by domain experts. This often requires to deeply understand the domain problem, the context and high-level domain goals. Discussions with domain experts and observations of those are needed to achieve these goals. Therefore, a domain understanding and at least a common language is needed. Preferably is even the possibility to communicate in the domain language and to act as an analysis partner. If, however, a domain expert can take over the task to abstract domain problems, a less comprehensive domain knowledge would be needed for the *Liaison*. Sedlmair *et al.* specify such a domain expert as a *translator* in their Design Study Methodology framework [Sedlmair *et al.*, 2012b]. Thus, a missing grounded domain knowledge, needs to be compensated with a second dedicated role, a domain expert acting as a translator. In contrast, a VIS team is a general part of a design study team, allowing to compensate a missing grounded visualization knowledge of the *Liaison* much easier.

Business analyst vs. *Liaison* A *Liaison* is similar to a consultant or business analyst. Consultants and business analysts can have diverse backgrounds, for instance, in computer science, design, psychology, business or even social science. In software projects these experts are often involved in requirement analysis and specification, as well as in negotiations of deliverables and operating plans. Consultants and business analysts are experts in analyzing and abstracting workflows and requirements. Their work effort often spans several month and includes a certain learning of the application domain and necessitates a basic knowledge of the technical feasibility. In this way a consultant or business analyst can be seen as a *Liaison*, as they become one by the learning during the project. The difference is more in the area of deployment. Due to the costs of such experienced experts, they are mostly deployed for large scale software development projects only. Even though exceptions exists (for instance in [Mittelstädt *et al.*, 2015b]), dedicated consultants or business analysts are, therefore, rarely deployed in research projects.

II. BRIDGING THE GAP TO DOMAIN EXPERTS - THE *LIAISON* ROLE

Nevertheless, in my opinion, methods from business analysis, software engineering and human computer interaction provide valuable resources for visualization research projects.

Chapter III

Requirement Analysis and Problem Abstraction

Note

This chapter is partly based on biological background information parts of the following two publications and parts of this chapter appeared or will appear in these publications [14, 13]¹.

However, Section [Problem Abstraction \(III-4\)](#) is (in this form) a new contribution made through this thesis. Subsection [III-4.2](#) was formulated retrospectively and is not directly based on publication [14]. Subsection [III-4.3](#) is based on publication [13] and parts of this subsection will appear in publication [13].

[14]: [Svenja Simon](#), Daniela Oelke, Richard Landstorfer, Klaus Neuhaus, and Daniel A. Keim. “*Visual Analysis of Next-Generation Sequencing Data to Detect Overlapping Genes in Bacterial Genomes.*” 2011 IEEE Symposium on Biological Data Visualization, October 23 - 24, Providence, Rhode Island, USA, 47-54, IEEE, DOI: [10.1109/BioVis.2011.6094047](https://doi.org/10.1109/BioVis.2011.6094047), 2011.²

[13]: [Svenja Simon](#), Sebastian Mittelstädt, BC Kwon, Andread Stoffel, Richard Landstorfer, Klaus Neuhaus, Anna Mühlig, Siegfried Scherer, and Daniel A. Keim. “*VisExpress - Visual Exploration of Differential Gene Expression Data.*” Information Visualization, 1-26, DOI: [10.1177/1473871615612883](https://doi.org/10.1177/1473871615612883), Published online before print December 14, 2015.³

¹For the division of responsibilities and work, as well as a statement of contributions in these publications, see [Visual Analysis of Next-Generation Sequencing Data to Detect Overlapping Genes in Bacterial Genomes](#) (p. 10) and [VisExpress - Visual Exploration of Differential Gene Expression Data](#) (p. 11).

²The Institute of Electrical and Electronics Engineers (IEEE) is the copyright owner of this work [14] but, as an author, I am permitted to re-use the work of this publication (verbatim and derivative) for my personal use. Link to the published article in IEEE Xplore.: <http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6094047>

³I own (with the co-authors) the copyright of this publication. The SAGE Publications Ltd holds the sole and exclusive right and license for publishing ([13]). The definitive version is available at <http://ivi.sagepub.com/> Direct link to the published article: <http://dx.doi.org/10.1177/1473871615612883>

III-1 Introduction

This chapter provides a requirement analysis and problem abstraction for analyzing RNAseq data from bacteria. After an introduction to the biological topic of genes and RNAseq data by next-generation-sequencing (NGS), I introduce the two main research gaps I identified for the (visual) analysis of RNAseq data to discover and describe genes in bacteria. First, assessing the trustworthiness of measurements and second, to discover and relate genes to identify their functions.

The contributions of

- a problem characterization and abstraction for the (visual) analysis of RNAseq data to discover and describe genes in bacteria,
- a corresponding set of tasks and
- a definition of *interestingness* and *uncertainty* to bypass, respectively understand, these challenges for the stated tasks

are the foundation for the two analysis systems described in Chapter IV and Chapter V, which address the identified two research gaps.

The given *problem characterization and abstraction* can also be used by other researchers who might develop alternative systems for the problem of (visual) analysis of RNAseq data to discover and describe genes in bacteria. Sedlmair *et al.* [Sedlmair *et al.*, 2012b] consider a *problem characterization and abstraction* as one of the three contribution types of problem-driven research and argue even to consider it “as a first-class contribution of a design study”¹.

III-2 Biological Background

The genome encodes the genetic information of organisms. The thousands of genes encoded on the genome are the information units - they encode proteins which perform a vast number of functions in cells. The protein hemoglobin, for example, transports oxygen in vertebrates². Depending on environmental conditions, a different composition of proteins is produced. If the oxygen content of the air is low (e.g., in high altitude on a mountain), for instance, more hemoglobin is needed and produced.

Even though the function of many proteins is known, a vast number of protein functions is still unknown. A better comprehension of protein functions and their interplay would facilitate the understanding of diseases and medical treatment and is, therefore, of major interest for biologists and physicians.

¹[Sedlmair *et al.*, 2012b]

²Animals with a vertebral column (also denoted as backbone or spine).

RNAseq by next-generation-sequencing (NGS) is a technology which allows making advancements in this direction. RNAseq is a high-throughput experimental method permitting measuring the 'production' of genes under a certain experimental condition. Thereby, new genes can be identified and functions of proteins that are so far unknown can be inferred.

III-2.1 Open Reading Frames and Genes

The genome consists of DNA which is a double-helix composed of the **nucleotides** adenine (A), cytosine (C), guanine (G) and thymine (T). In the double-helix certain nucleotides are complementary to each other (A-T and C-G) and are connected by hydrogen bridges. This complementary nature of the DNA allows representing the genome as one computer-readable sequence over the alphabet A, C, G, T ($\Sigma = \{A, C, G, T\}$).

Information is encoded by triples of nucleotides, the so-called codons. Codons build the genetic code. In bacteria genes start with a specific start codon and end with specific stop codon, the variable number of codons between start and stop codon encode the genetic information (see Figure III.1). A sequences of codons starting at a specific position is called a reading frame and the sequence between a start and a stop codon is called an **open reading frame (ORF)** .



Figure III.1: The six reading frames of a genome, with one highlighted open reading frame (ORF). The start codon of the ORF is indicated with blue, the stop codon with red.

However, not each ORF is a gene. ORFs can occur by chance, therefore, false gene candidates need to be ruled out to determine the genes of an organism. Genes are ORFs that encode proteins. Proteins carry out specific functions in cells. The protein collagen, for instance, is a main part of the connective tissue and, therefore, collagen has a structural function in cells and influences the strength and elasticity of the skin. As the necessity for many proteins is depended on the current environmental or experimental condition, regulatory elements are needed to control the production of proteins. During aging, for example, the protein collagen is produced less and less leading, among others, to wrinkles.

III. REQUIREMENT ANALYSIS AND PROBLEM ABSTRACTION

If a regulatory element triggers the production of a protein, first the corresponding gene is transcribed to so-called messenger RNA (mRNA) which is basically a copy of the gene. RNA also consists of nucleotides but the chain is built with ribose (RNA) instead of deoxyribose (DNA) and the nucleotide thymine (T) is replaced by uracil (U). In a next step, named translation, the mRNA is translated into the protein. A protein is a chain of amino acids that build a secondary structure to perform a specific function. Thereby, each codon of the mRNA encodes for one amino acid. In total 64 codons exist but only 20 amino acids. The genetic code is, therefore, redundant. Some amino acids are encoded by more than one codon. This assignment is the genetic code. The whole process of transcription and translation is also called gene expression.

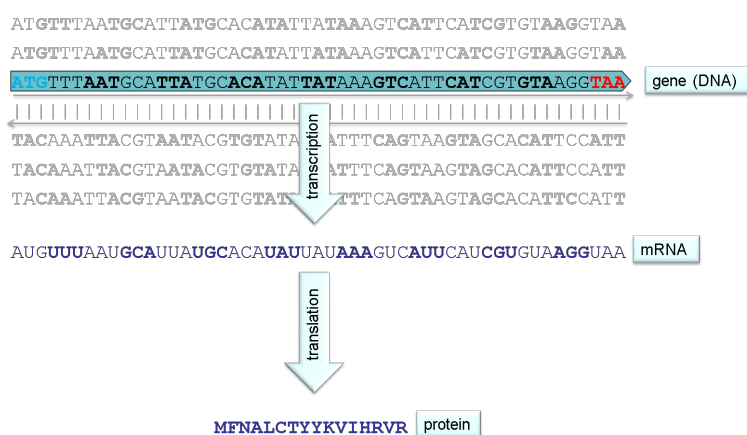


Figure III.2: Gene expression. A gene is first transcribed to a mRNA. Based on the genetic code, each codon of the mRNA is then translated to one amino acid in the protein sequence.

Beside genes, which encode for proteins, further information is encoded on the genome. Regulatory RNA like, for example, antisense RNA and small RNA as well as the specialized RNA types, ribosomal RNA (rRNA) and transport RNA (tRNA). rRNAs build the ribosomes which translate mRNA to proteins. tRNA transports amino acids to the ribosom to synthesize proteins.

Please note that I have simplified the biological background information. It is also important to mention here that the described scenario applies for bacteria only. Higher organisms have more complicated mechanisms and gene candidates cannot be determined in a straight forward way in higher organisms.

Determination and Annotation of Genes

As the structure of genes is clearly defined by start and stop codons in bacteria, ORFs are straightforward to determine. Different methods are applied to determine which of these ORFs

are genes (protein coding). First several computational gene prediction algorithms exist like GLIMMER [Salzberg et al., 1998, Delcher et al., 1999, Delcher et al., 2007], the GeneMark [Borodovsky and McIninch, 1993], [Lukashin and Borodovsky, 1998], [Besemer et al., 2001], [Besemer and Borodovsky, 2005], or Prodigal [Hyatt et al., 2010]. Most of these algorithms are part of comprehensive annotation pipelines. Madupu *et al.* [Madupu et al., 2010] and Angiuoli *et al.* [Angiuoli et al., 2008] provide an overview over several annotation pipelines.

Besides gene prediction itself, which assesses the coding potential of a sequence, predictions of further features (promoter predictions [Jacques et al., 2006, Ranganathan and Bansal, 2007, Wang and Benham, 2006, Ozoline and Deev, 2006, Shavkunov et al., 2009], terminator predictions [de Hoon et al., 2005, Kingsford et al., 2007, Lesnik et al., 2001, Silby et al., 2004] and predictions for translation initiation signals [Hu et al., 2009, Hyatt et al., 2010] and [Saeys et al., 2007]) are used in annotation pipelines. However, often manual curation is applied to get high quality annotations.

A further branch of prediction is based in sequence similarity. Some genes are essential for the survival capability of bacteria and they exist in all bacteria. Based on sequence similarity, which is interpretable as similarity due to a common ancestor, the function of known genes can be transferred to the ORF of a newly sequenced bacteria species. Beside essential genes, many other genes exist in at least one branch of related bacteria species which can be detected by similarity searches. If a function can be assigned, the gene is *annotated* with a function. However, often all ORFs that are considered as genes are tagged as *annotated*, even if no function is annotated based on experimental evidence but just a computational prediction exist.

However, for some genes no related annotated gene exist. In these cases, similarity search is not helpful. These can be genes that have been overlooked in all species so far. Either these sequences do appear not “gene-like”, are of short length, or they overlap with an annotated gene. Furthermore, so called orphan genes exist. Orphan genes exist just in one species and are expected to be important for the adaption to an ecological niche, leading to the emergence of new biological species.

Besides, computational predictions and sequence similarity, experimental methods exist which allow determining genes. The final evidence is given by detecting the existence of the protein encoded by the gene. In order to detect proteins mass spectrometry is applied, however, mass spectrometry is not sensitive enough to detect proteins which are produced in low concentrations. An alternative is to sequence the RNA (indirectly). RNAseq by next-generation-sequencing (NGS) has opened up the possibility to sequence in a high throughput fashion with falling costs. As RNAseq measures all RNA transcripts in a cell, also so far unknown genes can be detected. However, as all RNA transcripts are sequenced, some transcripts might not be protein coding mRNA but regulatory RNA. The transcriptome can be determined to account

III. REQUIREMENT ANALYSIS AND PROBLEM ABSTRACTION

for this. Therefore, active ribosomes are isolated which translate mRNA to proteins. After a digestion of mRNA, just the mRNAs covered by a ribosome remain. This RNA fragments are protein coding.

It is important to note that, beside the advantages of high-throughput sequencing, only active genes can be detected. As many genes are just active under specific conditions, the detection of all genes is still challenging.

Finally annotating a determined gene with a function is a further challenging and time consuming step. For this purpose, for instance, gene knock-outs are used. In order to knock out a gene of interest, for instance, a stop codon can be introduced in the gene sequence. The bacteria without an active version of the gene can then be tested under different conditions and compared to the wild type bacteria to identify the function of the gene. RNAseq is another possibility which can at least give hints for the function of the gene. Therefore, the RNA is measured under different conditions. If the production of a gene differs between two conditions a function related to this condition is likely. Furthermore, the gene can be compared to other genes. If a gene of interest behaves similarly to a gene with a known function, the gene of interest might belong to the same functional category. This analysis is called gene expression analysis or differential gene expression analysis if data from different experimental conditions is compared.

Overlapping Genes

Determining genes can become even more complex since reading frames can overlap. Six reading frames are possible - three in each direction (see Figure III.3). For viruses many cases of [overlapping genes \(OLGs\)](#) are known. In contrast, for bacteria only a few instances are known by now (less than one hundred; in comparison: in bacteria several million genes are known). For examples of overlapping genes, see e.g., [[Wang et al., 1999](#), [Behrens et al., 2002](#), [McVeigh et al., 2000](#)]. Furthermore, a new overlapping gene pair [7] was described in the course of the [FOG-Project](#).

Due to the small number of known overlapping genes in bacteria and information content constraints the consensus in biology is that overlapping genes are an exception in bacteria and exist in viruses just because of a selection pressure owing to space limitations of the viral capsid [[Chirico et al., 2010](#)].

Therefore, gene prediction discards overlapping genes. For each locus, the ORF with the best prediction value is selected. All overlapping ORFs are discarded, even if the prediction value is also high [[Delcher et al., 2007](#)]. As ORFs overlapping a gene are overlooked in the shadow of the annotated genes, they are named shadow ORFs (sORFs).

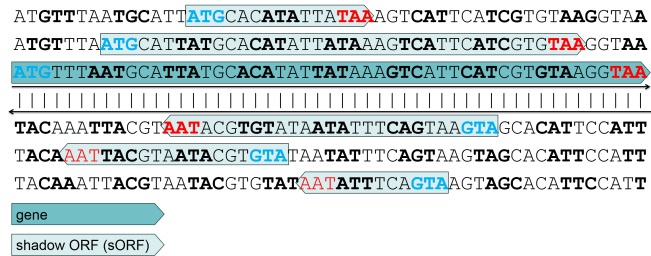


Figure III.3: In frame +1 a gene is annotated. At the same locus five other ORFs exist. As they overlap with the gene they are not annotated and named shadow ORFs (sORFs).

III-2.2 RNAseq Using Next-Generation-Sequencing

Next-generation-sequencing (NGS) is an umbrella term for high throughput sequencing methods which sequence DNA by synthesis. Competing methods of this basic principle have been developed by several companies but the most common used are 454 sequencing (now owned by Roche Diagnostics), Solexa sequencing (now owned by Illumina), and SOLiD sequencing (from life technologies, formerly Applied Biosystems). Common to all is a limited sequencing read length of a few hundred nucleotides. Often reads have just a length from 20-100 nucleotides. These short reads have to be puzzled into genomes (when sequencing *de novo*) or mapped onto an existing genome sequence, e.g., in the case of RNA Sequencing (RNAseq). RNAseq uses NGS to indirectly sequence RNA of cells. This allows to determine which genes are active under a specific condition and with which strength. As RNAseq is faster and not restricted to known genes, RNAseq replaces DNA micro-arrays.

In order to point out biases which are introduced in the experimental pipeline for next-generation-sequencing¹ I explain all steps. RNAseq data suffers from quality issues due to these biases and necessities a visual inspections to assess the trustworthiness (see [Assessment of the Trustworthiness of RNAseq Measurements](#) (p. 47)). Background for the aspects influencing the trustworthiness, needed to be addressed in the visual representation, is given in the following.

Experimental Protocol for Sequencing Library Preparation

The set of experimental procedures for sequencing is called protocol. The final nucleotide sequence fragments of an experiment, ready for the actual sequencing, are called a library. Depending on the NGS method used, the protocols for RNAseq library preparation differ to some extent but the main steps are similar (see also Figure III.4). In order to highlight data quality issues, bias sources are named for all steps.

¹This thesis covers strand specific single read sequencing only. Note that also paired-end sequencing exist which sequences fragments from both sides and links the resulting reads.

III. REQUIREMENT ANALYSIS AND PROBLEM ABSTRACTION

RNA extraction. The total RNA inventory of a sample is extracted. In the case of bacteria, it might originate from around 10^{10} cells.

mRNA enrichment. Next, ribosomal RNA (rRNA) is depleted since it does not code for a protein (see Section III-2.1) but constitute about 90 to 95 % of the cells RNA.

Bias Sources. If depletion of rRNA does not work sufficiently, large amounts of the sequencing capacity are used to sequence rRNA. Thus, the probability to sequence a transcript with a low concentration is reduced.

Fragmentation. Because only short pieces of nucleotide sequences can be sequenced, the remaining RNA is fragmented. Fragmentation can be obtained chemically due to autocatalytic fragmentation, physically by sonication or nebulization, or enzymatically by different RNases.

Bias Sources. In RNAseq data it is often observed that the shapes of a [read coverage](#) above a gene from different experiments are quite similar (given the same experimental protocol). Thus, high and low read coverage is observed at the same position within a gene for different experiments. This was also observed for the data of the [FOG-Project](#) (Fig. III.5). For our data the most likely explanation is a fragmentation bias due to the use of RNase III which is usually used for enzymatic fragmentation. This enzyme cleaves only double stranded RNA, releasing 3' overhangs with 2 to 3 nucleotides. Thus, enzymatic degradation requires the mRNA to fold into secondary structures that are non-random but depend on the actual sequence patterns of a given mRNA [Zuker, 2003]. Actually all fragmentation methods are biased since the local physical property of a nucleotide strand depends on its actual sequence causing the strands to brake non-random at preferential sites. Consequently, for all current fragmentation methods, unequal fragmentation patterns can be expected and are ob-

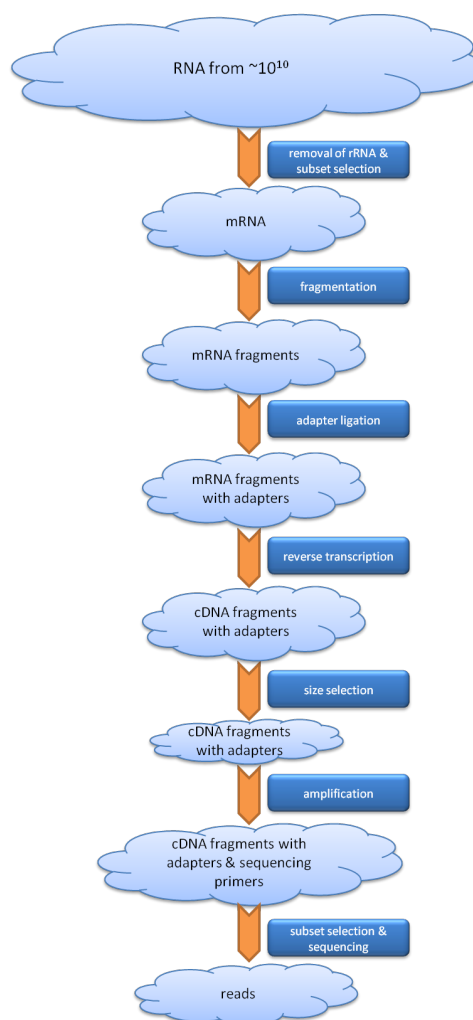


Figure III.4: Steps of the library preparation. From RNA to sequenced reads.

served [Quail et al., 2008, Fisher et al., 2011]. Therefore, it can be summarized that fragmentation is dependent on the actual sequence either directly or indirectly by the secondary structure.

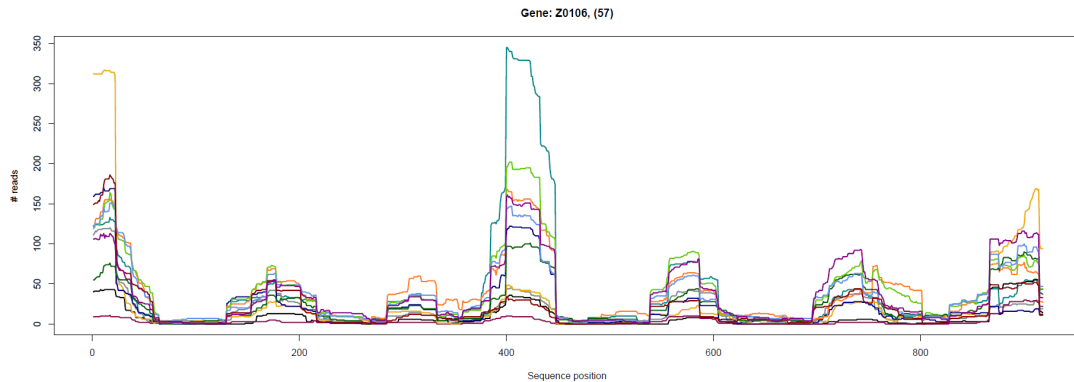


Figure III.5: The line chart shows the read coverage (number of overlapping reads per position) for twelve conditions of one gene. It highlights that peaks and valleys are located at the same positions for all experiment conditions.

Ligation of adapters. The obtained RNA fragments are further ligated with adapters. For strand specific sequencing, two different adapters are ligated to the 5'- and 3'-end strand specifically. These adapters are necessary for, first, the reverse transcription, and later the sequencing.

Bias Sources. The efficiency of adapter ligation might also be dependent on the sequence.

Reverse transcription. After adapter binding, the RNA is reverse transcribed, starting at the 3'-adapter, to copy DNA (cDNA) since only DNA can be sequenced.

Bias Sources. The efficiency of the reverse transcription might also be sequence-dependent.

Size selection. The cDNA is size selected by gel electrophoresis. Only cDNA of a certain size class is to be sequenced.

Bias Sources. Size selection is dependent on the fragmentation (see [Fragmentation](#) (p. 34)). This means, if fragmentation produces (with high probability) always the same fragments of one gene, fragments might not be observed if they are too short or too long. Therefore, also the length of a mRNA might influence if and which possible fragments are observed. If a short mRNA is fragmented, the fragment length might be too short such that the fragments are always filtered out. A long mRNA might result in long fragments that are filtered out as well.

III. REQUIREMENT ANALYSIS AND PROBLEM ABSTRACTION

Amplification by Polymerase Chain Reaction. Finally, the selected fragments are amplified by [Polymerase Chain Reaction \(PCR\)](#) using the adapter sequences. PCR allows an exponential multiplication of the fragments in the library. This is needed since, during the library preparation, many fragments are lost due to the inefficiency of enzymes (e.g., reads without adapters), [Size selection](#) steps (p. 35), or [Subset selection](#) (p. 36) or random loss in purification steps ([mRNA enrichment](#) (p. 34)). The PCR amplification is, therefore, indispensable to obtain a sufficient amount of fragments for sequencing. Usually between 11 and 18 PCR cycles are completed, depending on the specific protocol and the amount of fragments before amplification. It should be noted that in some protocols PCR amplification is performed before size selection.

Bias Sources. Theoretically, each PCR cycle duplicates the number of fragments. Thus, after 11 cycles a single cDNA strand is multiplied about 1000-times ($2^{(cycles-1)}$). However, the efficiency of the amplification is not constant during the PCR as Karlen *et al.* [[Karlen et al., 2007](#)] showed. They describe three phases, where the first phase is not treated as it is suboptimal. The second phase is the exponential phase, where most molecules are doubled. In phase three, the reaction saturates due to too many target molecules, degradation of necessary ingredients and an increase in waste products. In this phase, only a linear amplification might be observed, if at all. Since the efficiency of the amplification is not constant during the PCR, the PCR increases the variance in the data and, therefore, enhances any previous bias of the preceding steps.

Thus, visualizing [read coverage](#) above a genome might show regions with high read coverage implying a highly expressed gene which is actually based on a low number of different reads (see [Figure III.6](#)). Furthermore, the total number of reads overlapping a gene ([read counts](#)) can be influenced which hampers a comparison of read counts between genes. This phenomenon is illustrated in [Figure III.7](#). Even though cDNA1 was present in a higher concentration, reads from cDNA2 are amplified with a higher rate. Please note that all previously described bias sources also influence the read counts of genes differently since they are sequence dependent.

Subset selection. During the complete protocol, often only a fraction of the output from a previous step is used as input for the next step. Of the total amount of RNA isolated in the beginning, only a defined amount is used for rRNA depletion. Thereof about half of it is further processed in RNA fragmentation. Neglecting questions about fragmentation efficiency, a yet smaller amount of RNA is used for adapter ligation and reverse transcription. Next, depending on the protocol, the complete or partial amount of the adapter ligated cDNA is used for size selection and final PCR. At last, only a part of the finished library resulting from the PCR is used for sequencing since, after PCR, the amount of sequences is too large to be sequenced.

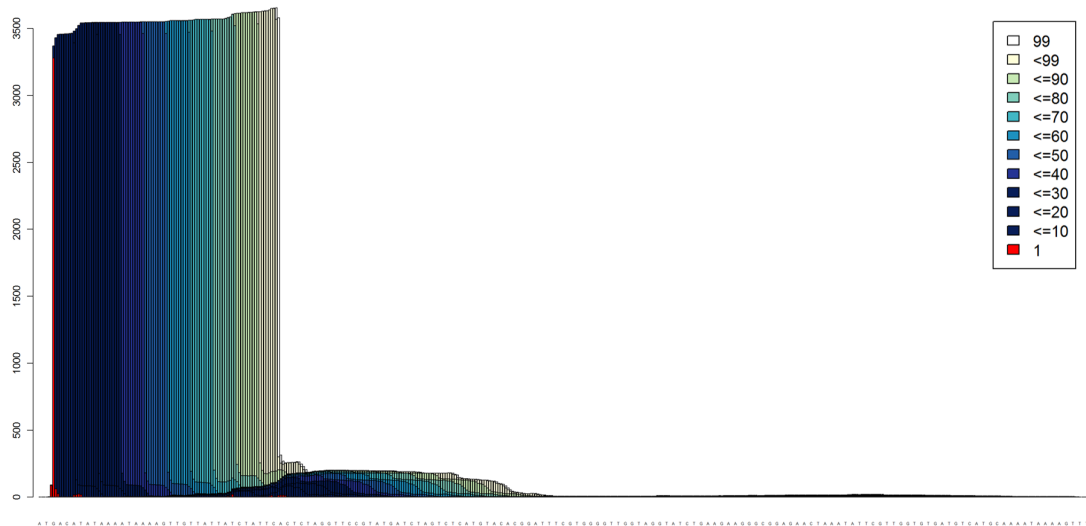


Figure III.6: Read coverage of the gene alr0402 (data: [Flaherty et al., 2011]). The number of reads overlapping each nucleotide is shown on the y axis. Colors represent the read positions (Dark colors encode positions at the front, red the first position). Most of the coverage is due to a high number of reads starting at one position.

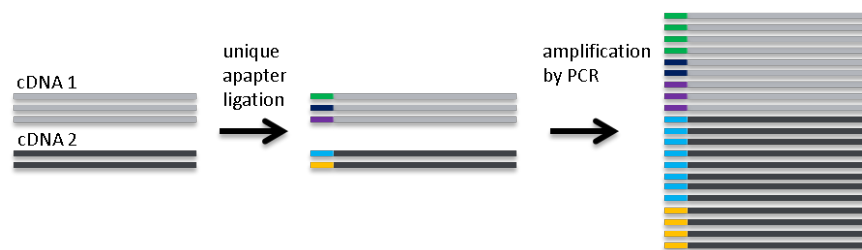


Figure III.7: Adapted from Fig. 1A in [Shiroguchi et al., 2012]. Two different cDNAs are shown which are all marked with a unique adapter in the first step. In the second step all cDNA fragments are amplified by PCR.

Sequencing and Mapping

Sequencing. After library preparation, as outlined above, the library is sequenced. The fragments contained in the library are physically separated and each fragment is amplified locally using PCR-like reactions (e.g., bridge-amplification, emulsion-PCR). Thus, at a given location, only one clonal PCR-product is found which is then sequenced. The multitude of fragments in one spot is necessary to obtain signals above detection background noise. Sequencing by synthesis is a synchronized chemical reaction that needs several rounds of blocking and de-blocking of the free DNA-ends which are sequenced. The final output from sequencing are called reads.

III. REQUIREMENT ANALYSIS AND PROBLEM ABSTRACTION

Bias Sources. Since the efficiency of the used enzymes is not 100%, free DNA-ends might not be blocked or de-blocked in time. This causes a decay in the signal as synchronization drops. Thus, the first part of a given sequence is generally more reliable than later parts (see Fig. III.8). However, other technical issues might cause an error any time. Therefore, a quality value for every sequenced nucleotide is given. These quality values are given as Phred quality scores Q . $Q = -10 * \log_{10}(p)$. p is the base-calling error probability. Thus, high values of Q represent a high sequencing quality and low values represent a high amount of uncertainty in this nucleotide. E.g., 50 stands for a 0.01% error probability, 10 for 10%, and 1 for around 80%.

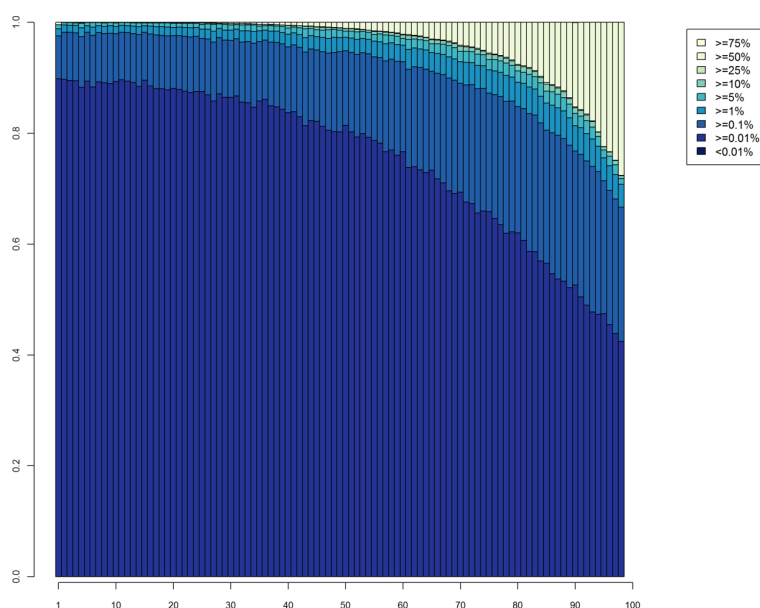


Figure III.8: Distribution of the nucleotide quality values above the read positions of all mapped reads. The Phred quality scores are divided into nine classes and transformed to the base error probability as shown in the legend. It is obvious that the amount of bases with low quality increases at the end of the reads.

Mapping. For RNAseq, the reads are mapped to an existing genome of the respective organism (see Fig. III.9). Common alignment tools (e.g., BLAST [Altschul et al., 1990], [Altschul et al., 1997]) are not designed for this task, in which huge amounts of short reads should be reliably matched to an existing sequence. Therefore, several new mapping algorithms have been proposed (for a review see [Schbath et al., 2012]).

Bias Sources. Due to errors in sequencing some reads cannot be mapped at all or the mapping is not unambiguous (see also [Assessing Mapping Quality](#) (p. 48)).

Visualization of RNAseq Data

Most common genome browsers can display RNAseq data. Furthermore, specialized genome browsers with respect to NGS and RNAseq data have been developed (see Section IV-3). A

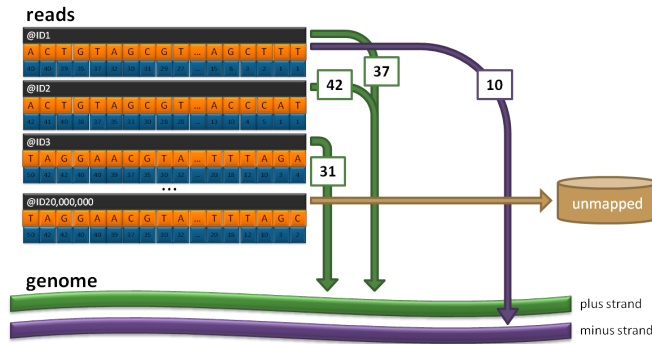


Figure III.9: Reads are sketched in the upper left part of the figure (often several million). The sequence itself and a Phred quality score for every position is given (see Sequencing). Lines sketches the mapping of reads to the genome. Reads might be mapped to one, several or no positions in the genome. Some mapping algorithms also give a mapping quality in Phred format (label on the arcs). Note: Parameter settings of mapping algorithms allow to skip reads which map at different positions or to map a read only to the "best" position (with respect to given parameters).

common way to visualize reads is to represent them as lines above the genome position they map to and to stack them (see Figure III.10). For strand specific reads a plot in two directions (up and down) is used. Alternative representations use color or arrows to encode the strand information. The stacked representation of reads is called scaffold view. They have the disadvantage that due to the stacking, visual gaps emerge. Another way to represent the reads is to represent the read coverage with line or bar charts. The height represents the number of reads overlapping a position in this representations (see Figure III.5 and Figure III.6). Both charts can be drawn strand specifically (up and down) but do not represent read start positions.

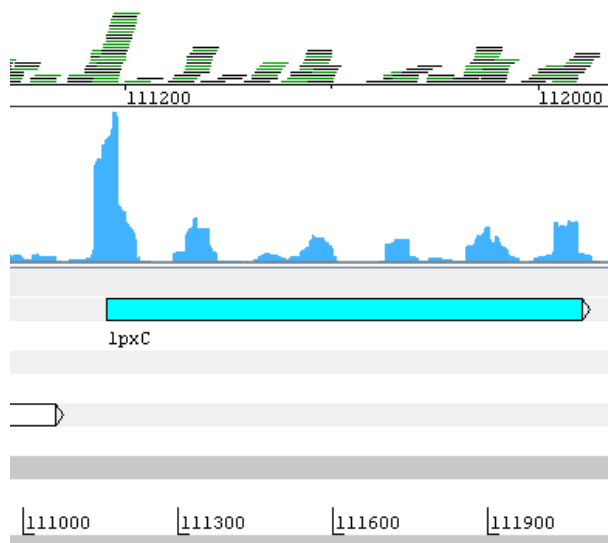


Figure III.10: Screenshot from the Artemis genome browser [Rutherford et al., 2000, Carver et al., 2012] which shows the read coverage as stacked reads and as a filled line chart.

III. REQUIREMENT ANALYSIS AND PROBLEM ABSTRACTION

Without any bias the read coverage would be distributed according to the relative transcript amounts found originally in the cells and the read coverage of a given gene would be uniformly distributed over the whole transcript length. Therefore, the read distribution could be approximated by a Poisson process with rate λ [Mortazavi et al., 2008, Marioni et al., 2008]. The assumption also leads to the introduction of the RPKM value, giving the value of reads mapping per kilobase genome per million mapped reads [Mortazavi et al., 2008]. A RPKM value, therefore, represents the normalized produced amount of a gene transcript.

However, the observed read coverage deviates from a uniform distribution (see Figure III.10). This is due to biases that are inherent in some of the steps listed in section III-2.2, the introduced biases lead to an uncertainty in the observed read coverage. Therefore, data trustworthiness should be considered in the analysis.

III-3 Data

In the following, the data related to RNAseq analysis will be described to provide the background for the visualization of this data. The aim of RNAseq is to measure genes, therefore, first a description of genomic data is given before RNAseq data sets are introduced. Genomic data comprises annotation data and, thus, the locations and functions of genes and [open reading frames \(ORFs\)](#), as well as meta data. RNAseq data sets imply the sequenced [reads](#) and sequencing quality information as well as the mapping of the reads to the genome sequence.

III-3.1 Genomic Data

Annotation Data

Known functional genetic elements are annotated for each organism. They are described by their location on the genome: by strand, start and stop, as well as by an identifier and annotation data like name (synonym), function and functional COG category [Galperin et al., 2015]. Beside genes, which encode for proteins, further genetic elements are annotated, e.g., rRNA and tRNA genes but also the family of so-called non-coding RNA (ncRNA). Non-coding refers to non-protein coding, ncRNA has other regulatory functions and comprises different sub-groups, e.g., anti-sense RNA, small RNA and others (see also [Determination and Annotation of Genes](#) (p. 30)).

Beside annotated genetic elements, the FOG-Project¹ addressed potential new [overlapping genes \(OLGs\)](#), so-called [shadow ORF \(sORF\)](#). As ORFs are clearly defined in bacteria, by

¹FOG-Project: “Finding new overlapping genes and their theory (FOG-Theory)”, part of the priority programme “Information and Communication Theory in Molecular Biology” (InKoMBio SPP 1395) of the German Research Foundation (DFG), 2010-2015.

a start and a stop codon in the same reading frame, the determination of all ORFs is straight forward. By matching ORFs against annotated genes, all not-annotated ORFs (naORFs) can be determined. Based on the definition, that a shadow ORF has to overlap at least 93 bp with a gene, all sORFs of an organism can be determined.

I summarize all genetic elements, including ORFs as *Annotation Data* in the following. The data can be represented as a table, with the attributes *ID*, *name*, *start*, *stop*, *strand*, *product*, *COG*, *type* and *sequence*. *Type* denotes gene, RNA, naORF, sORF and so on. The connection of genetic elements, e.g., which sORFs overlap a gene, can be represented in a second linked table, with the two IDs of the overlapping ORFs and the attributes *overlapping length*, *direction*, and *frame*. Direction is a boolean attribute with the conditions same-strand and opposite-strand. Frames are defined as in Figure III.11.

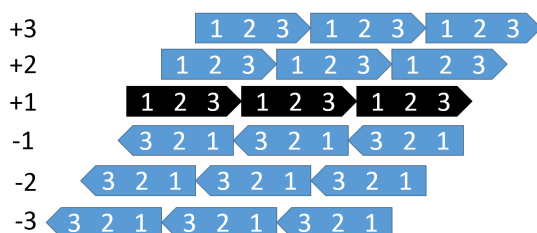


Figure III.11: Frame +1 is the reference frame. The frame of an ORF is always given with respect to a reference. Frame +2 is shifted one position compared to the reference on the same strand. Frame +3 shifted two positions compared to the reference on the same strand. Frame -1 is not shifted compared to the reference but is located on the opposite strand. Frame -2 and -3 are also located on the opposite strand but are shifted one resp. two positions in the opposite direction of the reference.

Meta Data

Additionally, meta data could exist for all genetic elements. This can be a set of scalar features per genetic element, for instance, molecular weight or number of polar amino acids; or more complex meta data, for instance, BLAST result tables [Altschul et al., 1990, Altschul et al., 1997].

BLAST Result Tables comprise all similar sequences (hits) to a query sequence as well as attributes describing the similarity. The most important is the expected value which describes the probability that a hit with such a score could have occurred by chance. The further important attributes are *organismID*, *organism name*, *taxonomicalID*, *proteinID*, *protein name*, *protein function*, *expect value*, *score*, *hit start*, *hit stop*, *query start*, *query stop*, *alignment length*, *number gaps*, *number identities*. Phylogenetic distance is a derived attribute from the whole taxonomy, based on the *taxonomicalID* of the query and the BLAST hit.

III. REQUIREMENT ANALYSIS AND PROBLEM ABSTRACTION

III-3.2 RNAseq Data

A sequencing run results in a set of sequence fragments, so called **reads**. The information of a read includes its sequence as well as the qualities of each sequenced **nucleotides**.¹

Mapping algorithms are used to map reads to a reference genome sequence (see Figure III.9). Based on (quality) parameters, reads can either be mapped with a certain quality or cannot be mapped at all². Some reads can be mapped to several positions, either due to repetitive genome sequences or poor mapping quality (see also **Mapping** (p. 38)). From the mapping of reads several data abstractions can be derived (see also Figure III.12):

For each read *mapping position, mapping quality and mismatched positions.*

For each genome position *reference nucleotide, number of overlapping reads, number of starting reads, set of sequenced nucleotides, set of sequence qualities, set of mapping qualities.*

For each open reading frame (ORF) *number of overlapping reads (counts), gene activity level, vector with the numbers of overlapping reads per ORF position (read coverage).* As well as derived attributes like *coverage* (the percentage of the ORF covered with reads) and *fit*. *Fit* describes how well the **read coverage** fits to the ORF (see Section IV-5.3 for more details).

A **gene activity level** can be determined from the reads overlapping a gene. Mortazavi *et al.* suggested the value **Reads Per Kilobase per Million mapped reads (RPKM)**, which is normalized for the gene length as well as the sequencing depth, that differ between experiments [Mortazavi *et al.*, 2008]. RPKM values can, therefore, be compared between experiments and theoretically also between genes in one experiment. However, due to sequence-dependent biases³, RPKM values might not be accurately between genes in one experiment (see Section III-2.2). For comparison of genes between experiments often the **counts** are considered without normalizing for the gene length, as just the same genes are compared. The ratios (**fold-change**) between conditions are then comparable between genes, as the strength of a bias for one gene is expected to be the same for different conditions. This analysis is called **differential gene expression** analysis.

¹In case of paired-end sequencing additionally, an ID links to the paired read. However, in this thesis just single, strand specific sequencing is considered.

²Some reads do not resemble any position in the genome, most likely due to many sequencing errors, or since they do not originate from this genome but from an unknown **plasmid**. These reads are not mapped and discarded.

³Bias which is dependent on the **nucleotide** sequence of the DNA. See, for instance, **Fragmentation** (p. 34).

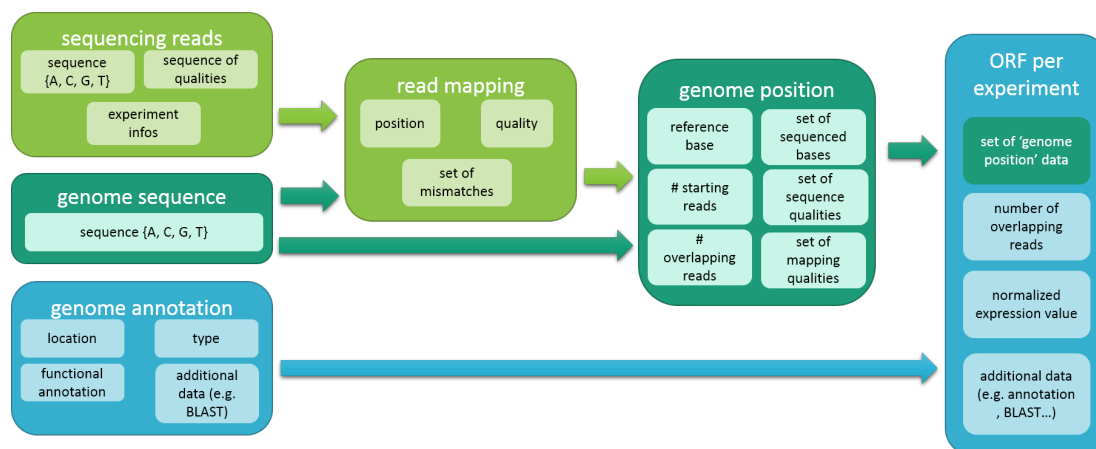


Figure III.12: This graphic illustrates the data complexity and structure of RNAseq data. Sequencing data is generated by next-generation-sequencing with sequencing machines. Sequencing machines generate nucleotide sequences as well as a quality value for each nucleotide. Genome sequence data is provided by public data bases like NCBI [Coordinators, 2013], as well as gene annotation data. Here genome annotation also includes not annotated open reading frames (e.g. sORFs). Sequencing reads are mapped by state-of-the-art short read mapping algorithms to the genome reference sequence. Mapping algorithms provide the position of the mapping and may provide a quality of the mapping. From the read sequence and the reference sequences, mismatches can be determined. Combining the data from the previous steps leads to a set of attributes available for each genome position. For each entity of the genome annotation (ORFs) further aggregation values can be calculated.

Differential Gene Expression Data

Differential gene expression analysis considers the relative comparisons between **gene activity levels** of the same gene under different experiment conditions (see Figure V.1). A gene (ORF) is called differentially expressed in two conditions: If the **gene activity level** (or **counts** of reads) of the gene (ORF) differs significantly between these conditions.

Rapaport *et al.* [Rapaport et al., 2013] provide a discussion of different approaches to determine differentially expressed genes. In this thesis, *edgeR* [Robinson et al., 2010] is used to determine differentially expressed genes, which determines the \log_2 **fold-change**¹ between the counts of different conditions, as well as the p-value of the statistical test. The fold-change is denoted as **gene activity ratio (GAR)** in this thesis and the set of GARs resulting from a pairwise comparison of several conditions is denoted as the **gene activity ratio (GAR) pattern** of a gene (short GAR pattern). All pair-wise comparisons (n:n) are considered which is in contrast to many studies. Mostly a (1:n) comparison is considered, i.e., several conditions are compared against one reference condition.

¹The logarithm of the fold-change makes up- and down-regulation comparable since $|\log_2(\frac{a}{b})| = |\log_2(\frac{b}{a})|$.

III. REQUIREMENT ANALYSIS AND PROBLEM ABSTRACTION

The non-uniform [read coverage](#) over genes [Li et al., 2010] influences the fold-change calculations. Therefore, a measure is provided to estimate the fold-change variance over the gene to indicate the quality of the fold-change estimation. The quality can be understood as the trustworthiness of the correctness of the fold-change (GAR).

The fold-change is calculated in a sliding-window fashion over the gene. The variance and mean are calculated from the fold-changes of the slices. The coefficient of variation expresses the quality of the fold change. Also other measures can be applied, e.g. the statistical significance (p-value of the statistical test) of fold-changes. Or, furthermore, a measure of the [gene activity levels](#) could act as a quality measure as low activity levels influence the assessment of the trustworthiness by the biologists.

III-4 Problem Abstraction

This thesis will cover the tasks for [RNA Sequencing \(RNAseq\)](#) data raised in the [FOG-Project](#). In this project the aim was to use RNAseq data to verify new [overlapping genes \(OLGs\)](#), as well as to detect relations and/or similarities to annotated genes (with a known function) in bacteria. However, for a broader applicability, I will cover genes in general in this thesis, i.e., new overlapping genes but also new genes ([open reading frames \(ORFs\)](#) not recognized as genes yet) and genes without an annotated function.

A better understanding of bacteria genes and proteins leads to a better understanding of cellular mechanisms and gene networks. Comprehension of these is necessary to clarify human [pathogenicity](#) and to develop new medical treatments (medicines) on the one side. On the other side, many bacteria are used in biotechnology to produce substances and medicines like insulin. A better understanding will allow improving the production yield and can open up possibilities for the industrial production of new substances. In this respect, especially a detection and consideration of new genes, is of high importance. Concluding new genes might have direct effects on human pathogenicity, antibiotic resistance or might be missing entities in gene networks which hamper a comprehensive understanding of annotated genes.

Regarding the aim to discover and describe genes with the support of visual analytics, I identified the following two research problems:

Assessment of the trustworthiness of RNAseq measurement. The RNAseq measurements of new gene candidates as well as of annotated genes needs to be verified by an expert as the process of RNASeq is error-prone (see Section [III-2.2](#)). Thus, experts must be provided with tools to capture these uncertainties in order to assess the trustworthiness of RNAseq measurements and to verify a (new) gene as active. See Chapter [IV](#) for the *NGS overlap searcher*, a

system that addresses this problem with visual analytics and further identifies and fills research gaps. Section III-4.2 discusses tasks and Section IV-2 derives requirements based on these tasks.

Comparison of gene activity levels between different experiment conditions. The reaction of genes in different experiment conditions and a comparison between conditions allows to relate genes with a function. Since this data is high-dimensional, complex, and large, and the task is of exploratory nature, solutions must provide a guided analysis and handle the large volumes of data. See Chapter V for *VisExpress*, a system that addresses this problem with visual analytics and further identifies and fills research gaps. Section III-4.3 states tasks and Section V-2 derives requirements based on these tasks.

In the next section, a task taxonomy is introduced which is used to describe the tasks related to the aforementioned research gaps in a formal and abstracted manner (III-4.2 and III-4.3).

III-4.1 Task Taxonomy

The problem and task abstraction in the next section is based on “A Multi-Level Typology of Abstract Visualization Tasks” of Brehmer and Munzner [Brehmer and Munzner, 2013] and the “Visualization Analysis and Design.” book of Tamara Munzner [Munzner, 2014]. See Figure III.13 for an overview to explain **Why?** and for which user goals visualizations are used. This taxonomy distinguishes between *Actions* and *Targets*. *Actions* define users goals and *Targets* describe which data aspects are of interest to the user.

The user goal with a visualizations is either to *consume* information or to *produce* information. On the *consume* side the goals to *discover* new information and to *present* information can distinguished, for instance, to communicate the result of a data analysis in a graphic. Lastly, the goal can be to *enjoy*, for example, a infographic in a newspaper or a blog. On the *produce* side, users might have the goal to *annotate* a graphic, for instance, with labels. Users might also want to *record*, for example, users can record findings with screenshots, or record their analysis with a graphical history of actions for reproducibility. Lastly, users may want to *derive* new data based on the original data in a visually guided fashion.

All goals aim at a set of elements of interest which are *searched* for. Here it can be distinguished if the location and/or the targets are know or unknown (see Figure III.13 for the definition of *Lookup*, *Locate*, *Browse* and *Explore*). I changed this definition slightly and extended it with the data characteristics. See Figure III.14 for the definition I use in this thesis. In the context of genes, the target would be a gene of interest. Users might either search for a specific target, gene x, for instance, or they search for a gene with a specific characteristic, e.g.,

III. REQUIREMENT ANALYSIS AND PROBLEM ABSTRACTION



Figure III.13: This figure appeared as Figure 3.1 in “Visualization Analysis and Design. Tamara Munzner, with illustrations by Eamonn Maguire. A K Peters Visualization Series, CRC Press, 2014.” [Munzner, 2014]. The figure is released under the Creative Commons Attribution 4.0 International license (CC BY 4.0) [Creative Commons, 4].

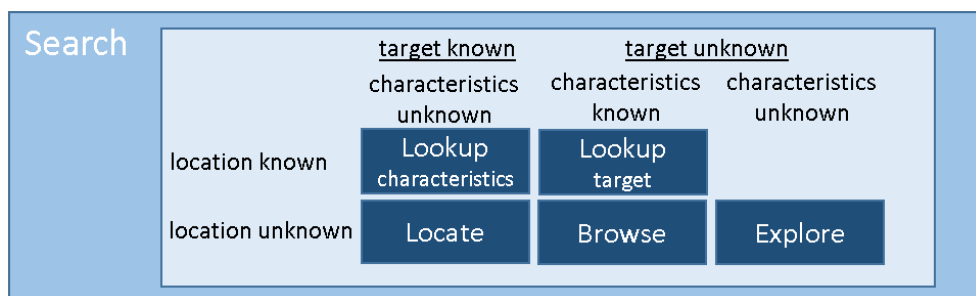


Figure III.14: This graphic explains the *search* component to accomplish user goals (see Figure III.13). Three influencing factors are distinguished. If the location is known, if the target (ID) is known and if the characteristics of the element of interest are known. Depending on these three aspects the *search* is named *lookup*, *locate*, *browse* or *explore* (see also Section III-4.1).

a gene with a specific [gene activity level](#). The location of an element of interest is either known or unknown. In genome browser presentations the location of genes is known and, thus, users can *lookup* the characteristics of the gene (e.g., the *gene activity level*). If genes are represented in an unordered list or spreadsheet, users need to *locate* the gene of interest first. The element of interest can also be described by specific characteristics only and the locations is known all the same. If, for instance, a list is ordered with respect to a specified characteristic, the element of interest is the first in the list and the user can *lookup* the target (e.g., the gene name). If the location is not known, users need to *browse* to find a target (gene) with the specified characteristics. If neither the target is known nor the location, and the user is just looking for unexpected characteristics, the user needs to *explore* the whole data set. As unexpectedness is not definable (otherwise characteristics could be defined), the location can never be known for this case of search.

The *actions* performed on identified elements are classified as *identify*, *compare* and *summarize (query)*. *Identify*, addresses one element, *compare* two or more elements and *summarize* addresses all elements of interest. Users can, e.g., *identify* the annotation data of a gene, or *compare* the *gene activity level* of two genes. To *summarize* all genes requires data overviews.

Data aspects (*targets*), that can be of interest for a user, are (high-level) *trends* or patterns, *outliers*, and other data *features* defined by the application. On the attribute level (low-level), the *distribution* of one attribute (including *extremes*) might be of interest, or the *dependency*, *correlation* and *similarity* between many attributes. For specific data types like networks, more specific data aspects might be of interest, e.g, the *topology* of a network.

III-4.2 Assessment of the Trustworthiness of RNAseq Measurements

As explained in section [III-2.2](#), the whole data preparation process, the sequencing process and the read mapping are error prone. Therefore, one task is to assess the trustworthiness of [RNAseq measurements](#). The term “measurement” is defined as the reads that are mapped to one [open reading frame \(ORF\)](#). I use the word “trustworthiness” here to emphasize the aim of the task. Due to missing generally accepted measures or thresholds, domain experts have to **trust** in the results to claim findings in publications or run further time and cost intensive verification experiments. The main tasks are (T-I) to assess the trustworthiness of RNASeq measurements and (T-II) to identify interesting ORF candidates worthwhile a further inspection. I will describe these tasks in detail in the following.

III. REQUIREMENT ANALYSIS AND PROBLEM ABSTRACTION

T-I

Assessing the trustworthiness of RNAseq measurements is the main task. As the trustworthiness is dependent on many aspects, the data representation has to incorporate these. In the following, I discuss the different aspects influencing the trustworthiness of RNAseq measurement. First I will assess the mapping quality and conclude with further reasons for a low trustworthiness.

Assessing Mapping Quality. One point that potentially influences the trustworthiness is the mapping. Reads might be mapped to a wrong location in the genome. This might happen due to many sequencing errors, a sequence may not be covered in the reference genome (e.g., from a so far unknown [plasmid](#)), or by repetitive sequences in the genome. In the first two cases, many mismatches in the mapping can be expected (low mapping quality), in the last case the read can be mapped to several genome locations (see also [Mapping](#)(p. 38)).

In order to assess the mapping quality, I visualized the mapping quality as error probability for selected genes (see Figure [III.15](#)). Therefore, I considered the mapping information for each genome position (see [For each genome position](#) (p. 42)). In order to represent the quality distribution, I binned the mapping qualities per position based on the given Phred Score (see Table [III.1](#)). Phred quality scores Q , with the error probability P are defined as: $Q = -10 * \log_{10} * P$

Table III.1: Thresholds for mapping quality bins applied for Fig. [III.15](#). Thresholds are meant as upper bounds, lower bounds are given by the next bin. Phred Score $Q = -10 * \log_{10} * P$

Bin Number	Phred Score	Error Probability
1	> 40	< 0.01%
2	≤ 40	≥ 0.01%
3	≤ 30	≥ 0.1%
4	≤ 20	≥ 1%
5	≤ 13.0103	≥ 5%
6	≤ 10	≥ 10%
7	≤ 6.0206	≥ 25%
8	≤ 3.0103	≥ 50%
9	≤ 1.249387	≥ 75%

Data inspections revealed that the mapping error probabilities are low. Landstorfer *et al.* [9] showed, furthermore, that mis-mappings are unlikely. Out of 7 million reads of the EHEC genome just one mapped to the mouse Y-chromosome. Therefore, mapping errors were neglected in this thesis (see Fig. [III.15](#)).

Further Reasons for a low Trustworthiness. Further reasons for wrong measurements of an ORF are **leaking transcription** from a neighboring gene and background transcription.

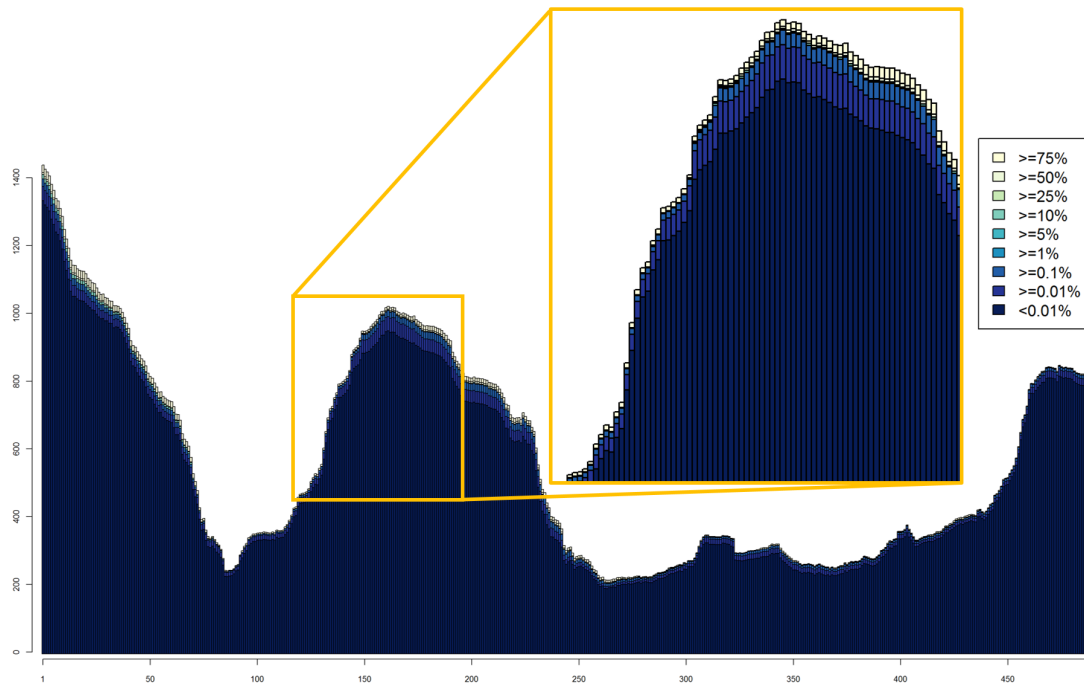


Figure III.15: Shown is a stacked bar chart, representing the number of overlapping reads per position of the gene *alr0022* (data: [Flaherty et al., 2011]) by bar height. Color represents the distribution of reads belonging to an error probability bin (see Tab. III.1 for the thresholds). Dark blue represents low mapping error probability. Obviously the mapping quality is very high for the whole gene.

This means the **untranslated region (UTR)** of a neighbor ORF might cover adjacent ORFs. Thus, an ORF might be covered at the beginning with reads actually belonging to the upstream gene. The signal “leaks out” in the following ORF or starts in the preceding ORF (see Fig. III.16(a) and III.16(b)). The positions of read starts can be helpful to assess if the majority of reads overlapping an ORF actually start before the ORF and might, therefore, originate from the preceding ORF.

Background transcription is another reason for wrong measurements and has been reported for prokaryotes and eukaryotes (see e.g., [Clark et al., 2011, Bruno et al., 2010] and [Vivancos et al., 2010]). However, due to the small probability, the abundance of background transcription is expected to be low overall. Even though amplification bias could still lead to high counts for some locations (see **Amplification by Polymerase Chain Reaction** (p. 36)). This could be due to amplification (PCR) duplicates or other biases (see Fig. III.16(c)).

Furthermore, **weak transcription** signals can occur. This means either even but low, or sparse and scattered (see Fig. III.16(b)). Users might also discard such candidates.

III. REQUIREMENT ANALYSIS AND PROBLEM ABSTRACTION



(a) In the middle a short gene (*STM14_2238*) is shown. The read coverage is low at the beginning and increases at the end of the gene. It is likely that this increase is due to the following gene since this region could be the **UTR** of *STM14_2239*. Therefore, it is doubtful if *STM14_2238* is active.

(b) In the middle a short gene (*STM14_0432*) is shown, the transcription is in a very low range over the whole gene but higher at the beginning and end. Therefore, it is likely that the gene signal is not valid but belongs mostly to the preceding and following gene *STM1_0431* and *STM14_0433*.

(c) Here a gene with one distinct peak is shown which indicates that a high number of reads start at the same position. This might indicate a PCR artifact.

Figure III.16: The upper panel shows the transcriptional signal (coverage curve) of a gene ± 200 nt as a line chart, the data is normalized for the library size and multiplied with the mean library size. The differentially expressed conditions are shown in blue and red, the further four conditions are shown in gray to keep the context. The panel in the middle shows the location of the gene of interest and the neighboring genes in part. The lower panel shows the log scaled data.

In order to solve the task **T-I**, the challenge is to design an expressive and effective visual representation of the read coverage to support the assessment of the different reasons for untrustworthy measurements. The visualization is the basis for domain experts assessing the trustworthiness of ragged measurements. Here, to *discover* the data is the main user goal and users need to verify **open reading frames (ORFs)** of interest. Either users verify all or specific ORFs (by *lookup*) and *identify* their characteristics (gene neighborhood, **read coverage** of the ORF (**RNAseq measurement**) and **Annotation Data**). The data aspects of interest (*targets*) are, in this case, the *distribution* of the read coverage (leaking transcription, weak transcription) and *extrema* (possible amplification bias). Furthermore, users might aim to *explore* the whole data set with the aim to *compare* genes and to get an overview of gene measurements. Besides, the aforementioned high level data aspects, *correlation* and *dependency* between ORFs (possible **operon**) and *similarity* between ORFs are of interest, as well as overall data patterns (*trends* and *outliers*), e.g., “Is there a weaker RNAseq measurement signal at the end of all ORFs?”. See Figure III.17(a) for the actions and targets, and Section **Task Taxonomy** for more details about the used terms to abstract domain experts’ tasks.

T-II

Identify ORFs worthwhile a further inspection. The number of sORFs is in the range of tens of thousands for many bacteria species. Thus, a detailed visual assessment of all is not feasible and a limitation of the analysis time is needed.



(a) **T-I** - Assess the trustworthiness of RNAseq measurements.

(b) **T-IIb** - Identify ORFs worthwhile a further inspection by filtering out discardable candidates.

Figure III.17: These figures are build upon the material which appeared in Figure 3.1 in “Visualization Analysis and Design. Tamara Munzner, with illustrations by Eamonn Maguire. A K Peters Visualization Series, CRC Press, 2014.” [Munzner, 2014]. The figure is released under the Creative Commons Attribution 4.0 International license (CC BY 4.0) [Creative Commons, 4].

T-IIa *Identify ORFs worthwhile a further inspection by increasing the scalability.* One option is to provide a more scalable visualization which allows to quickly assess if an ORF is worthwhile a further inspection. This option has the advantage that all ORFs are still visualized.

T-IIb *Identify ORFs worthwhile a further inspection by filtering out discardable candidates.* Alternatively, users can *produce* a reduced data set by *deriving* a binning in reasonable candidates and discardable candidates. This can be achieved by setting a threshold to distinguish reasonable and discardable candidates. Discardable candidates are filtered out. As explained in Section **T-I** to assess the trustworthiness of an ORF measurement, different aspects need to be considered. All these aspects need to be considered simultaneously and, thus, a definition of a single threshold is not possible and a combination of thresholds is not straightforward. This is especially the case as thresholds might vary between species and experiments. As reasonable parameter settings are not known, users need to adjust parameters. Therefore, users need to *explore* the resulting binning and *summarize* the data to steer the effects of parameter changes. Data aspects of interest are thereby, the *distribution* of gene attributes (see [For each open reading frame \(ORF\)](#) (p. 42)),

III. REQUIREMENT ANALYSIS AND PROBLEM ABSTRACTION

as well as the *dependency*, *correlation* and *similarity* of gene attributes. See Figure III.17(b) for the actions and targets and Section [Task Taxonomy](#) for more details on used terms.

III-4.3 Comparison of Gene Activity Levels between Different Conditions

Note

This section is based on the following publication and parts of this section will appear in this publication [13]¹:

[13]: [Svenja Simon](#), Sebastian Mittelstädt, BC Kwon, Andread Stoffel, Richard Landstorfer, Klaus Neuhaus, Anna Mühlig, Siegfried Scherer, and Daniel A. Keim. “*VisExpress - Visual Exploration of Differential Gene Expression Data*.” Information Visualization, 1-26, DOI: [10.1177/1473871615612883](https://doi.org/10.1177/1473871615612883), Published online before print December 14, 2015.²

¹For the division of responsibilities and work, as well as a statement of contributions in this publication, see [VisExpress - Visual Exploration of Differential Gene Expression Data](#) (p. 11).

²I own (with the co-authors) the copyright of this publication. The SAGE Publications Ltd holds the sole and exclusive right and license for publishing ([13]). The definitive version is available at <http://ivi.sagepub.com/>
Direct link to the published article: <http://dx.doi.org/10.1177/1473871615612883>

The overall aim to apply RNAseq in the course of the FOG-project² was to discover and describe new [overlapping genes \(OLGs\)](#). Many (new) genes are expected to be active under non-standard laboratory conditions only, whereby these genes have been overlooked under standard laboratory conditions so far. The analysis of a number of different non-standard conditions is, therefore, the mean of choice to discover new genes. Beside the screening for [RNAseq measurements](#) for each single [open reading frame \(ORF\)](#) to assess the trustworthiness of the measurements (see Section III-4.2), several experiment conditions also allow to compare the [gene activity levels](#) of ORFs between conditions. In order to fully exploit the RNAseq data set with several experiment conditions we, therefore, decided to use [differential gene expression data](#) as *derived* data from the [gene activity levels](#). Differential gene expression data opens up new analysis possibilities. If the measurements between two conditions differ significantly, it can not only be inferred that the ORF is indeed transcribed and active but also that its function is related to the changed conditions. A comparison to genes, which react similar to the changed conditions, but for which the function is known, allows generating further hypotheses about the function of the ORF and to draw further conclusions.

²FOG-Project: “Finding new overlapping genes and their theory (FOG-Theory)”, part of the priority programme “Information and Communication Theory in Molecular Biology” (InKoMBio SPP 1395) of the German Research Foundation (DFG), 2010-2015.

For this thesis, I extend the task of discovering and describing new overlapping genes to discovering and describing genes in general. Description denotes here to learn more about the function of genes in the long run and to generate and test hypotheses about gene functions or their correlation with certain other genes or conditions, in the short run. The last sentences describe the more specific aims and tasks of domain experts, when analyzing and comparing the RNAseq data from different experimental conditions. Thereby the challenges are to handle the large data volume and to allow an expressive exploration of the data. Thus, patterns need to be interpretable for domain experts, similar patterns need to be grouped and to be related to known gene functions. Additionally, the trustworthiness needs to be considered, regarding RNAseq measurements (see also Section III-4.2), as well as regarding differential gene expression. The following set of tasks can be summarized which all have the goal to *discover* the data set (see Section [Task Taxonomy](#) for more information on the used terms):

T1 *Generate hypotheses about the function of genes.* In this exploration task, biologists want to find new hypotheses about genes and their potential functions. In order to generate these hypotheses, they *explore* the data set for genes¹ with an unexpected function in a set of genes with similar [gene activity ratio \(GAR\) patterns](#). Unexpected is interpreted here with respect to the meaning of the examined conditions and the interpretation of the GAR pattern. The first part of this task is to build a subset of similar genes which is also relevant for other tasks (see Figure III.18(a)). Therefore, gene characteristics² need to be *identified*, genes need to be *compared* and similar ones need to be *summarized*. Important data aspects are, thereby, the GAR patterns (*trend*).

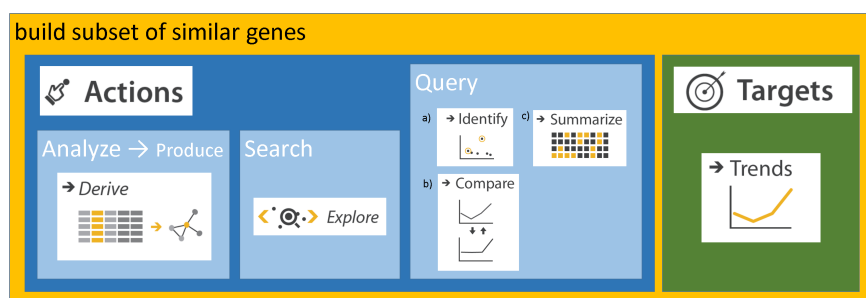
The second part of this task is *browsing* for genes with an unexpected function (see Figure III.18(b)). Their data characteristics need to be *identified* and interpreted with respect to the involved conditions. Furthermore, these genes should be *compared* to other genes in the subset with an expected function and *summarized* to reveal common features, e.g., the same unexpected function. Thereby, the GAR patterns (*trend*) and the *distribution* of functions assigned to the genes in the subset are important.

T2 *Test hypotheses about the function and reaction of genes.* In this task, biologists make an assumption about the reaction of genes to the experimental conditions. Through [differential gene expression](#) analysis, they can confirm or reject their hypotheses, if genes with particular functions have an expected or unexpected [gene activity ratio \(GAR\) pattern](#). Here, biologists have a data characteristic in mind they want to *browse* for (see Figure III.18(c)). Thereby, the

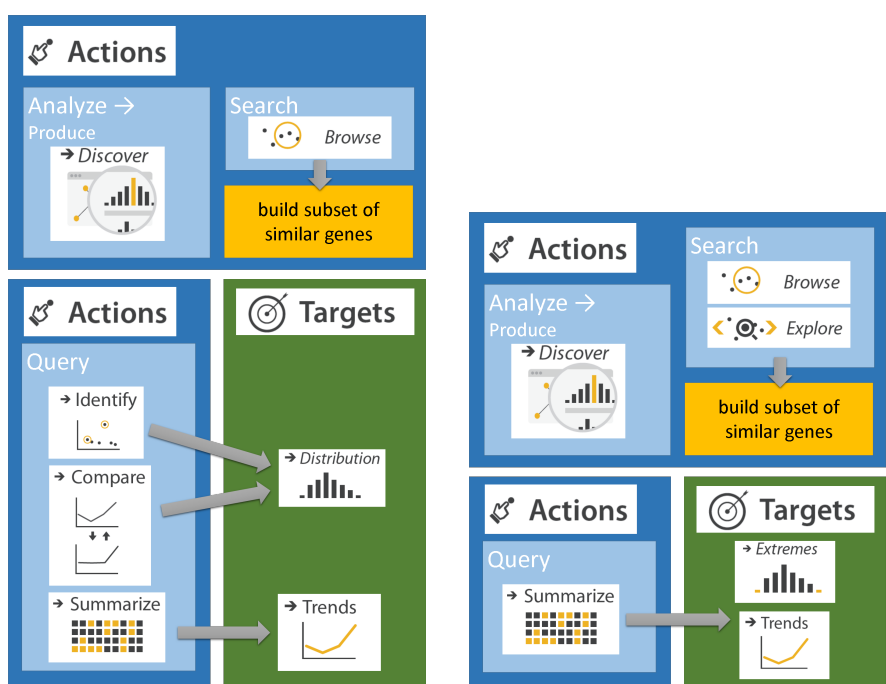
¹genes or ORFs

²gene activity ratio (GAR) patterns and [Annotation Data](#) (p. 40)

III. REQUIREMENT ANALYSIS AND PROBLEM ABSTRACTION



(a) Building a subset of similar genes. This task is important for all other tasks.



(b) Tasks T1. Generate hypotheses about the function of genes.

(c) Tasks T2. Test hypotheses about the function and reaction of genes.

Figure III.18: These figures are built upon the material which appeared in Figure 3.1 in “Visualization Analysis and Design. Tamara Munzner, with illustrations by Eamonn Maguire. A K Peters Visualization Series, CRC Press, 2014.” [Munzner, 2014]. The figure is released under the Creative Commons Attribution 4.0 International license (CC BY 4.0) [Creative Commons, 4].

data aspects are the GAR patterns (*trend*) and the *extremes* of functions assigned to the gene subset (most prominent function). In order to draw conclusions, biologists need to build sets of similar genes (see Figure III.18(a)) again. However, in contrast to T1, where biologists need to understand and interpret the GAR pattern, biologists have a GAR pattern in mind, they just want to find. T2 also comprises hypotheses about experimental conditions. For instance: “Condition 1 and 2 should reveal the same GAR to the other conditions for most of the genes.” For this

subtype of T2, data characteristics¹ are partly known for which users can *browse*. Additionally, they need to *explore* the whole data set for counter examples. The remaining actions and targets are the same for this subtype of T2. Remark: For T2 a (1:n) comparison of conditions is not sufficient since this involves the interrelation of all conditions. Therefore, a (n:n) comparison of conditions is required (see [Differential Gene Expression Data](#) (p.43)).

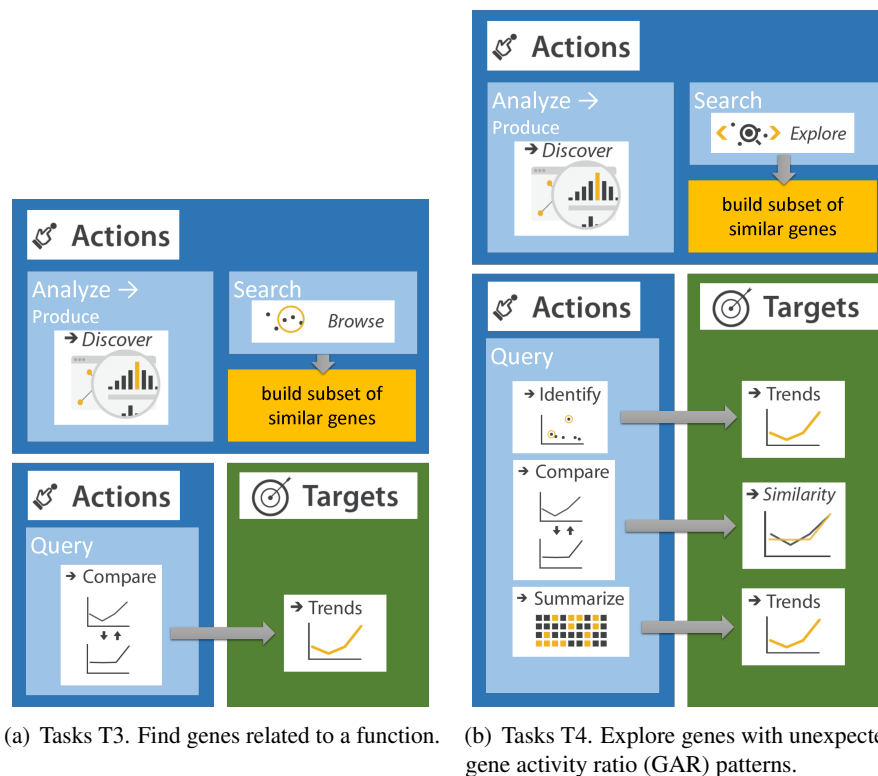


Figure III.19: These figures are built upon the material which appeared in Figure 3.1 in “Visualization Analysis and Design. Tamara Munzner, with illustrations by Eamonn Maguire. A K Peters Visualization Series, CRC Press, 2014.” [Munzner, 2014]. The figure is released under the Creative Commons Attribution 4.0 International license (CC BY 4.0) [Creative Commons, 4].

T3 *Find genes related to a function.* When biologists analyze a single function, they are interested in identifying genes, yet unknown, to be related to this function. In order to find these genes, they need to *browse* for genes with the characteristic of a specific function (see Figure III.19(a)). For a further interpretation biologists first need to build sets of similar genes (see Figure III.18(a)). Secondly, the need to *compare* the gene activity ratio (GAR) patterns of genes not related to the specified function with genes that comprise this function. Genes with the most similar GAR pattern will become potential candidates for further investigations, thereby, the

¹gene activity ratio (GAR) pattern

III. REQUIREMENT ANALYSIS AND PROBLEM ABSTRACTION

data characteristic are the GAR patterns (*trend*). In contrast to T2 biologists search for the data attribute “function” and not for the data attribute “GAR pattern”.

T4 *Explore genes with unexpected gene activity ratio (GAR) patterns.* If unexpected GAR patterns are *explored* in the data set, all genes with similar GAR patterns need to be *identified* (see Figure III.18(a)). Genes with unexpected patterns then need to be *identified* and *compared* in order to examine their *similarities* to other genes and their functions. Data aspects are here the GAR pattern (*trend*) and the distribution of gene functions (see Figure III.19(b)). In contrast to T1, where the focus is to generate hypotheses about gene functions based on the functions of similar genes and the interpretation of the GAR pattern, the focus in T4 is on a higher level. The aim of T4 is to get an overview (to *summarize*) how genes react to changed conditions. A subsequent task would be T1 (see Figure III.18(b)).

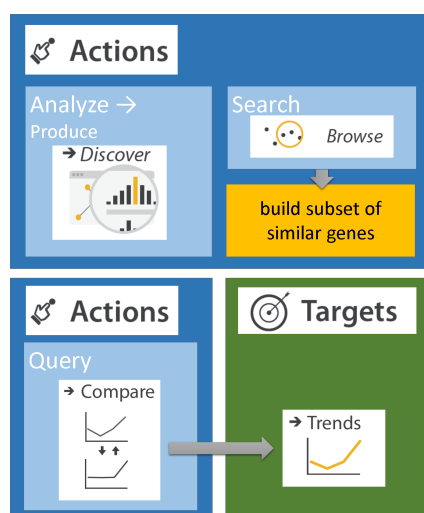


Figure III.20: Tasks T5. Relate new genes candidates to genes with known functions. (This figure is built upon the material which appeared in Figure 3.1 in “Visualization Analysis and Design. Tamara Munzner, with illustrations by Eamonn Maguire. A K Peters Visualization Series, CRC Press, 2014.” [Munzner, 2014]. The figure is released under the Creative Commons Attribution 4.0 International license (CC BY 4.0) [Creative Commons, 4].)

T5 *Relate new gene candidates to genes with known functions.* Open reading frames (ORFs) that show a differential expression are likely to be genes. In order to understand their function they need to be compared and related to genes which similar gene activity ratio (GAR) patterns and known functions. This task is highly related to task T1 (see Figure III.18(b)). However, here the targets are known and need to be *located*. Next a subset of similar genes needs to be built (see Figure III.18(a)) and the gene candidate needs to be *compared* with the known genes in the subset. This includes to *summarize* the functions of known genes in the subset. The important data characteristics are again the GAR patterns (*trend*) and the *distribution* of functions assigned to the genes in the subset (see Figure III.20).

Chapter IV

Visual Analysis for the Trustworthiness Assessment of RNAseq Measurements

Note

This chapter is mainly based on the following publications and parts of this chapter appeared in these publications [14, 10]¹:

[14]: Svenja Simon, Daniela Oelke, Richard Landstorfer, Klaus Neuhaus, and Daniel A. Keim. “*Visual Analysis of Next-Generation Sequencing Data to Detect Overlapping Genes in Bacterial Genomes.*” 2011 IEEE Symposium on Biological Data Visualization, October 23 - 24, Providence, Rhode Island, USA, 47-54, IEEE, DOI: [10.1109/BioVis.2011.6094047](https://doi.org/10.1109/BioVis.2011.6094047), 2011.²

[10]: Oelke, Daniela, Halldór Janetzko, Svenja Simon, Klaus Neuhaus, and Daniel A. Keim. “*Visual Boosting in Pixel-Based Visualizations.*” Computer Graphics Forum 30, no. 3: 871-80, DOI: [10.1111/j.1467-8659.2011.01936.x](https://doi.org/10.1111/j.1467-8659.2011.01936.x), 2011.³

Please note that I will use “we” throughout this chapter instead of “I”, as this chapter is based on publications¹.

¹For the division of responsibilities and work, as well as a statement of contributions in these publications, see [Visual Analysis of Next-Generation Sequencing Data to Detect Overlapping Genes in Bacterial Genomes](#) (p. 10) and [Visual Boosting in Pixel-based Visualizations](#) (p. 12).

²The Institute of Electrical and Electronics Engineers (IEEE) is the copyright owner of this work [14] but, as an author, I am permitted to re-use the work of this publication (verbatim and derivative) for my personal use. Link to the published article in IEEE Xplore: <http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6094047>

³I own (with the co-authors) the copyright of this publication, EUROGRAPHICS and Blackwell Publishing, hold the exclusive license for publishing ([10]). The definitive version is available at <http://diglib.eg.org/> and www.blackwell-synergy.com.

Direct link to the published article: <http://diglib.eg.org/handle/10.1111/v30i3pp0871-0880>

IV. VISUAL ANALYSIS FOR THE TRUSTWORTHINESS ASSESSMENT OF RNASEQ MEASUREMENTS

IV-1 Introduction

RNA Sequencing (RNAseq) by next-generation-sequencing (NGS) allows measuring (indirectly) all RNA sequences in a high-throughput fashion. The sequencing of all transcribed genetic elements allows not only to analyze known genes but also to discover new genes as well as regulatory (non-coding) RNA. However, analyzing the large amounts of data generated in NGS is a serious challenge which requires novel data analysis and visualization methods to support the discovery new genes. Current genome browsers do not follow the visualization design guidelines and hamper an expressive and effective reading of the visual design. The exploration in genome browsers is, furthermore, restricted to time consuming and non-targeted browsing in the genome. Users cannot filter directly for all genes of interest based on filtering parameters.

Due to uncertainty issues in the generation of RNAseq data, (see Section III-2.2), RNAseq measurements need to be verified before it can be concluded that an open reading frame (ORF) is, indeed, transcribed and active. For known genes, it is generally assumed that a measurement indicates the activity of the gene but in the case of gene candidates, the trustworthiness of measurements needs to be verified. Due to the large number of new gene candidates - in the range of tens of thousands - a visual inspection of all candidates by an expert is not feasible. Therefore, we provide a visual analysis solution that overcomes the shortcomings of genome browsers and addresses the issues of “visual design” and “filtering” in this chapter. Filtering is complex since RNAseq data suffers under many biases sources. Thus, no strict thresholds can be defined for the activity of an ORF (see also Section III-4.2) and an expert judgment is required to adjust parameters.

After stating the requirements derived from the task descriptions in Section III-4.2, we will discuss a Pixel-based Representation of RNAseq Reads Coverage and introduce The NGS Overlap Searcher - An Enhanced Genome Browser as a system to assess the trustworthiness of RNAseq measurements.

The contributions of the NGS Overlap Searcher are:

- An effective and efficient representation of read coverage without introducing artifacts.
- A visualization of RNAseq measurements in the open reading frames (ORFs) representation allowing to determine how well the region of read coverage fits to the ORF.
- A filter functionality to focus on interesting ORFs to handle the large volumes of data.
- An overview representation to adapt filter parameters based on visual feedback as well as to navigate to ORFs of interest.

The usefulness of the NGS Overlap Searcher is demonstrated with a case study in the area of overlapping gene detection.

IV-2 Requirements

The tasks regarding the aim to assess the trustworthiness of [RNAseq measurements](#) are:

- **T-I:** *Visual assessment of the trustworthiness of RNAseq measurements.*
- **T-II:** *Identification of [open reading frames \(ORFs\)](#) worthwhile a further inspection.*
- **T-IIa:** *... by increasing the scalability.*
- **T-IIb:** *... by filtering out discardable candidates.*

Based on these tasks we have derived the following requirements (see [III-4.2](#) for more details about the tasks).

R-I *System should resemble state-of-the-art tools for RNAseq data.* In order to reduce initial training, a solution should resemble state-of-the-art tools for RNAseq sequencing data that biologists are already acquainted to.

R-II *Expressive & effective visualization of the read coverage.* The visualization of [read coverage](#) has to support the assessment of the trustworthiness of [RNAseq measurements](#) effectively (**T-I**). Especially the representation of the trend is important here.

R-III *Visualizing the surround.* In order to assess where the RNA transcript of an ORF possibly starts and ends, not only read coverage of the ORF location but also of its surround needs to be visualized. Thus, at least the [untranslated region \(UTR\)](#) need to be covered. In order to assess if the read coverage of one ORF might originate from an adjacent ORF, a larger surround might be needed, which includes adjacent genes (**T-I**).

R-IV *Visualizing contextual information of ORF locations.* When the surround of an ORF is visualized, the representation of the ORF location and adjacent genes are needed as contextual information. This includes an effective assignment of [RNAseq measurements](#) to the ORF location, i.e., to assess how well the RNAseq measurement fits to the ORF region (**T-I**).

R-V *Visualizing read start positions.* In order to assess the distribution and number of reads at one location, the start position of mapping reads is important (**T-I**).

R-VI *Providing filtering capabilities.* Interesting sites need to be automatically determined according to a user-defined interestingness function, based on several parameters, to help users to deal with the large amount of data (**T-II**).

R-VII *Steering of the effects of parameter changes.* Several parameters need to be combined to define ORFs of interest. In order to adjust parameters users need to steer parameter changes (**T-II**).

IV. VISUAL ANALYSIS FOR THE TRUSTWORTHINESS ASSESSMENT OF RNASEQ MEASUREMENTS

R-VIII *Strand specific data representation.* In order to identify new genes and especially overlapping genes a strand specific representation of the read data is needed (T-I).

IV-3 State of the Art and Related Work

In general RNAseq data is visualized with so-called genome browsers. Genome browsers represent the whole genome in a linear fashion and align meta information and further information (e.g. measurements) to the genomic coordinates as so-called tracks which are vertically stacked (see Figure IV.1). Due to the large size of the genome, the current view of a genome browser always shows a segment of the genome. For navigation, users can zoom and browse in up- and down-stream direction. In order to inspect a gene of interest, users can search for the gene name. Examples for popular genome browsers are: the UCSC browser [Karolchik et al., 2014], the Integrative Genomics Viewer (IGV) [Thorvaldsdóttir et al., 2013], the Integrated Genome Browser (IGB) [Nicol et al., 2009] and Artemis [Rutherford et al., 2000, Carver et al., 2012].

Besides, there are genome browsers whose graphical representations aim at supporting special tasks. LookSeq [Manske and Kwiatkowski, 2009], for instance, uses a stack view which enables the identification of deletions and insertions by using paired-end reads. The Integrative Genomics Viewer (IGV) [Thorvaldsdóttir et al., 2013] integrates different data types and supports array-based and next-generation sequencing data as well as clinical and phenotypic data. The UCSC browser [Karolchik et al., 2014] is a web-based genome browser which also provides access to many publicly available data sets. In order to analyze SNPs, genome browsers typically visualize the differences compared to the reference genome. However, application specific genome browsers like the IBrowser [Afitos et al., 2015] have also been designed for SNP visualization. For further readings see Nielsen et. al. [Nielsen et al., 2010] who discuss techniques and challenges of visualizing genomes.

RNAseq measurements are shown in genome browsers either as stacked reads which represent each reads. Or the number of reads that overlap each genome position (read coverage) is represented either as line charts or bar charts. See Figure IV.3 for an example of visual representations of reads in the Artemis genome browser [Rutherford et al., 2000, Carver et al., 2012].

The advantage of the stacked read view is that not only the coverage at a specific position is visible but also the position of a specific read sequence. However, an issue are artificial gaps that emerge as artifacts between stacks. Thus, boundaries between genes cannot be determined accurately (see top of Figure IV.2). Further, the stacked read view is not scalable. If the number of reads mapping to one gene is large, not all reads can be represented (see Figure IV.3(a)). This can even happen if identical reads are aggregated as in Figure IV.3(a). In this case identical reads are aggregated to one green read. However, it is not clear how many reads are merged.

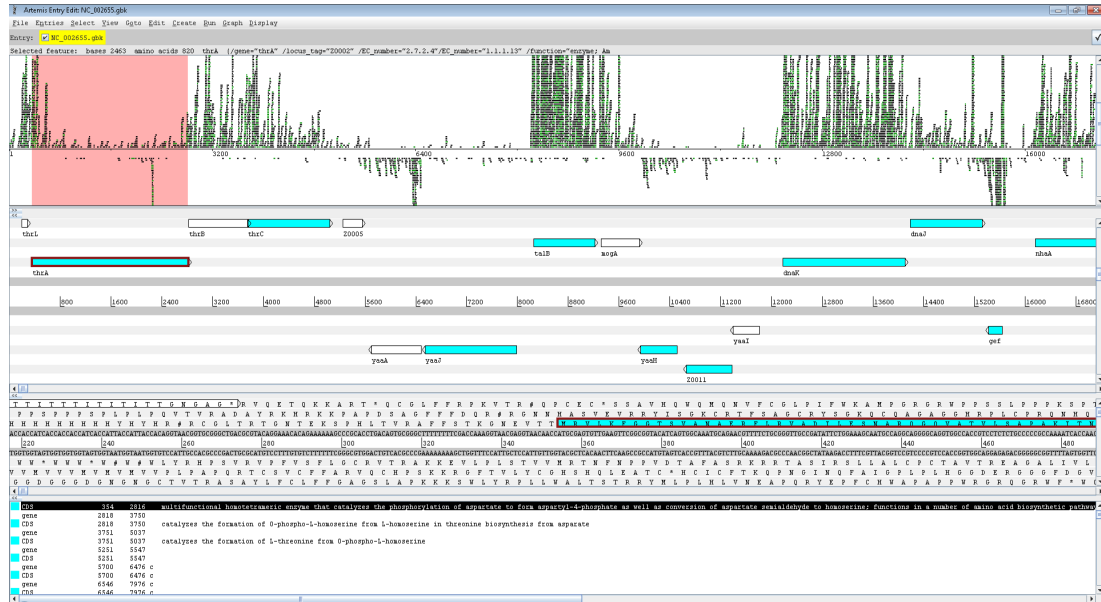


Figure IV.1: Screenshot of the genome browser Artemis [Rutherford et al., 2000, Carver et al., 2012]. Four tracks are shown. At the top the *RNAseq data track* is shown. Reads mapping to the reverse strand are drawn upside down below the middle line. The red region highlights a selected gene. Below the *RNAseq data track* the linked *annotation track* with annotated genes (cyan and white rectangles) in this gene segment is shown. The gene with the red border is selected. Arrows indicate the strand (right: forward strand, left: reverse strand). Below the *annotation track*, the zoomed-in part of the selected gene, is shown. This track shows the DNA sequence as well as its six-frame translation to amino acids. The track at the bottom shows the gene annotation information of the genes in the view. The selected gene is highlighted in black.

Line and bar charts have a better scalability, as they can be normalized for the available screen space (see Fig. IV.3(c)). However, genes can have diverse coverages with reads. Thus, a scaling to the gene with the highest coverage in the genome would suppress the visibility of line charts for many genes. An alternative is to scale the current view, like in the Artemis genome browser [Rutherford et al., 2000, Carver et al., 2012]. However, a gene with a low coverage might still be overlooked next to a highly covered gene. The change of the scaling is, furthermore, confusing and can lead to the wrong impression that two genes have a similar expression strength.

One issue with all genome browsers is that exploration is restricted to time consuming and non-targeted browsing in the genome. Users cannot directly filter for all genes of interest based on filtering parameters.

IV. VISUAL ANALYSIS FOR THE TRUSTWORTHINESS ASSESSMENT OF RNASEQ MEASUREMENTS

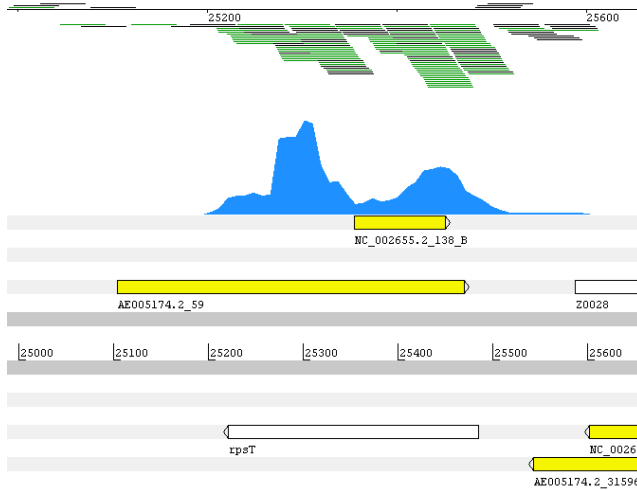
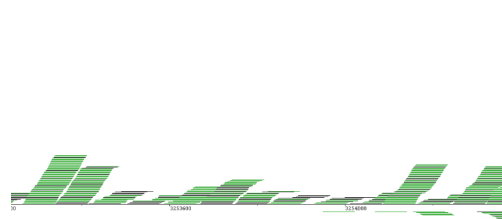
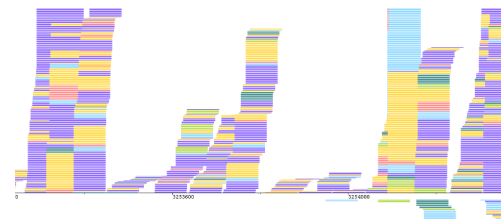


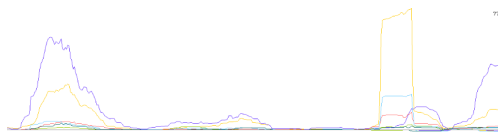
Figure IV.2: Stack view in the Artemis tool [Rutherford et al., 2000]. An annotated gene (rpsT) is irregularly covered by read stacks. Identical reads are merged to a single green read. Due to the stack representation artificial gaps emerge as artifacts between stacks. The corresponding coverage is displayed by the blue graph but not strand specific. Overlapping ORFs of ≥ 93 bp are displayed by yellow bars. This graphic appeared in [14] ©IEEE.



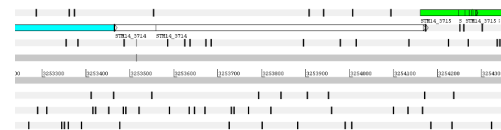
(a) Stacked Reads. Six conditions are shown at the same time. Identical reads are collapsed to one green read. Reads mapping to the forward strand are drawn above the middle line, reads mapping to the reverse strand are drawn below the middle line (see right corner).



(b) Stacked Reads. Identical reads are not collapsed. Color codes for the condition. The view is truncated and shows only part of the data compared to IV.3(c). Reads mapping to the reverse strand are drawn below the middle line (see right corner).



(c) Line Charts for each of the six conditions. View is scaled to the highest data value in the view. The height encodes the number of the overlapping reads on this strand. For the reads mapping to the reverse strand, the line chart is drawn bottom up below the middle line (see right corner).



(d) The annotation track shows the location of genes (white, green or cyan rectangles) as well as stop codons (vertical black lines). Above the middle line the forward strand is shown, below the reverse strand.

Figure IV.3: This graphic shows different visualization options of the genome browser Artemis [Rutherford et al., 2000, Carver et al., 2012] to represent the reads mapped to the genome (subfigures IV.3(a)-IV.3(c)). Subfigure IV.3(d) shows the annotation track.

IV-4 Pixel-based Representation of RNAseq Reads Coverage

Note

This section is based on the following publication and parts of this section appeared in this publication [10]¹:

[10]: Oelke, Daniela, Halldór Janetzko, Svenja Simon, Klaus Neuhaus, and Daniel A. Keim. “Visual Boosting in Pixel-Based Visualizations.” *Computer Graphics Forum* 30, no. 3: 871-80, DOI: [10.1111/j.1467-8659.2011.01936.x](https://doi.org/10.1111/j.1467-8659.2011.01936.x), 2011.²

Please note that I will use “we” throughout this chapter instead of “I”, as this chapter is based on a publication¹.

¹For the division of responsibilities and work, as well as a statement of contributions in this publication, see [Visual Boosting in Pixel-based Visualizations](#) (p. 12).

²I own (with the co-authors) the copyright of this publication, EUROGRAPHICS and Blackwell Publishing, hold the exclusive license for publishing ([10]). The definitive version is available at <http://diglib.eg.org/> and www.blackwell-synergy.com.

Direct link to the published article: <http://diglib.eg.org/handle/10.1111/v30i3pp0871-0880>

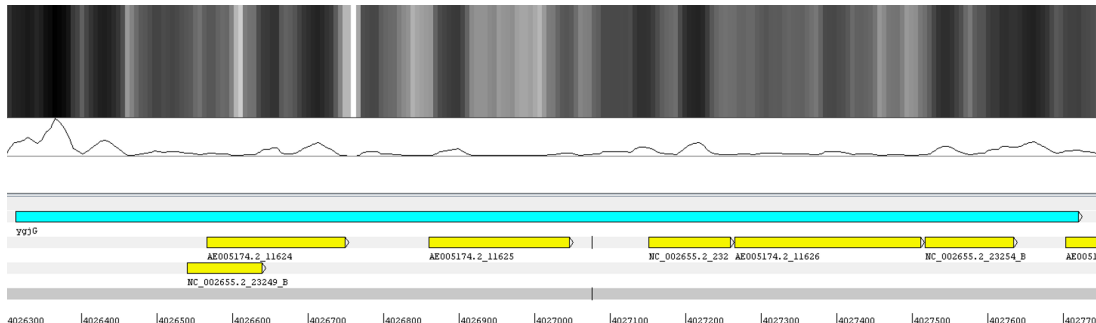
An important point in the visualization of **RNAseq measurements** is the scalability. Task **T-II** requires identifying **open reading frames (ORFs)** worthwhile a further inspection. A possible option to achieve this is to provide a scalable data representation which allows to quickly assess if an ORF is worthwhile a further inspection (**T-IIa**). In this section, we will address task **T-II** with scalable pixel-based visualizations.

Note that this work was preceding the *NGS Overlap Searcher* discussed in the next section. The prototype presented here was implemented with the focus on highly scalable visualizations to overcome the scalability issues of genome browsers. We came to the conclusion that the solution is a valuable extension to genome browsers in general, however, it is not sufficient for our specific task to visually assess the trustworthiness of RNASeq data. Therefore, we abstracted the shortcomings and lessons learned from this prototype to redefine the requirements and designed the *NGS Overlap Searcher* that is successfully applied for this domain problem.

The common representation in genome browsers uses a linear display. This leads to a high aggregation on zoom-out or a focus on a small fraction of the genome in the current view. The **reads** are either shown as stacked bars, line charts, bar charts or heatmaps (see Figure **IV.3**). Heatmaps can also be understood as pixel-based visualizations¹. The **read coverage** is represented by colored rectangles (see Figure **IV.4**).

¹Pixel-based visualizations refer to the use of a colored rectangle (mostly quadratic) as the unit of representation rather than a screen-pixel.

IV. VISUAL ANALYSIS FOR THE TRUSTWORTHINESS ASSESSMENT OF RNASEQ MEASUREMENTS



(a) Below the heatmap representation, the read coverage is shown as a line chart.



(b) Clipping of the heatmap.

Figure IV.4: These graphics show the heatmap representation of the read coverage in gray scales. Artemis genome browser [Rutherford et al., 2000, Carver et al., 2012].

The advantage of heatmaps is their high scalability in the y-axis direction compared to line charts and especially stacked bars (see Figure IV.4(b)). A disadvantage of color is the fact that the number of values which can be distinguished is limited. However, also the accuracy of line charts is limited by the available y-axis space. Due to the high value ranges of read coverage, genome browsers often normalize the data in the view to the data maximum of the view. A logarithmic scaling is an alternative which still allows a comparison of different regions.

In order to exploit the screen space efficiently, heatmaps are advantageous compared to line charts due to their higher scalability (see Figure IV.4(b)). Without taking genome annotation data into account, which needs a lot of space, screen-filling layout alternatives are possible. However, when ignoring the gene locations, a representation of read coverage loses context. We identify the gene location as the minimum information which is needed to interpret the data. In the following, we refer to pixel-based visualization instead of heatmaps.

In the paper Visual Boosting in Pixel-based Visualizations [10], we discuss several alternatives to use boosting in pixel-based visualizations to enhance relevant information and gave guidance when to use which technique¹.

We identified the following methods to boost information in pixel-based visualizations:

- halos: add a colored surround to a pixel to highlight it.

¹The main research idea, to address the question how to boost interesting and important information in pixel-based visualizations and to provide a guideline, was identified by Daniela Oelke. Daniela Oelke contributed the text application scenarios, Halldór Janetzko contributed the geospatial use case and I contributed the biological usage scenario. Possible boosting techniques were collected in discussions. The guidance, which technique works best in which usage context, is also based on discussions. I suggested thereby the distinction between image-driven and data-driven boosting for the comparison of boosting techniques. Klaus Neuhaus and Daniel Keim helped with fruitful discussions and advises. See also work distribution in [Visual Boosting in Pixel-based Visualizations](#) (p. 12).

IV-4 Pixel-based Representation of RNAseq Reads Coverage

- color: an appropriate color map highlights values of interest, e.g., high values.
- distortion: pixels of important resp. unimportant values are widened resp. narrowed.
- hatching: important pixels are highlighted with a hatching.
- shapes: important pixels are marked with a shape, e.g., an arrow.

All boosting techniques have pros and cons depending on the data (sparse or dense) and the definition of interestingness (image-driven or data-driven). Either a specific data range is of interest, for instance, the highest data values, then the information is already given in the image. Or the interestingness is defined by meta information, for example, the information where a gene starts and ends. In this case the information is not given in the image. See [10] for a discussion of boosting techniques and usage guidelines.

Read coverage data is an example where data-driven boosting is needed. Color cannot be applied in this case as the read coverage needs to be encoded with color already. As a whole passage needs to be boosted, also halos are no option, as they necessitate surrounding blank pixels which is per definition not given for a passage. The same holds for shapes which extend the size of the pixel. The passage should be perceived as a continuity, therefore, only shapes fulfilling this requirement would be possible. However, to overlay the pixels as little as possible, we decided to use hatching with one dash as well as distortion. As scalability is important for visualizing read coverage, we narrowed unimportant nucleotides which we define as **nucleotides** in intergenic regions without read coverage.

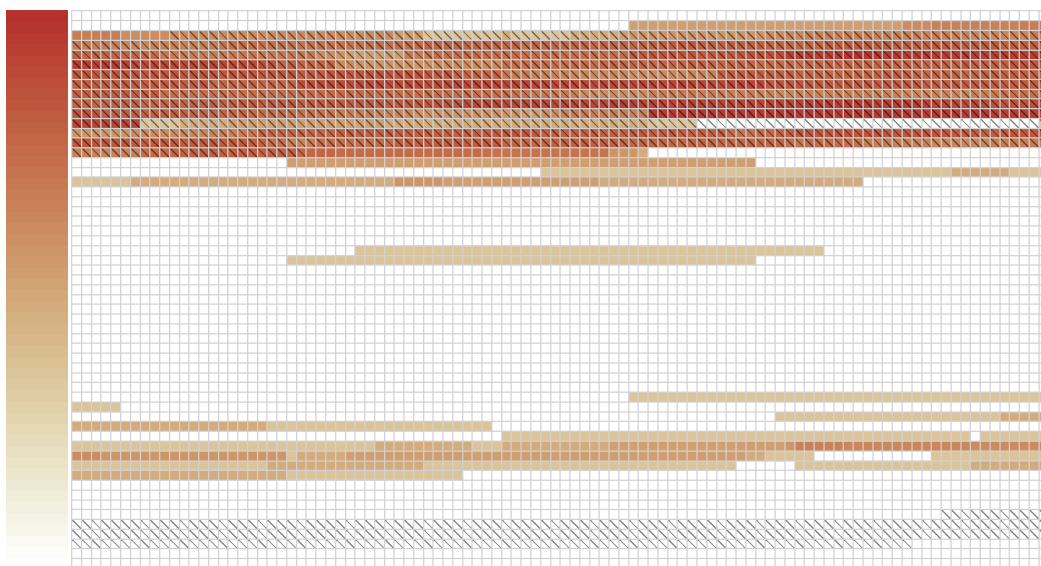
See Figure IV.6 for a boosting of gene regions with a horizontal dash “-”. Figure IV.6(a) without distortion and Figure IV.6(b) with distortion. Pixel borders are drawn in a light gray here, leading to a continuous impression of read coverage values (a closer look is needed to perceive pixel borders). Figure IV.5 shows a boosting with a tilted dash - a backslash “\”. Figure IV.5(a) without distortion and Figure IV.5(b) with distortion and a diverging color scale. The diverging color scale highlights additionally the gene region with read coverage (red). Pixel borders are drawn in a darker gray here. Thus, pixels are easier to distinguish. In Figure IV.6(b), the distorted regions appear gray leading to a less pronounced pop-out effect of the gene without read coverage compared to the design in Figure IV.5(b). The examples in Figures IV.6 and IV.5 demonstrate the applicability of the boosting techniques hatching and distortion to highlight genes in a screen-filling pixel-based representation of read coverage data. Due to the good scalability R-III is given, and the hatching provides the contextual information (R-IV). However, even though the read coverage is well represented with color, the line breaks due to the limited number of nucleotides per line, hamper a coherent impression of the read coverage of genes. This contradicts with the expressiveness and effectiveness requirement (R-II).

Requirement R-V to visualize read starts could be fulfilled by applying using a 2D color map to represent read coverage and reads starts simultaneously or by visualizing the difference

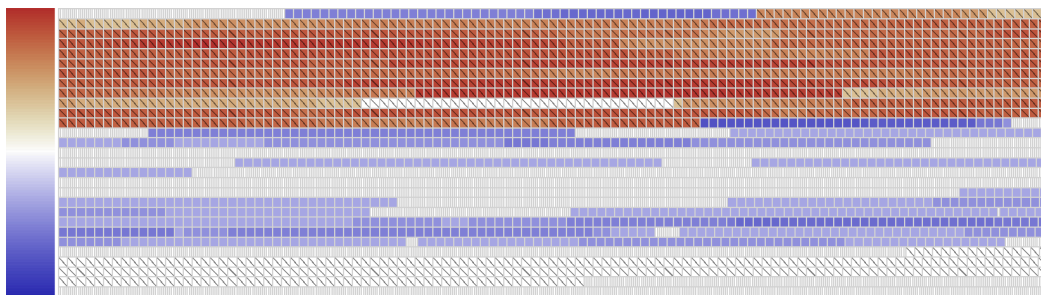
IV. VISUAL ANALYSIS FOR THE TRUSTWORTHINESS ASSESSMENT OF RNASEQ MEASUREMENTS

or ratio of both values instead. A strand specific data representation (R-VIII) could also be achieved with a 2D colormapping. The challenge with other representations of both strands is the feasibility of a mental mapping. In case of separate juxtaposed views for both strands, it would be hard to map locations. In case of paired lines of pixels for forward and reverse strand line breaks would hamper to keep track of the strands.

Due to all these reasons, this design alternative was discarded leading to a redesigned system that satisfies all these requirements: the *NGS Overlap Searcher* discussed in the next section.



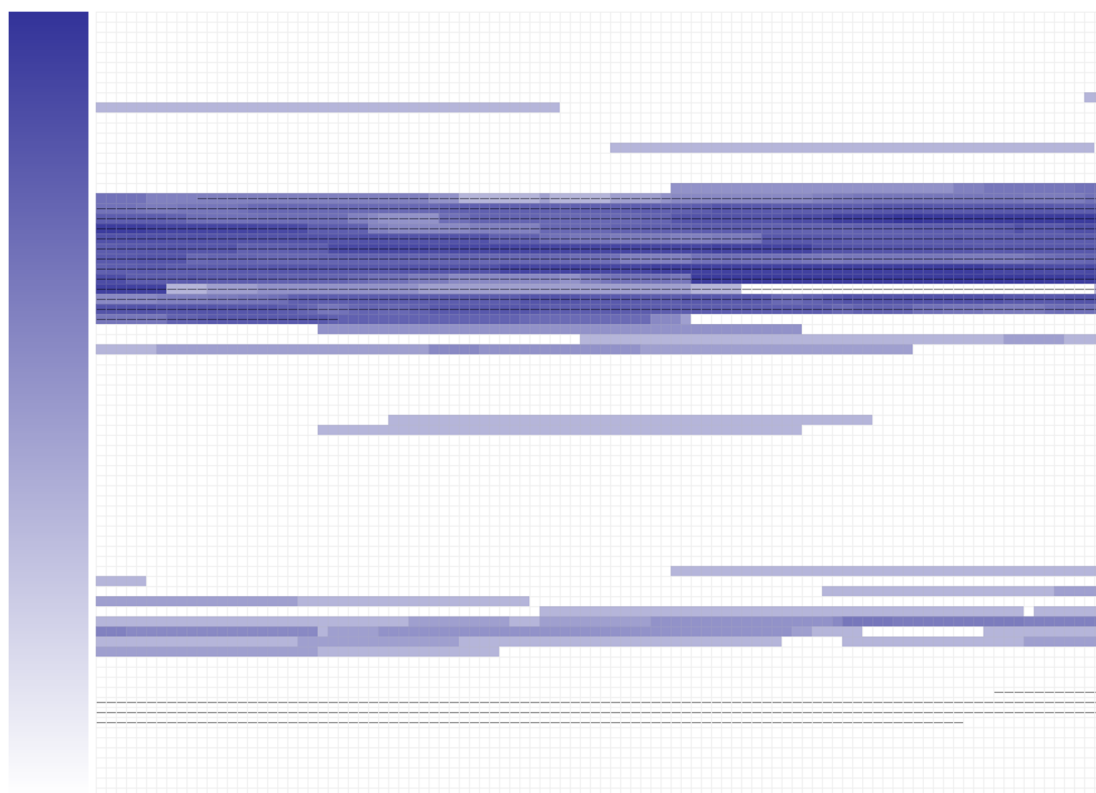
(a) Without distortion.



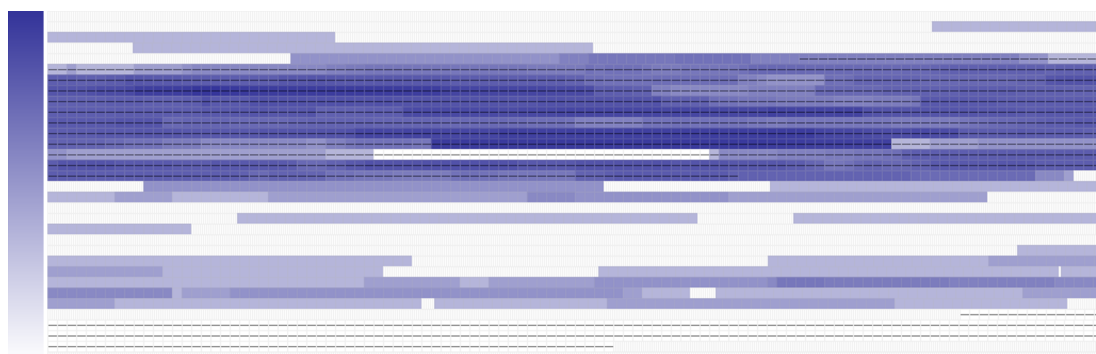
(b) Pixels of nucleotides, which are neither transcribed nor part of a gene, are distorted (reduced rectangle width).

Figure IV.5: Each nucleotide is represented by one pixel (rectangles with gray boarder). Nucleotides are arranged per line. (a) The continuous color scale encodes the read coverage (dark blue: high read coverage, white: no read coverage). (b) Two continuous color scales encode the read coverage (dark red: high read coverage within a gene, dark blue: high read coverage outside of genes, white: no read coverage). The regions of genes are marked with “\”. Two genes are shown. The upper one has read coverage, i.e., the gene is transcribed (red color). The lower one is not transcribed (white color). Between both genes some nucleotides are covered with reads. These graphics appeared in [10].

IV-4 Pixel-based Representation of RNAseq Reads Coverage



(a) Without distortion.



(b) The pixels representing nucleotides, which are neither transcribed nor part of a gene, are distorted (reduced width of the rectangles).

Figure IV.6: Each **nucleotide** is represented by one pixel (rectangles with light gray boarder). Nucleotides are arranged per line. The continuous color scale encodes the **read coverage** (dark blue: high read coverage, white: no read coverage). The regions of genes are marked with “-”. Two genes are shown. The upper one has read coverage, i.e., the gene is transcribed (blue color). The lower one is not transcribed (white color). Between both genes some nucleotides are covered with reads.

IV-5 The NGS Overlap Searcher - An Enhanced Genome Browser

Note

This section is based on the following publication and parts of this section appeared in this publication [14]¹:

[14]: Svenja Simon, Daniela Oelke, Richard Landstorfer, Klaus Neuhaus, and Daniel A. Keim. “Visual Analysis of Next-Generation Sequencing Data to Detect Overlapping Genes in Bacterial Genomes.” 2011 IEEE Symposium on Biological Data Visualization, October 23 - 24, Providence, Rhode Island, USA, 47-54, IEEE DOI: [10.1109/BioVis.2011.6094047](https://doi.org/10.1109/BioVis.2011.6094047), 2011.²

Please note that I will use “we” throughout this chapter instead of “I”, as this chapter is based on a publication¹.

¹For the division of responsibilities and work, as well as a statement of contributions in this publication, see [Visual Analysis of Next-Generation Sequencing Data to Detect Overlapping Genes in Bacterial Genomes](#) (p. 10).

²The Institute of Electrical and Electronics Engineers (IEEE) is the copyright owner of this work [14] but, as an author, I am permitted to re-use the work of this publication (verbatim and derivative) for my personal use. Link to the published article in IEEE Xplore: <http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6094047>

IV-5.1 System Architecture

Requirement **R-I** requires designing a system which resembles state-of-the-art tools for RNAseq sequencing data. The state-of-the-art tools are genome browsers. In order to fulfill **R-I**, our system will resemble standard genome browsers and depict the genome as a linear sequence.

This design also fulfills, in general, the requirements **R-III** to visualize the surround and **R-IV** to visualize contextual information. However, as genome browsers are not designed for specific tasks, but for a broad applicability, the design is not tailored for the requirements stated in Section **IV-2**. Furthermore, the visual design choices of genome browsers do not follow information visualization design guidelines. See Section **IV-3** for a discussion of the disadvantages of the stacked read view. A further disadvantage is the high spatial distance between read data and ORF representations (see Figure **IV.1**).

The issue of the spatial distance violates **R-IV** as read coverage cannot be assigned to ORF locations efficiently. We, therefore, decided to represent the read coverage between the three reading frames of the forward and the three reading frames of the reverse strand. Read coverage is represented for the forward and reverse strand separately to fulfill **R-VIII**. Open reading frames are depicted as boxes and positioned in the corresponding reading frame. In contrast

IV-5 The NGS Overlap Searcher - An Enhanced Genome Browser

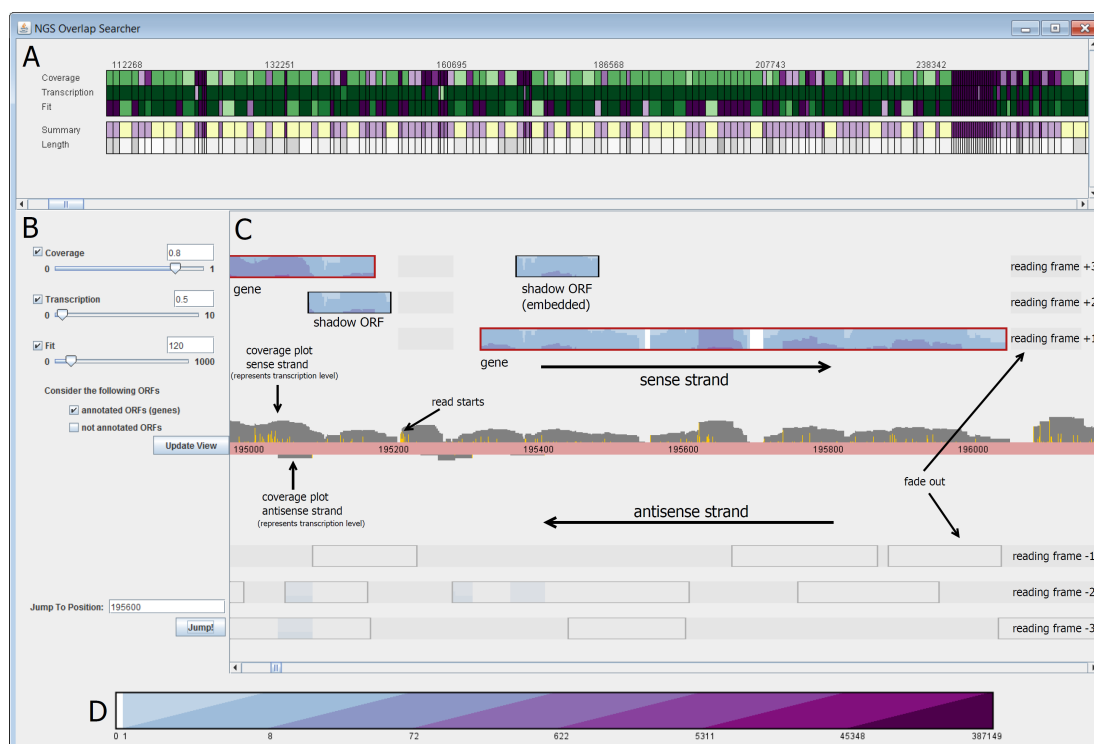


Figure IV.7: A shows the Genome Overview Bar. Every **open reading frame (ORF)** (here only annotated ORFs (genes)) is represented by a column. Each line represents one parameter of the interestingness function and the coloring of the cells encodes whether the given threshold was exceeded or not (see Fig. IV.9 for the color legend). Distortion is used to highlight interesting regions. In B the interestingness measure can be parameterized. Furthermore, the search may be restricted to genes or not-annotated ORFs only. The genome view C consists of a plot for the read coverage in the middle, plus the six reading frames of the sense and antisense strand. ORFs with a transcription value that does not allow them to exceed the transcription thresholds are faded out. D shows the color scale for the read coverage plot in the ORFs. In the plot, a gene is shown which can be considered as active since nearly its whole region is covered with reads. There are only two small gaps which decrease the coverage value (percentage of ORF covered with reads). This graphic appeared in [14] ©IEEE.

to most sequence viewers, our task requires depicting *all* ORFs (in our example with at least 93bp in length) and not just already annotated genes. In order to distinguish annotated from not-annotated, the former are shown with a red frame. See Figure IV.7 for a screenshot of the system.

In the following section, we will discuss how to visualize read coverage.

IV-5.2 Visualization of RNAseq Read Coverage

As described in Section IV-3, the stacked read view has several disadvantages, making it ineffective to represent **RNAseq measurements (read coverage)** (R-II). Line and bar charts

IV. VISUAL ANALYSIS FOR THE TRUSTWORTHINESS ASSESSMENT OF RNASEQ MEASUREMENTS

have a normalization issues. However, beside these issues, they are much more effective. As mentioned before, logarithmic scaling can be applied to solve the normalization issue. With a logarithmic scaling, the application of different scales between genome segments can be avoided and genes with a low read coverage can still be represented. Although reading the data values is more difficult with logarithmic scales than with (different) linear scales, the trends of the data can be efficiently compared (according to a single common scale) which allows to intuitively detect interesting patterns such as correlation between genome segments. Thus, line or bar charts with a logarithmic scaling fulfill R-II.

Following these considerations, we use a bar chart to represent the complete genome coverage and the coverage values are logarithmically scaled with a consistent scale for the whole genome (see Figure IV.7 C). As the information of the start of reads are, furthermore, important, we plot in orange the start positions of the reads in the bar chart to fulfill R-V. Highlighting the end of a read is not necessary in this case since all reads have the same length¹.

Considering R-IV and the required effective assignment of the [read coverage](#) to [open reading frame \(ORF\)](#) locations, we decided to additionally map the read coverage values directly to the ORF representations (rectangles above the read coverage bar chart in Figure IV.7 C).

From the perspective of visualization, the challenge is to find a representation that permits fitting the [read coverage](#) line chart (that reveals if an ORF is active) directly into the rectangles that represent the ORFs. Coloring the whole rectangle according to [gene activity level](#) of the ORF would cause a loss of necessary information, e.g., if the complete ORF is covered with reads or not. Standard line charts do not work as well because the space in the y-direction of the graph is too limited to depict the read coverage fluctuations truthfully.

The solution to this is two-tone coloring [[Saito et al., 2005](#), [Heer et al., 2009](#)], in which each value is represented by two discrete colors (see color map in Fig. IV.7 D). Using this technique, values can be read quite precisely even if not much space is available for drawing. By directly showing the read coverage color coded within the ORF rectangles (see Figure IV.7 D), we are able to reduce the mental effort to determine if an ORF fits to the region with read coverage or not. Since we are specifically interested in [overlapping genes \(OLGs\)](#), that are typically not annotated (without a red border in Figure IV.7 C), it is not only the question if an ORF fits to a region with read coverage but also to which of a few ORFs the read coverage fits best. Therefore, correct data interpretation is critical. Because the read coverage is rugged and uneven, inspection by an expert is mandatory.

Only the combination of the ORF rectangles and the read coverage bar chart can satisfy the requirement R-IV. The bar chart alone would not ease the mental load of mapping the data to the

¹Depending on the experimental protocol (see Section III-2.2) reads can also have variable length. However, in the [FOG-Project](#), reads have always the same length.

rectangles. Otherwise, coloring only the ORF rectangles would camouflage the fact that often not only the area of a specific ORF is transcribed but part of the sequence before and after it as well (UTR). We also considered coloring the background of the chart instead of using bar charts to encode this information. However, this would have interfered with the two-tone coloring that is used in the ORF representation. Furthermore, the bar chart comes with the advantage that it is a common way of visualizing sequencing data.

IV-5.3 Providing an Interestingness Function

NGS experiments provide vast amounts of data which complicates data analysis. Due to technical (random sampling of reads) and biological reasons (background transcription, multiple promoter sites, etc.) the [read coverage](#) always appears in rugged course of values (see also Section [III-2.2](#)). For genes with low expression and rare RNA transcripts, it is difficult to tell whether a gene is or is not transcribed and active. In order to differentiate the following criteria are taken into account by experts:

- **Coverage:** Percentage of bases of an ORF with a [count](#) of at least one. A low value means either coverage by only a few reads (background transcription) or overlap with a read coverage from the [untranslated region \(UTR\)](#) of an adjacent ORF. A high coverage value indicates a good fit of a certain ORF to the region with a read coverage.
- **Transcription:** Average number of [counts](#) of the bases of an ORF. In order to ensure that the numbers for different experimental conditions are comparable, this value has to be normalized¹. The higher the transcription value², the more likely it is that the ORF was indeed transcribed to RNA and, therefore, active.
- **Fit:** Absolute value of the difference of the transcript length (region with [read coverage](#)) and the ORF length. In case of annotated ORFs (genes), a high fit value indicates that this gene is part of an operon; in case of not-annotated ORFs, which overlap a gene on the same strand, this is an indication that the coverage is only due to the [untranslated region \(UTR\)](#) of the gene (which, in turn, would decrease coverage).

All three criteria can be parameterized and thresholds can be adaptively defined by an expert (Fig. [IV.7 B](#)). Combining these three criteria, we set up an interestingness function to highlight interesting ORFs ([R-VI](#)). Any other ORF, that does not satisfy the thresholds, is faded out. Alternatively, we may also restrict the analysis to regions with annotated or not-annotated ORFs. Furthermore, we distinguish between the two reading directions because the read coverage is available for both strands separately ([R-VIII](#)).

¹Normalized by the *total number of counts*: sum of counts over all genome positions, less rRNA reads.

²A type of a [gene activity level](#)

IV. VISUAL ANALYSIS FOR THE TRUSTWORTHINESS ASSESSMENT OF RNASEQ MEASUREMENTS

IV-5.3.1 Genome Overview

Having a flexible and expressive interestingness function is beneficial since the automatic filter focuses on interesting areas that need further inspection (all other regions are filtered out). Still, scrolling through the whole genome sequence would be necessary. Displaying only regions of interest would reduce the amount of data but with the expense of losing context information (violating R-IV).

Our overview representation of the genome reveals for all [open reading frames \(ORFs\)](#) all values of the interestingness function (see Figure IV.7 A). In this tabular view each horizontal line represents a single parameter of the interestingness function (coverage, transcription, and fit). Each column represents an interesting ORF (only annotated ORFs (genes), only not-annotated ORFs or both) of the genome. A cell is colored in a range of green to purple. The more saturated the green is, the more the threshold was exceeded. Similarly, intensifying shades of purple encode increased fail of a threshold (see Figure IV.9(a)).

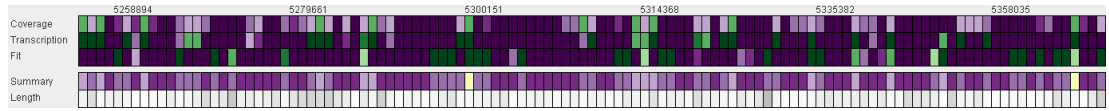
A summary line below reduces this information to a single cell to enable an extremely quick overview and a steering of the effects of parameter changes. Again, sections failing one or more thresholds are shown in shades of purple. Regions exceeding all thresholds are displayed in yellow (see Fig. IV.9(b)). Additionally, bars are distorted such that columns which satisfy more criteria are assigned more space. Because the length of the different ORFs in each column varies significantly and longer ORFs are often considered as more interesting, the length of each region is encoded in a gray scale below. Alternatively, length could also be used as the variable that determines the distortion factor of a column or as a further filtering parameter. By clicking on a column in the overview bar the corresponding ORF is centered in the genome view. Thus, the overview bar allows to browse the genome faster and in a more focused and task specific way.

IV-5.3.2 Genome Overview Bar to Steer Effects of Parameter Changes.

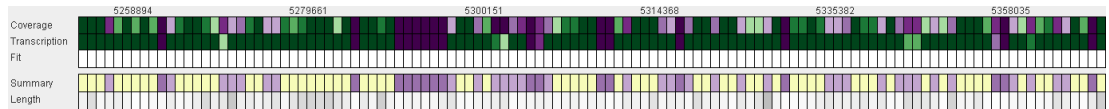
Because random sampling is involved in the sequencing process, the resulting [read coverage](#) is inevitably rugged. For the same reason also gaps have to be expected, even in regions of clearly active and transcribed genes. The probability of gaps depends on the sequencing depth in general and the [gene activity level](#) in particular. The latter depends on the experimental conditions. Consequently, it is not possible to specify default threshold for filtering. Thus, every analysis process starts with the challenge of choosing meaningful parameter values. This is an interactive process that imperatively needs an expert analyst with some experience in evaluating next generation sequencing data. Our genome overview representation supports this task.

Figure IV.8(a) shows part of the resulting display when the expected coverage is set to 100% and the threshold for the transcription value is 10. Furthermore, the threshold value for the fit is

IV-5 The NGS Overlap Searcher - An Enhanced Genome Browser



(a) Overview for a stringent parameter setting: Only two genes in this subsection of the Genome Overview Bar exceeded all three thresholds (shown in light yellow in the summary row). Some of the cells in the first line are colored in bright purple. This suggests that these regions are only slightly below the threshold for this criterion. Similarly, for the ones in bright purple in the summary row, only a single criterion was below the threshold, hindering them to get through.



(b) Overview of a relaxed parameter setting not using the fit criterion. The majority of genes exceed these thresholds.

Figure IV.8: *Genome Overview Bar* for different parameter settings. Only genes are shown and distortion is not applied. These graphics appeared in [14] ©IEEE.

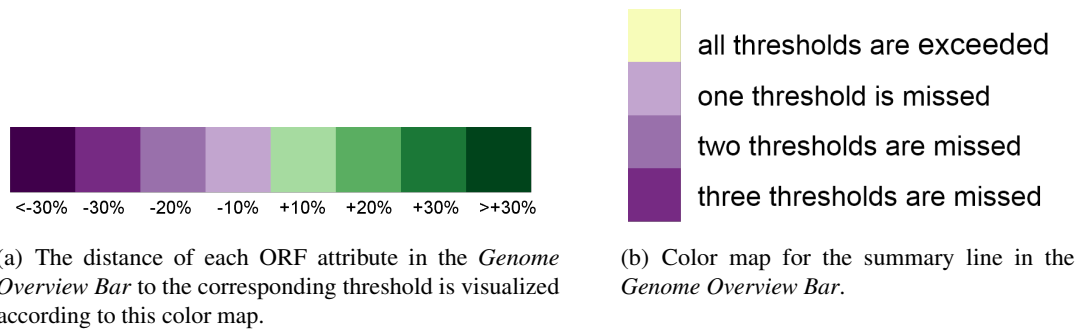


Figure IV.9: Color Legend of the *Genome Overview Bar* (see Fig. IV.8) in the *NGS overlap searcher* (see Fig. IV.7).

set to 45 nucleotides. Overall these are quite stringent settings. Consequently, only few regions in the genome are able to exceed all three thresholds.

However, through the coloring (many cells in this line are in light purple) it becomes apparent that slightly lowering the coverage threshold would let quite a couple of more regions exceed this threshold. On the other hand, it might be more advisable to obtain a couple of very good hits that can be tested in wet lab experiments which are time consuming. Thus, depending on the goals, the rate of false positive or false negative hits can be adjusted. E.g., it might be of interest to get an impression of the amount of not-annotated ORFs which are transcribed and active. This could be achieved by lowering the thresholds to not miss too many interesting ORFs. In order to further support this task not-annotated ORFs can be ignored. This way, adequate parameters can be assessed according to genome regions already better researched.

IV. VISUAL ANALYSIS FOR THE TRUSTWORTHINESS ASSESSMENT OF RNASEQ MEASUREMENTS

The influence of adjusted parameter settings can be viewed in the genome overview bar. By clicking on a column in the genome overview bar, the genome view jumps to the corresponding ORF. In this way promising ORFs can be investigated but also columns representing ORFs that have not exceeded the thresholds can be inspected for a better understanding of the parameter settings.

It is also possible to exclude parameters from the search. In Figure IV.8(b) the minimum coverage was set to 60% and 'transcription' to 0.5. The parameter 'fit' could be excluded as bacteria can have genes that are located very close to each other. For those cases, the fit value is not meaningful anymore. On the one hand, to find transcribed not-annotated ORFs that overlap with a known gene, the fit value is very helpful. On the other hand, cases in which the read coverage of an ORF is actually part of the gene it overlaps, would lead to many false positive hits.

IV-5.4 Evaluation

We have evaluated the *NGS Overlap Searcher* with a case study searching for [overlapping genes \(OLGs\)](#). Many not-annotated [open reading frames \(ORFs\)](#) overlap with a known gene in a different frame. How many of them encode proteins is debated but surely more than anticipated before (see also [Overlapping Genes](#) (p. 32)). Our goal is to find such incidences by analyzing the [RNAseq measurements](#) of a genome under different experimental conditions. Thus, we are searching for transcribed ORFs¹ that are overlapping with annotated genes².

With the help of the interestingness measure, it is easy to locate such regions in the genome. However, false positives appear as well. In most of all cases where the [read coverage](#) covers a "same strand overlapping ORF", the read coverage belongs to the gene (annotated ORF) and not the overlapping "same strand ORF" (see Figure IV.10 for an example). Thus, "same strand overlapping ORFs" are excluded.

For the analysis, the thresholds were set as follows: Coverage = 80%, Transcription = 0.5, Fit = 120 (see Figure IV.13). It can easily be seen in the summary line of the Genome Overview Bar that only few regions in the genome are able to exceed all thresholds. Next, the highlighted regions can be inspected one-by-one by clicking on them, to assess if they are indeed meaningful.

The example in Figure IV.11 shows an ORF that does meet all the criteria. It is located in the shadow of a large gene that did not exceed all the thresholds and, thus, was classified as inactive by our algorithm. A closer inspection reveals that there are many islands of [read coverage](#) which

¹An ORF is transcribed if it has a [read coverage](#) which is trustworthy.

²Note that not every transcribed ORF necessarily also encodes a protein but it provides some evidence that this might be the case.

are separated by regions without read coverage (gaps). However, the total trend of the read coverage suggests that the gene is actually active and not one of the overlapping ORFs. It is due to the gaps in the read coverage that the ORF was also not filtered out by the fit criterion.

Figure IV.13 shows an example for a region that might indeed contain an overlapping gene. Three not-annotated ORFs that are encoded in the sense strand do meet the specified criteria. One of them, which would then overlap with the already known gene of the antisense strand, could indeed encode a protein. Further evidence can be gained by comparing the specific region for different experimental conditions tested, taking additional meta-data into account, and finally by testing the assumption in wet lab experiments. Figure IV.12 shows two further examples for promising findings.

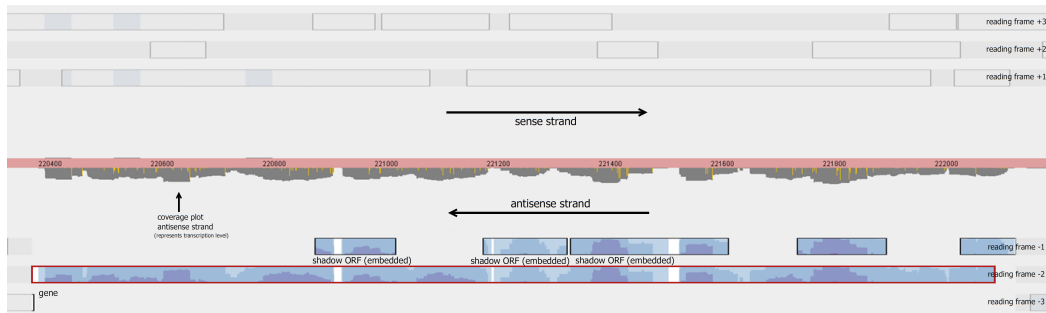


Figure IV.10: The read coverage clearly belongs to the gene (red bordered rectangle) and not to the overlapping ORFs (reading frame -1). This graphic appeared in [14] ©IEEE.

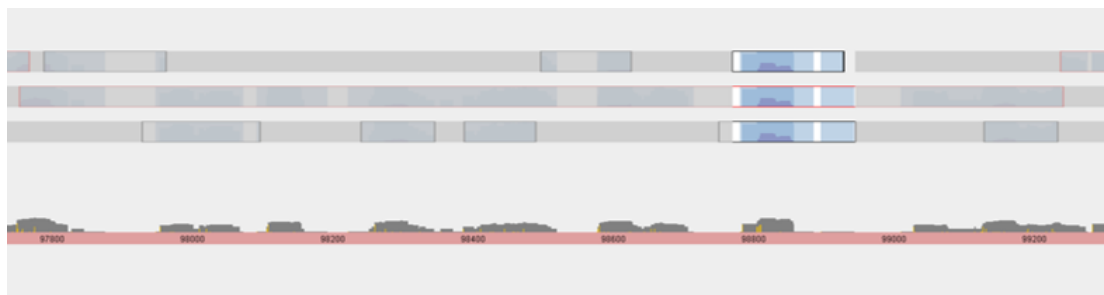


Figure IV.11: Example in which the read coverage on the sense strand belongs to a long gene (rectangle with the red border in reading frame +2) and not to the overlapping ORFs (reading frames +1 and +3). The gene did not exceed the thresholds due to many gaps in the read coverage.

IV-6 Discussion & Lessons Learned

Through our experiences in cooperating with biologists using the system, we have gained the following insights.

IV. VISUAL ANALYSIS FOR THE TRUSTWORTHINESS ASSESSMENT OF RNASEQ MEASUREMENTS

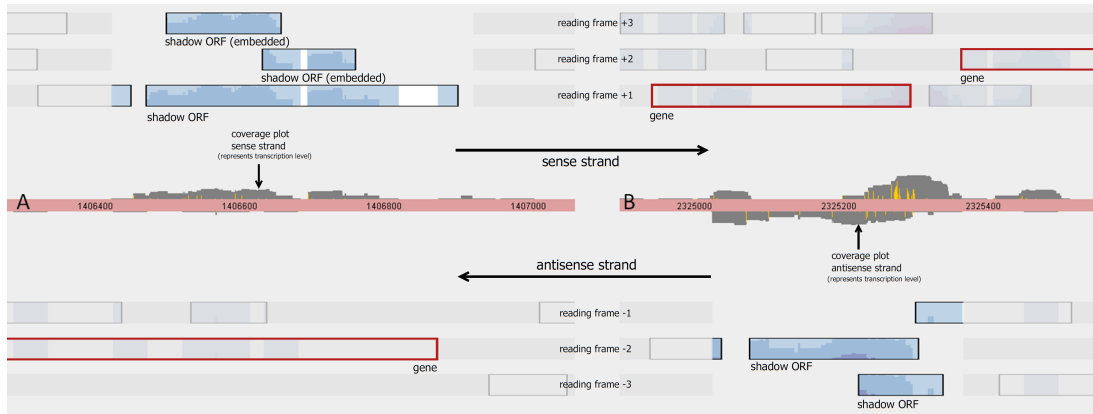


Figure IV.12: **A** On the sense strand two ORFs are shown which may be promising overlapping gene candidates for future wet lab experiments (in reading frames +1 and +3). The longer ORF in reading frame +1 might be the better candidate but the sloping coverage and the gap at the end of the ORF is indicative for the smaller ORF in reading frame +3. **B** On the antisense strand in reading frame -2 an ORF is shown which is completely covered with reads. Since this ORF is located opposite to a gene (in reading frame +1), this is also a promising candidate for future examination by wet lab experiments. This graphic appeared in [14] ©IEEE.

Comparison between different experimental conditions Beside the assessment of the trustworthiness of measurements, a comparison of different experimental conditions would be important to judge if an transcribed ORF is protein-coding or not. An overlapping gene might be found weakly expressed under one condition but highly expressed under another condition. A differential expression analysis of multiple RNA-seq experiments would be beneficial. Based on the gained insights, we developed the *VisExpress* System (see next Chapter V).

Evaluation When working with biologists in the course of the *NGS overlap searcher*, it turned out that it is difficult for them to describe their course of action when deciding on whether an ORF should be considered as transcribed or not. Some important criteria only became clear, when working together with them and discussing their analysis results. These experiences led to the definition of the *Liaison* role (see Chapter II). Subsequent and additional evaluations and tasks analysis lead to the *VisExpress* System (see next Chapter V).

IV-7 Limitations & Future Work

Finding appropriate thresholds Setting the right thresholds in the interestingness function is difficult but critical for the analysis. Few NGS transcriptional studies have been published so far and general thresholds have not been established. Future wet lab confirmations might form a feedback-loop which helps to determine meaningful thresholds.



Figure IV.13: The region of the ORF in reading frame +3 is almost completely covered with reads. There are no genes on the same strand which could potentially be responsible for this read coverage. On the antisense strand we can see a gene (red bordered rectangle) which overlaps this ORF. Thus, this ORF would be a promising candidate (potential **overlapping genes (OLGs)**) for future examination by wet lab experiments. This graphic appeared in [14] ©IEEE.

Scalability Finally, the scalability of the tool is an important issue. In all genome analysis projects, long linear sequences have to be processed which are difficult to display. In our research, we address this problem by an overview representation which also eases navigation. However, the genome overview bar still can not display all genes at once. A data aggregation of ORFs with similar values might lead to an improvement here.

Including additional data sources Additional meta data will support the analysis. If some regions are found to carry more overlapping gene transcripts than others, it would be important to see, for example, if these regions belong to genome-integrated bacterial viruses (prophages) or "normal" genome regions. Furthermore, ORFs that have a significant BLAST hit, are more likely to indeed encode a functional protein. Other protein identifying features, such as Shine-Dalgarno

IV. VISUAL ANALYSIS FOR THE TRUSTWORTHINESS ASSESSMENT OF RNASEQ MEASUREMENTS

sequence, promoters, terminators, regions with signal peptides or a good secondary structure prediction would also support the hypothesis of a new gene.

Chapter V

Visual Analysis of Differential Gene Expression

Note

Parts of this chapter will appear in the following publication [13]¹:

[13]: Svenja Simon, Sebastian Mittelstädt, BC Kwon, Andread Stoffel, Richard Landstorfer, Klaus Neuhaus, Anna Mühlig, Siegfried Scherer, and Daniel A. Keim. “*VisExpress - Visual Exploration of Differential Gene Expression Data*.” *Information Visualization*, 1-26, DOI: [10.1177/1473871615612883](https://doi.org/10.1177/1473871615612883), Published online before print December 14, 2015.²

Please note that I will use “we” throughout this chapter instead of “I”, as this chapter is based on a publication¹. “I” will only be used to refer to my role as an experimenter in the pair analytics study or my role as a *Liaison*.

¹For the division of responsibilities and work, as well as a statement of contributions in this publication, see [VisExpress - Visual Exploration of Differential Gene Expression Data](#) (p. 11).

²I own (with the co-authors) the copyright of this publication. The SAGE Publications Ltd holds the sole and exclusive right and license for publishing ([13]). The definitive version is available at <http://ivi.sagepub.com/>
Direct link to the published article: <http://dx.doi.org/10.1177/1473871615612883>

V-1 Introduction

Biologists are keen to understand the processes in bacteria in detail and how these processes react to environmental changes. Bacteria react to their environment, such as temperature, light, or food sources, by producing a variety of proteins. An understanding of proteins and cell processes supports, for instance, the understanding the **pathogenicity** of EHEC and is, therefore, of major interest. However, the functions of many proteins are still unknown and it is suspected that several protein-coding genes have been overlooked so far, for instance, overlapping genes but also short genes in inter-genetic regions.

RNA Sequencing (RNAseq) by **next-generation-sequencing (NGS)** has opened up the possibility to measure the whole transcription in a high-throughput fashion and to (indirectly) measure the protein activity level in cells under specific experimental conditions in parallel (see

V. VISUAL ANALYSIS OF DIFFERENTIAL GENE EXPRESSION

Figure V.1). Measuring the whole transcription is an advantage over DNA micro-arrays which can only measure known genes. Thus, RNAseq can also be used to identify new genes which have been overlooked so far.

In the last chapter, we discussed visualization and visual analysis techniques to analyze RNAseq from one experiment at a time. In this chapter, we address the combined analysis of several RNAseq experiments. Such an analysis has the advantage that first, a significant difference in the transcription of gene candidates is a strong indicator for functionality. Second, the relation of known genes to a specific condition allows inferences about the functional classification of genes. This analysis and comparison of [RNAseq measurements](#) from different experiments is named [differential gene expression](#) analysis.

Beside the possibilities of differential gene expression analysis, the analysis is also challenging. First, quality is an issue since the whole data generation process is error prone and introduces biases and uncertainties in the measurements (see Section III-2.2). A quality aware analysis to reduce false positive findings is, therefore, desirable. Second, scalability issues arise. The large number of genes and gene candidates is even further increased by the experiment comparisons. Third, a data perspective that focuses on all pairwise condition comparisons (n:n) instead of a condition to reference comparison (1:n), requires new visualization metaphors to allow a comprehensive view on the data. This involves an expressive overview and cognitively effortless recognition & interpretability of patterns. This last point is specific for the [FOG-Project](#) questions and is not covered in state-of-the-art visual analysis systems for differential gene expression data which focus on (1:n) comparisons (different conditions against a reference) only. In order to identify new genes, a set of non-standard conditions was analyzed in the FOG-project. Exploiting the full potential of the data set regarding the identification of new genes, also comparisons between these non-standard conditions are of interest, leading to (n:n) comparisons (each condition against all other conditions). Beyond the identification of new genes, (n:n) comparisons are of interest as, so far unknown, functional involvements can be revealed. Biologists can test, for example, the hypothesis that membrane proteins react similar to acid and nitrate stress.

After applying state-of-the-art analysis tools and performing a comprehensive literature search, we detected that currently no system meets the (n:n) comparison and quality awareness requirement. We, therefore, conducted a design study to build an interactive visualization system that covers these points, as well as the scalability issue. As a proof-of-concept the system was designed for known genes only. Extensions to incorporate gene candidates will be discussed in Section V-5.2. During the design study a VIS team of four visualization experts collaborated with three domain experts via a *Liaison* [12] to characterize the problem and to evaluate the system with a pair analytics [[Arias-Hernandez et al., 2011](#)] study on a real world data set (see

also Chapter II). From the visualization perspective, this problem domain provides an interesting and complex data exploration and hypotheses generation problem since expert hypotheses and background knowledge need to be integrated in the analysis process. The challenges for information visualization and visual analytics [Keim et al., 2012] are *scalability* due to the large amount of complex data and the challenge of *uncertainty* due to quality issues of the underlying data (see also Section I-4).

As the result of the design study, this chapter presents *VisExpress*. *VisExpress* uses a gene fingerprint visualization which allows a recognition & interpretability of patterns by (n:n) comparisons of experiments with low cognitive effort. Further, it integrates the data quality in the visual representation to address the uncertainty challenge. An expressive treemap-based overview supports the user to identify patterns, revealing connections, and generating new hypotheses in an overview and, thereby, reduces the analysis complexity by a divide-and-conquer approach which addresses the scalability challenge of the large volumes of [differential gene expression](#) data.

The three participants of the pair analytics study mentioned that the analysis of the real world data set would have required several days with the systems of their current use. With *VisExpress*, the domain experts got a comprehensive overview of the whole data set within an hour. Furthermore, they detected interesting findings and generated hypotheses for patterns that are easily overlooked by state-of-the-art systems. They identified the intuitive, comprehensive and quality aware overview as major improvements over the state-of-the-art.

In summary the contributions of the *VisExpress* are:

- The validated visualization design of *VisExpress*, based on an overview to detail visualization approach and *gene fingerprints* to explore [differential gene expression](#) data.
- A pair analytics study to validate the design of *VisExpress*.
- A discussion of the resulting biological findings.

This chapter is structured in the following way. First, we state the requirements derived from the tasks for differential gene expression analysis stated in Section III-4.3. Based on these requirements, we will discuss related work in the bioinformatics and biological visualization literature. In the following sections, we describe the design process, introduce the architecture of *VisExpress* and discuss the visual and interaction design decisions. The chapter is concluded with the evaluation, discussion and lessons learned, as well as limitations and future work.

V. VISUAL ANALYSIS OF DIFFERENTIAL GENE EXPRESSION

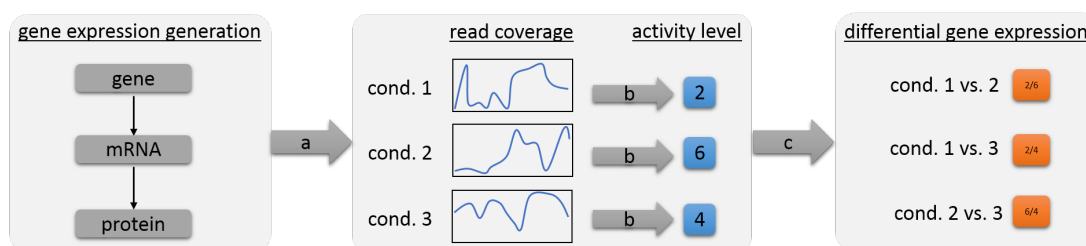


Figure V.1: Gene Expression is the production of proteins. Depending on the experimental condition, a larger or lower amount of specific proteins is needed. (a) Next-generation-sequencing is a method used to indirectly measure the amount of proteins in cells, by measuring the intermediate step (mRNA). Due to biases, the measured signal (read coverage) of a gene is ragged. (b) For further analysis steps the **read coverage** is expressed by a single normalized **gene activity level**. (c) The comparison of the **gene activity levels** is called **differential gene expression** and is expressed as the ratio (fold change) between conditions. Biologists use **differential gene expression** to relate genes with unknown functions with potential functions. See also Section **Biological Background** (p. 28).

V-2 Requirements

The tasks in the visual analysis of differential gene expression are:

- **T1:** *Generate hypotheses about the function of genes.*
- **T2:** *Test hypotheses about the function and reaction of genes.*
- **T3:** *Find genes related to a function.*
- **T4:** *Explore genes with unexpected **gene activity ratio (GAR)** patterns.*
- **T5:** *Relate new gene candidates to genes with known functions.*

Based on these tasks, we have derived the following requirements (see III-4.3 for more details about the tasks).

R0 *Interpret GAR patterns of genes.* Users need to *identify* the characteristics of the target gene which are expressed by **gene activity ratio (GAR) patterns**. A GAR pattern is the change of the activity levels of a gene under different experimental conditions. The representation of the activity ratios of a gene needs to allow the identification of each pairwise (n:n) comparison between conditions to interpret the GAR pattern (**T1**, **T2**, **T4**).

R1 *Compare GAR patterns of genes.* The tasks (**T1**, **T2**, **T3**, **T4**) require the ability to *compare* the GAR patterns of genes. Comparisons between single genes, between groups of genes, and between a single gene and a group of genes must be possible.

R2 *Summarize the functions of genes.* The system should be able to *summarize* the functions associated with a gene or a group of genes. When users identify an interesting gene or find a

group of genes with a similar GAR pattern, they need to know which functions are associated with them (T1, T2, T3, T4).

R3 *Explore genes according to GAR patterns.* The system should allow exploring the data to enable users to generate new hypotheses about genes (T1, T3, T4). The exploration should be guided by the GAR patterns to easily spot genes with similar behavior.

R4 *Support different comparison measures.* Different measures can be used to compare the activity level of genes that are based on different properties. The analysis results are more trustworthy if different measures produce similar analysis results.

R5 *Assess the trustworthiness of (automatic) results.* Automatic analysis results are useful to get an overview and to quickly come up with hypotheses but biologists do not trust them unconditionally. When they find an answer through the automatic evaluation, they want to assess the trustworthiness by analyzing the raw sequencing output and meta data by themselves, leading to several sub requirements (see [Detail: Gene Board](#) (p. 98)).

R6 *Highlight the quality of activity ratios.* According to our study, biologists do not trust automatic analysis results on the one hand; on the other hand they also want to reduce exploration space without loss of information. Therefore, they want to assess the quality of GAR patterns.

R7 *Highlight new gene candidates.* Genes candidates need to be highlighted to determine new genes and to set them in context with known genes with an annotated function.

VisExpress was designed according to R1-R6, as a proof-of-concept for genes only. We will discuss possible extensions to incorporate gene candidates and to fulfill R7 in Section [V-5.2](#).

V-3 State of the Art and Related Work

Gehlenborg *et al.* [[Gehlenborg et al., 2010](#)] provide a broad discussion of visualization systems for gene expression data. Many systems were established for (differential) gene expression data from DNA micro-arrays, e.g. TM4 and Mayday [[Saeed et al., 2003](#), [Battke et al., 2010](#)]. DNA micro-arrays used to be the state-of-the-art for gene expression before the rise of next-generation-sequencing (NGS) technologies and the possibility to sequence DNA in a cheap and high-throughput fashion without any pre-knowledge.

The state-of-the-art visualizations of (differential) gene expression data are heatmap-based visualizations (see [Fig. V.2](#)). Rows represent genes and columns encode experimental condition comparisons or the experiment data. Interactive heatmaps provide the possibility to select parts of the heatmap for further analysis (e.g., in INVEX [[Xia et al., 2013](#)]). Mayday

V. VISUAL ANALYSIS OF DIFFERENTIAL GENE EXPRESSION

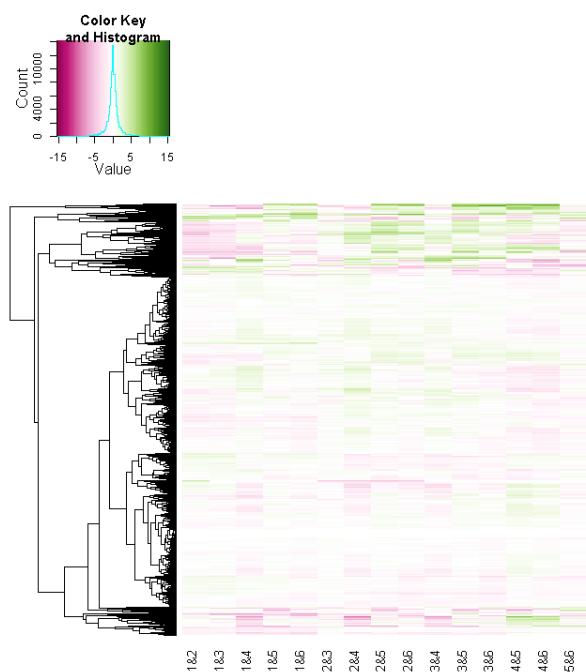


Figure V.2: State-of-the-art heatmap of the differential gene expression data used in this study (created with the R function `heatmap.2` [Warnes et al., 2014]). Genes are depicted by means of the rows and experimental conditions are illustrated by the columns. The clustering of rows is indicated by a dendrogram. All genes are included (around 5000). Two large clusters at the top and at bottom stand out. However, no clear pattern that separates the clusters or conditions stands out which increases the efforts of visual analysis. The colormap was adapted from ColorBrewer.org [Harrower and Brewer, 2003] (saturation: high gene expression ratio; white: low ratio; hue: direction).

[Battke et al., 2010] uses an enhanced heatmap which integrates metadata to emphasize relevant genes by, e.g., scaling of matrix rows and an additional color gradient [Gehlenborg et al., 2005].

Heatmaps are an appropriate and reasonable visualization for pure experiment data or if a set of conditions is compared to one reference (1:n comparison) which is the focus of many biological studies. However, pairwise comparisons of all experimental conditions (n:n comparison) are not well supported (see Section *Design rationale* on page 88). However, NGS technology advancements and falling costs lead to more and more complex experiment designs with (n:n) comparisons of different conditions. Furthermore, quality of the underlying data is not addressed sufficiently, if covered at all. Thus, a pre-processing or post-processing has to ensure quality. In our study, the analysis focuses on a quality aware (n:n) comparison and, therefore, the systems mentioned above cannot satisfy our requirements.

For gene expression time series data, parallel coordinates (profile plots) are often used to represent the changes over time. In order to analyze differences between clusters, these can be indicated by color-coding in one chart or by small multiples of parallel coordinates, such as in BiGGEsTS [Gonçalves et al., 2009] and Mayday [Battke et al., 2010]. MulteeSum supports the inspection of gene expression data not only over time but also in conjunction with the spatial cell location within an organism [Meyer et al., 2010a].

Clusterings are typically used in differential gene expression analysis to group genes with similar patterns (e.g., in [Saeed et al., 2003, Battke et al., 2010, Xia et al., 2013]). Different clustering

methods have been used and proposed on that account. In heatmaps the clustering is mostly indicated by an ordering of the genes based on clustering results and along with a dendrogram next to the heatmap (see Figure V.2). BicOverlapper [Santamaría et al., 2008] focuses on the visualization of biclustering results from gene expression matrices. Biclusters are represented as undirected complete subgraphs. Differential expression analysis and functional enrichments are added in BicOverlapper 2.0 [Santamaría et al., 2014].

Functional enrichment (or gene set enrichment) analysis is often a subsequent step after the identification of a set of potentially relevant genes (see [Hung et al., 2012] for an overview). An enrichment search refers to finding pathways or networks where a set of genes is significantly over-represented. BicOverlapper 2.0 [Santamaría et al., 2014] visualizes functional annotations of groups of genes as word clouds. Systems such as GENeVis [Westenberg et al., 2008] map gene expression data directly to networks. Gene expression is represented as bars inside network nodes (for an overview and alternatives see Gehlenborg *et al.* [Gehlenborg et al., 2010]). Pathline combines visualizations of multiple genes, time points, species, and pathways by introducing a linearized metabolic pathway representation and curve-maps representing the temporal expression data [Meyer et al., 2010b]. The data and focus of Pathline is different to our problem definition as we only analyze one bacteria species.

The pure visualization of a functional enrichment analysis or pathway analysis is not the focus of *VisExpress*. We focus on the visual exploration of differential gene expression patterns in relation to gene functions, providing quality awareness and (n:n) comparisons with expressive overviews and visual representations that allow a cognitively effortless recognition & interpretability of patterns. An integration of functional enrichment analysis will be part of future work.

V-4 The *VisExpress* System

V-4.1 Design Process

Deploying visualizations for real-world problems is problem-driven research. The aim of design studies is to abstract and/or generalize domain problems as well as designing visualization systems that are validated with real experts and real data. In this process, a collaboration with domain experts (real users) is vital. However, performing problem-driven research and working with domain experts can lead to many pitfalls. In order to avoid them, as well as to structure our design study project, we followed the nine-stage design study methodology framework of Sedlmair *et al.* [Sedlmair et al., 2012b] (see references therein for alternative approaches and a comparison of methodologies) which also lists 32 common pitfalls.

V. VISUAL ANALYSIS OF DIFFERENTIAL GENE EXPRESSION

Precondition phase.

This design study was conducted in the settings of a well-established, long-term cooperation between me and a group of biologists from the [FOG-Project](#). The whole design study team consisted of a BIO (three front-line analysts) and a VIS team (four VIS experts; including me). I acted as a *Liaison* between the BIO and the rest of the VIS team (see Chapter II). Just the *Liaison* had contact with the BIO team to keep the rest of the VIS team independent.

Core phase.

Discover stage - problem characterization & abstraction. Starting with interviews and observations of the current workflows of the BIO team, I subsequently collected relevant state-of-the-art systems based on my professional expertise as a bioinformatician and VIS expert. In the second step, the drawbacks of these systems were discussed and the problem characterization was refined. In the third step, the VIS team discussed these, concretized tasks and requirements, and improved the problem abstraction. The *Liaison* ensured in the whole process that the problem abstraction was still valid from the domain users' perspectives.

Initial prototyping and expert feedback. We created a low-resolution prototype to receive feedback from the BIO team. This initial design enabled the BIO team to precisely point out important aspects that the system should cover which were translated and merged with the identified requirements.

Design refinements. Based on experts' feedback, we stepped back to the design phase. In order to fully exploit the expertise of the four VIS team members, we took the following approach to create and implement design ideas: 1) every team member created a set of alternative solutions as paper mock-ups; 2) these solutions were selected, merged and refined in a critique-and-creation round; 3) we discarded or refined ideas by evaluating them against tasks and requirements. This entire process iterated until all VIS team members were satisfied.

Formative assessment and final design implementation. In this process, the VIS team improved design details based upon formative assessment conducted by the *Liaison* with one member of the BIO team. Functionalities of the system were explained and demonstrated. The constructive feedback led to design improvements and an optimized user interface to resolve some usability issues.

Summative assessment and design refinement. For validation of our design, the *Liaison* performed a pair analytics study [[Arias-Hernandez et al., 2011](#)] with the BIO team in order to

verify our design decisions for target tasks. Based upon the evaluation results, we refined our system designs and reflected our findings.

V-4.2 Architecture of *VisExpress*

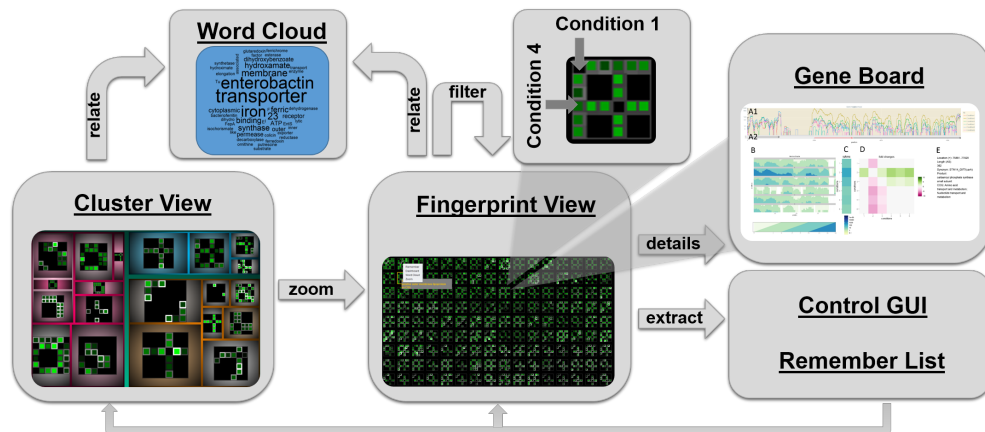


Figure V.3: Schematic work flow of the three views in *VisExpress* (based on the visual information seeking mantra of Shneiderman [Shneiderman, 1996] 'Overview first, zoom and filter, then details-on-demand'). A user can **overview** the whole data in the first level with a treemap that reveals the clusters in the data (*Cluster View*). By selecting a cluster in the treemap the user can **zoom** to the second level which overviews all gene fingerprints in one cluster (*Gene Fingerprint View*). Users can further **filter** out genes of interest and open them in a new *Gene Fingerprint View*. The third level gives **details-on-demand** about selected genes (*Gene Board*). Further, the user can **extract** interesting genes to a remember list for later analysis. In order to **relate** the gene fingerprints with gene functions the user can open a word cloud of gene functions as a further *details-on-demand* view. The user is also able to switch between different designs that support different analysis foci in the control GUI (see Figure V.10).

VisExpress is designed following the classical visual information seeking mantra of Shneiderman [Shneiderman, 1996] 'Overview first, zoom and filter, then details-on-demand' in order to support a divide and conquer approach for exploration of multiple genes but also for detailed investigations of genes of interest.

VisExpress uses matrix fingerprints to provide a visual summary of a gene in order to make gene activity ratio (GAR) patterns interpretable (R0; see Figure V.3). The matrix layout enables to visualize conditions as rows and columns and, therefore, reveals the activity of genes in different experimental conditions (n:n comparisons). The first-level of *VisExpress* (*Cluster View*) uses these fingerprints and word clouds to overview all clusters of similar genes in a treemap. This reveals common characteristics of the clusters (R1: comparison) as well as their biological functions (R2). The second-level (*Gene Fingerprint View*) visualizes all genes of a selected cluster in a scalable space filling layout for visual exploration of large amounts of genes (R3).

V. VISUAL ANALYSIS OF DIFFERENTIAL GENE EXPRESSION

The third-level (*Gene Board*) provides details—on—demand for single interesting genes. This view reveals detailed information related to the gene’s functions as well as gene activity level trends and allows manual assessment of findings (R5). The intended work-flow of *VisExpress* is illustrated in Figure V.3.

The three levels are seamlessly connected for smooth transition of analysis via a multiple view system. Each level can also be instanced multiple times with different data and settings. All instances are linked to a central instance which synchronizes the configuration of the designs and handles interactions between instances and levels (see also Figure V.10). The system’s visual components were implemented with JAVA Swing Components. An interface to R and Bioconductor [R Core Team, 2013, Gentleman et al., 2004] is used for preprocessing, statistical analysis, and machine learning algorithms.

The next sections will describe the following in detail: why and how we visualize **gene activity ratio (GAR) patterns** (Section V-4.3); the system components of *VisExpress* (Section V-4.4); and the user interaction design (Section V-4.5)).

V-4.3 Visualizing GAR Patterns

Biologists aim to generate and verify hypotheses about the behavior of genes. The main information units are, thereby, the **gene activity ratio (GAR) patterns** (focus of the tasks T1-T4). Heatmaps are the state-of-the-art for visualizing differential gene expression data (see [Gehlenborg et al., 2010] for an overview). Thereby, GAR patterns are represented as rows in heatmaps (see Figure V.2). Gene activity ratios are represented as color-coded pixels. All comparisons are shown next to each other and all genes are stacked horizontally. However, this representation supports requirements R0 (interpretability of GAR patterns) and R1 (comparison of GAR patterns) only partially:

1. A linear representation of GARs does not allow to directly identify the involved conditions (R0; see Fig. V.4 A1 & A2).
2. A linear representation of GARs does not sufficiently capture salient patterns (compare A1 & A2 with D2 in Fig. V.4).
3. It is hard to compare and explore genes (see Figure V.2) since single genes are hard to identify in a simultaneous representation of several thousand genes (R1, R3).

Fingerprinting

Based on these considerations, we decided to represent the **gene activity ratio (GAR) patterns** of each gene as a single entity (glyph) which we will name *gene fingerprint*. Our design goal

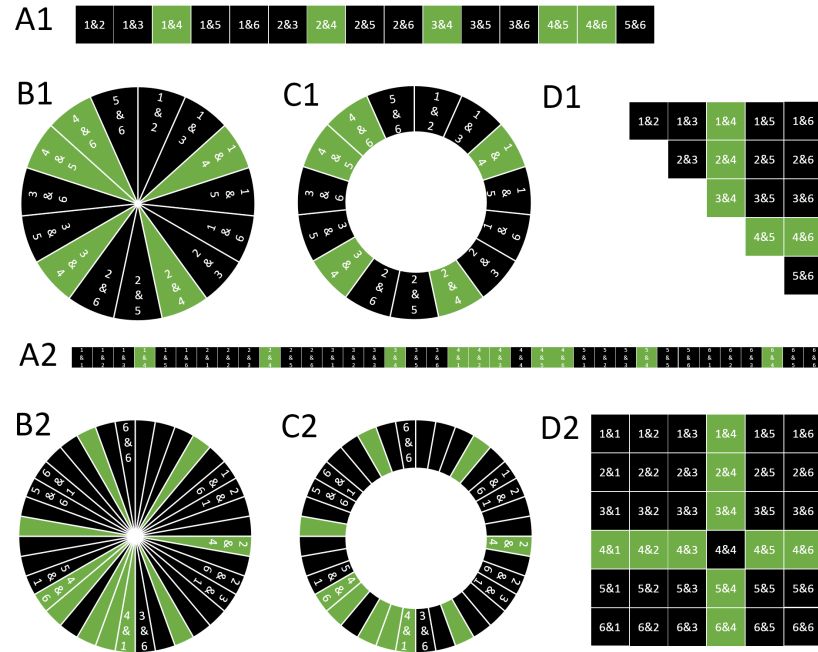


Figure V.4: Design alternatives for gene fingerprints. All sub-figures illustrate the same **gene activity ratio (GAR) pattern** of the pairwise comparison of six conditions (black: low value, green: high value). (1) shows all 15 unique comparisons and (2) all 36 comparisons with 6 conditions. (A) shows a linear ordering similar to a heatmap (see Figure V.2), (B) a circular layout, (C) a ring layout and (D) a matrix layout. In the illustrated data, condition 4 is different to all other conditions (which would be an important finding since this indicates that this gene and its function is related to this condition). This is hardly readable from (A), (B) and (C). Even though (B1) and (C2) show a pattern (black-green-black-green), the pattern is not interpretable and not salient. The pattern (condition 4 is different to all other conditions) is most salient in (D2).

of gene fingerprints is to provide a visual summary of a gene which can be used to compare the GAR patterns effectively (R1). The idea of fingerprinting is based upon the work of Keim and Oelke of literature fingerprinting [Keim and Oelke, 2007]. Each gene consists of a tuple of a gene activity ratio $r_{k,l}(g_i)$ and a quality $q_{k,l}(g_i)$ as well as functional description (plain text) for contextual information. Gene fingerprints should support identification and comparison of GAR patterns (R0, R1), and the assessment of quality (R6). Therefore, we discussed dividing the tuple into *measure* and *quality* in order to focus the visualization on the GAR measure.

The quality could be handled by threshold-filtering and/or details-on-demand such that only GAR patterns with a high quality are visualized. However, the BIO team preferred to see all genes and to perform quality-aware analysis (R6). Even patterns with low quality can be interesting and there is no fixed threshold that can define interestingness which rejects the idea of threshold-filtering. The challenge is to find visual metaphors that can encode both GAR value

V. VISUAL ANALYSIS OF DIFFERENTIAL GENE EXPRESSION

and quality and also satisfy R0, R1, R3 (interpret, compare and explore GAR patterns). In the following, we discuss design alternatives for gene fingerprints.

Design of Gene Fingerprints

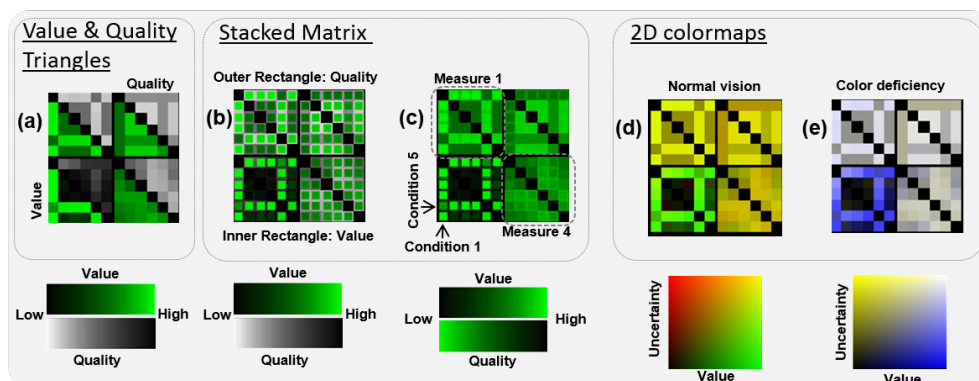


Figure V.5: Design alternatives for matrix visualizations of gene fingerprints. Four different measures to characterize a gene are illustrated for each design (see (c)). (a) Two triangular portions in a matrix representing the value (bottom left) and the quality (upper right) of a gene. (b) and (c) Stacked Matrices with inner and outer rectangles encoding value and quality, respectively. (d) and (e) Two dimensional color maps for normal and dichromatic visions, respectively. The color mapping in (c) highlights high values and low quality.

Due to the exploration requirement (R3), the visualization design has to be scalable. Highly scalable techniques are pixel-based visualizations such as Recursive Patterns [Keim et al., 1995] or Pixel Bar Charts [Keim et al., 2001]. Therefore, the VIS team discussed several alternatives to visualize GAR patterns with pixel-based or pixel-cell-based techniques such as circular, ring, or matrix representations. As in the linear arrangement of a heatmap, identification of the involved comparisons is not effective for circular or ring representations which violates the interpretability requirement (R0) (see Figure V.4 and Figure V.2). Matrices support the identification of the involved conditions since the matrix element at row x and column y indicates the activity ratio value of the x -th condition and the y -th condition (see Figure V.4 and Figure V.5). Biologists can, therefore, interpret the GAR pattern between conditions within a single gene by inspecting elements of a matrix (R0). Subsequently, they can compare the GAR patterns between multiple genes by inspecting the distribution of patterns across multiple matrices (R1).

Design alternatives for gene fingerprint matrices Each matrix has to represent a summary of a single gene’s activity ratio values and their qualities for different experimental conditions. Since there are several variants to encode the data with the visual metaphor of a matrix, the VIS

team came up with several design alternatives (see Figure V.5) which will be discussed in detail in the following paragraphs.

Two symmetric or triangular matrices for value and quality. One solution is to visualize the quality of each metric as an additional matrix juxtaposed to the corresponding value matrix. Though this design may ensure more accurate perception of both values, there are some significant drawbacks: 1) it wastes valuable display space and 2) it is hard to visually align value-quality pairs. Therefore, this design does not guarantee effective inspection on the *gene activity ratio (GAR)* and the quality (R6) by burdening biologists with cognitive efforts to find and check two locations for a single comparison. The VIS team, therefore, excluded this design.



Value & quality triangles. Similar to the aforementioned design, Figure V.5 (a) shows a design where each of two triangular portions represents the activity ratio and its quality, respectively. This solution was discussed among the VIS team and with the BIO team as well. We concluded that the cognitive efforts to find and check two locations for a single comparison is still a burden for the analysis.

Resizing matrix. A further possibility to encode the quality would be to encode the GAR ratio with color and quality with the size of matrix cells. However, this solution is not scalable and the saliency of patterns is highly dependent on the size and, thereby, on quality which might suppress important patterns in the data. The VIS team, therefore, excluded this design.



Stacked matrix. Another approach is to use a *Stacked Matrix*. This approach is inspired by work of Oelke *et al.* [Oelke *et al.*, 2009], here a stacked resizing matrix is used to represent user opinions on printers. The *Stacked Matrices* in Figure V.5 b) and c) use the outer rectangle for encoding the quality and the inner rectangle for encoding the value. The size of the inner rectangle is fixed. The *Stacked Matrix* with two different color maps perceptually separates the inner and outer rectangles. This design is different from Oelke *et al.* since the inner and the outer rectangle do not represent the same measure in our design and the size is fixed. The proximity between two values enables biologists to read the activity ratio and its quality accurately and, thus, it supports the interpretability (R0) and quality requirement (R6). However, this design may suffer when many fingerprints are shown in a small space. Thus, zooming and panning interactions should be used when the task requires exploration of many genes (T1-T4).

V. VISUAL ANALYSIS OF DIFFERENTIAL GENE EXPRESSION



One might also consider using the same color map for the activity ratios and quality (see Figure V.5(c) upper matrices). Due to the Gestalt Laws of Similarity and Pragnanz, we perceive regions of similar color as a whole large rectangle, instead of several stacked rectangles with different shades of green (see Figure V.5(c)). This supports the detection of row and column patterns (R3) which are important in the tasks of building and associating groups (T1-T4). This design alternative of a *Stacked Matrix* has a higher scalability and can, therefore, be used in overviews with larger amounts of fingerprints.

“In addition to the matrix structure, color maps should be carefully selected because they encode the activity ratios and qualities in our design. The selection of color maps impacts upon the performance of all tasks (T1-T4) because our visual cognition system is steered by several attention effects. Our vision tends to focus on strong contrasts especially when colors are fully saturated and intense on dark backgrounds. Warm colors will suppress cold ones if they are spatially close. Therefore, lightness, saturation, and temperature of colors must be considered [Wang et al., 2008]. We suggest using a perceptually uniform color map that varies from black to green. In this way, values are perceived more prominently in comparison to the qualities which are encoded with a perceptually uniform gray scale.”¹



2D colormap matrix *“Two dimensional color maps can also be used as illustrated in Figure V.5 (d) and (e). Two dimensional color maps are not suited for accurate value perception [Wainer and Francolini, 1980] but these color maps support the quick assessment of quality differences between different genes (R6) in data exploration (R3). Thus, it is recommended to use this where biologists want to quickly estimate values of multiple genes with a reasonable accuracy (R3).”¹*



“Furthermore, one should note that two dimensional color maps fail to function as intended for people with color vision deficiencies. Addressing this issue, we used opponent chromatic channels to encode the dimensions (normal: red-green, dichromatic: blue-yellow). As illustrated in Figure V.5 (d) and (e), the lower left matrix is clearly different from the other matrices. This is extremely useful to compare GAR patterns with the quality in mind (R6) which is only partially supported by other designs. Furthermore, this design is highly scalable in overviews of vast amounts of fingerprints (see Figure V.12).”¹

Triangle vs. symmetric matrices and reordering. The *Stacked-Matrix* and the *2D Colormap Matrix* designs can be used with a full (symmetric) matrix or even a triangle matrix since half of the matrix comparisons is redundant. The advantage of a triangle matrix would be to

¹[13], written by Sebastian Mittelstädt (see also work distribution in [VisExpress - Visual Exploration of Differential Gene Expression Data](#))

save the space of redundant information. However, after a series of discussions among the VIS team and a consultation of the BIO team, we concluded that a symmetric matrix strengthens the visual saliency of patterns. The BIO team perceived the pattern in Figure V.6 A2, for example, less salient than that in A1 even though the two figures show the same pattern. Further, some patterns might appear more interesting than others with the symmetric layout (e.g., the cross in A1 appeared more interesting than in B1 for the biologists on the first sight). However, the BIO team always reflected the meaning of a pattern and had no concern to realize that B1 reflects the same pattern as A1 (one condition is different to all others; condition 1 for B1 and condition 4 for A1). Rows and columns represent specific experimental conditions which need to be maintained as references in order to assess other matrices. Therefore, the idea of the VIS team to use ordering emphasizing interesting patterns was rejected. Inconsistent ordering may confuse biologists to interpret the comparison of results between multiple genes (R0, R1).

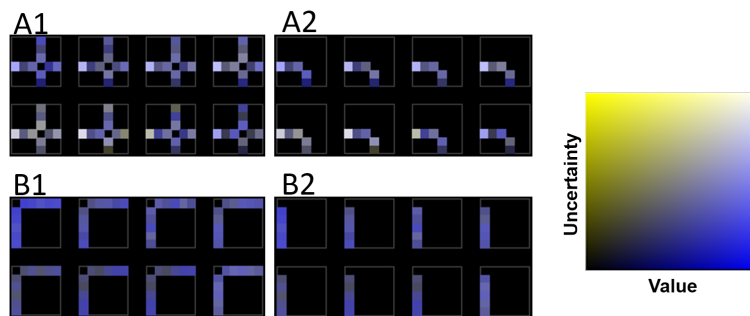


Figure V.6: The Figure illustrates the perceptual differences between (1) a symmetric gene fingerprint matrix and (2) a triangular gene fingerprint matrix. A) shows an example where the fourth condition is different from the rest; B) shows an example where the first condition is different from all the others.



Support of different comparison measures One requirement (R4) is to ‘support different comparison measures’ because multiple measures can increase the level of trust in findings and provide different views on the data set. Reasonable measures are the [fold-change](#) and the significance of the fold-change since they are the state-of-the-art for differential gene expression data. Further useful measures are, for example, the euclidean distance (indicating the difference of activity levels) and dynamic time warping [[Berndt and Clifford, 1994](#)] (indicating the similarity of activity levels) adapted from time series analysis. We use a small-multiples design and, thus, each matrix of a gene fingerprint represents one measure (see Figure V.5 (c)). This allows easy comparison within and between genes and, therefore, also satisfies (R0, R1, and R4).

V-4.4 Components of *VisExpress*

VisExpress provides an overview of gene expression data with a *Cluster View*. The second level visualizes gene clusters with gene fingerprints (*Gene Fingerprint View*), whose design alternatives were discussed in the previous section. The *Gene Board* provides a detailed view of a selected gene (see Section Architecture of *VisExpress* and Figure V.3). In the following, we will introduce and discuss the design of the components of *VisExpress*.

Overview: *Cluster View*

Our overview aims to provide a snapshot of genes grouped with similar **gene activity ratio (GAR) pattern** so that users can immediately grasp the pattern distribution across genes, select interesting group of genes, and delve into details. Therefore, the system must provide a visualization that allows an overview of the clusters (GAR patterns) in the data set, thereby, fulfilling **R0, R1 and R3** (interpretability, comparison and exploration). In order to account for **R2**, the overview should also show a summary of the gene functions of the clusters.

Alternatives for cluster overviews In order to build sets of genes with similar GAR patterns heatmap-based approaches such as [Battke et al., 2010, Saeed et al., 2003, Xia et al., 2013] use clustering. Genes naturally form hierarchical clusters if the genes operate with the same regulatory mechanism (regulon). In heatmap-based visualizations, the hierarchical clustering is used to order rows and a dendrogram is visualized next to the heatmap to represent the clustering (see Figure V.2). However, this representation does not clearly show which different clusters exist in the data set since: 1) clustering is ill-defined and, therefore, clusters are often not visually separable and 2) small clusters might be overlooked. Thus, these approaches do not fulfill the comparison and exploration requirements (**R1,R3**).

There are space-filling visualization techniques such as self-organizing maps (SOM) or treemaps that can be used to overview gene clusters. However, SOM clustering does not preserve the natural hierarchy. Large clusters will span over large parts of the map, whereas small clusters are suppressed. Further, the creation of cluster centroids will refine the centroids of big clusters but suppress centroids of small clusters such that interesting GAR patterns of small clusters are consumed. This violates **R0, R1 and R3** (interpretability, comparison and exploration).

Treemap Overview We choose to visualize groups of genes with a squarified treemap [Bruls et al., 2000] showing the hierarchical clusters. The number of cluster items is encoded by its node size. This enables to assess the importance of clusters but also small clusters are preserved. Inside the treemap either a centroid gene fingerprint of the corresponding cluster

V. VISUAL ANALYSIS OF DIFFERENTIAL GENE EXPRESSION

Explore: *Gene Fingerprint View*

The comparison and exploration of genes according to GAR patterns (R1,R3) requires inspecting sets of genes with similar GAR patterns (R0) and their functions (R2). Sets of genes with similar GAR patterns are given by the clusters in the treemap. The layout of the *Gene Fingerprint View* has to represent large volumes of gene fingerprints. Furthermore, to effectively scan through GAR patterns of a cluster to compare and explore genes (R1, R3), the cognition load needs to be minimized. Therefore, the layout has to use the display space effectively and also provide a structured view on the GAR patterns. Furthermore, quality issues need to be highlighted (R6).

Alternative layouts for gene fingerprint overviews One way to structure the view is a sorting by interestingness function: For instance, by sorting gene fingerprints by their GAR values and/or their qualities, or by the similarity of GAR patterns. The selection of the interestingness function depends on the analysis task and can be changed by the user on-the-fly (see Section V-4.5).

Using an interestingness function allows several alternatives for a structured layout. The most straightforward alternative is, for instance, to layout fingerprints line by line according to the interestingness. However, this does not preserve local proximity (e.g., the two first objects of the first and second row are spatially close but very distant in the interestingness or data similarity). Hilbert curves [Hilbert, 1891] preserve local proximity but cannot guarantee a globally ordered layout since curves might start and also end at the top depending on the number of objects. This violates intuition because intuitively all interesting genes are on the top and the least interesting ones are on the bottom.

Layout of gene fingerprints We used the recursive pattern algorithm of Keim *et al.* [Keim *et al.*, 1995] that is particularly suitable to arrange sorted data points in dense pixel displays. This algorithm lays out the pixels with recursive levels of arrangements (hierarchical “Z”-arrangements) that have specific widths and heights. Thereby, recursive patterns can preserve local proximity and global (intuitive) interpretation. Recursive patterns can guarantee to show the interesting GAR patterns on the top area and similar patterns in proximity.

As shown in Figure V.8, the system arranges the fingerprints on the first level by 4 columns to the right, one row down, 4 columns left, one row down, and 4 columns right to complete the “Z”. This pattern is then repeated 4 times to the right and then 4 times to the left in the lower row. In each level the ordering of the interestingness is preserved which preserves local proximity and (intuitive) interpretation of the whole layout (top: the most interesting ones (green); bottom: the least interesting ones (red)). A disadvantage of the technique is that

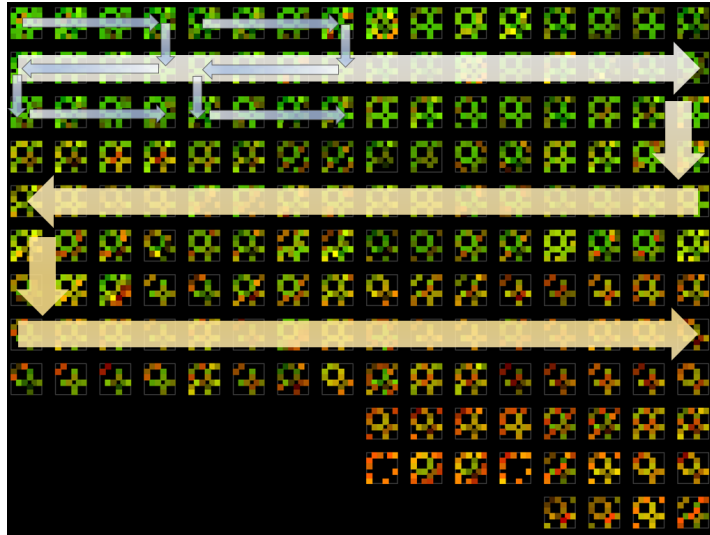


Figure V.8: Overview of gene fingerprints. Matrices are sorted according to the interest of the user and layouted in recursive patterns [Keim et al., 1995].

parameters of the algorithm have to be selected in advance. The problem is to find a good combination of widths and heights (e.g., four steps in the example above) for each recursive level. Keim *et al.* [Keim et al., 1995] suggest determining the arrangements by interaction. However, this would disturb the exploration process and we decided to determine the parameters automatically by applying an optimization algorithm to this combinatorial problem.

Optimization details of the recursive pattern layout “Here we describe our combinatorial optimization process. The optimization goal is to find a combination that 1) layouts all fingerprints; 2) uses as much of display space as possible; and 3) assigns quadratic size to the fingerprint matrices. A combination can be evaluated with multi objective cost functions with: f_1 being the number of elements that cannot be visualized with the combination; f_2 is the number of unused pixels; f_3 is the maximum ratio of the width and height of the fingerprints. The cost functions are computationally cheap which led us to choose ant colony optimization [Dorigo et al., 2006] that tests stochastically selected solutions and converges against the global optimum by the power of randomness. The “ant workers” randomly select the widths and heights for the hierarchy levels. As soon as all fingerprints can be visualized ($f_1 = 0$), the “ant worker” stops and evaluates its solution with f_2 and f_3 . Good solutions will influence other “ants” and the algorithm converges.”¹

¹[13], written by Sebastian Mittelstädt (see also work distribution in [VisExpress - Visual Exploration of Differential Gene Expression Data](#))

V. VISUAL ANALYSIS OF DIFFERENTIAL GENE EXPRESSION

Detail: *Gene Board*

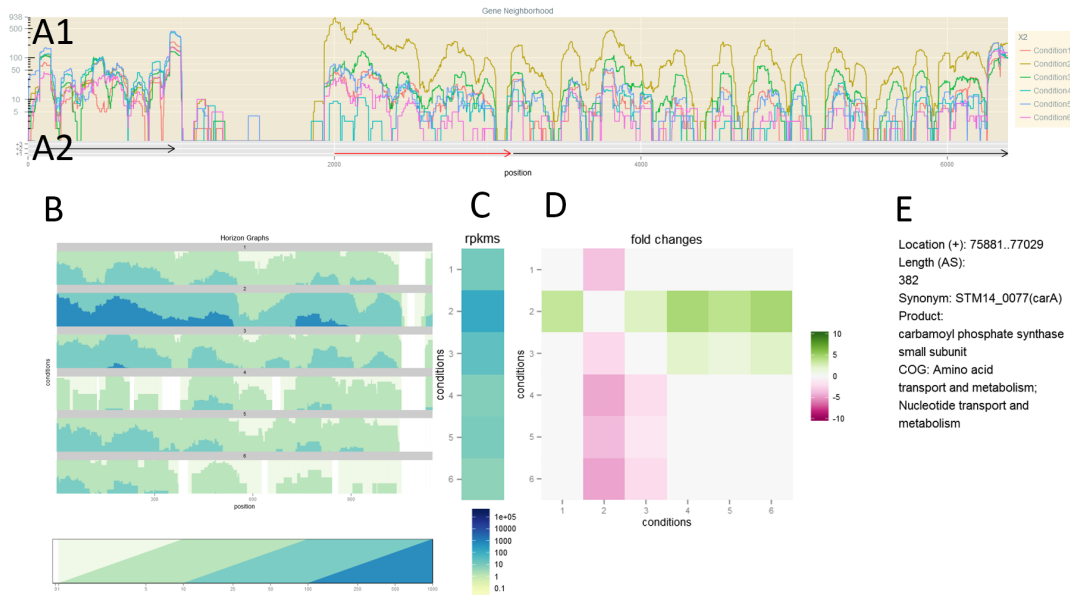


Figure V.9: An example of a *Gene Board* is shown. (A) shows the trend of the gene activity levels for the gene (red arrow in (A2)) and gene neighbors (black arrows in (A2)). (B) shows the trend of the gene activity levels for the gene region with horizon graphs. (C) shows the normalized gene activity levels. (D) shows the GAR pattern and (E) summarizes gene descriptions and gene functions. (B), (C) and (D) are closely arranged to set their data into context. In detail: (C) shows that condition 2 has the highest normalized activity level. Compared to other genes, this value is in a medium range (see color legend). (B) The activity level drops before the end of the gene (probably due to a technical artifact). (D) The horizontal green line indicates that condition 2 is up-regulated in comparison to the other conditions. However, (B) and (C) show that the gene is active in all conditions.

This level supports detailed information about a single gene for the manual assessment of the trustworthiness and a detailed inspection (R5). The design of the *Gene Board* was not the focus of this paper but was highly tailored by the given application specific specifications (sub-requirements of R5, see also Chapter IV) and closely coordinated with the BIO team (see Figure V.9).

The baseline for the design was a gene activity level view with genome annotations (A1), following the style of genome browsers, we use line charts which are a common representation of gene activity level trends in genome browsers. A focus on ratios in the data representation improves the interpretability as the BIO team is mainly interested in the *gene activity ratio (GAR)s* between conditions (achieved by a log scaling). Position of the gene (red) and neighboring genes are indicated with arrows (A2). As the strengths of the activity levels and their

trend over the gene are a major assessment criterion, we decided to additionally show the trend of the activity levels as horizon graphs (B). Horizon graphs are a visualization for sequential data that enables easy comparison between multiple conditions [Heer et al., 2009]. This enables the biologists to see at a glance which conditions have a high activity level and to easily assess the trend over the gene. Next to the horizon graphs, the normalized *gene activity level* (Reads Per Kilobase per Million mapped reads (RPKM) values) are represented as color-coded pixels (C). We use a global color-coding to allow a comparison between genes. In this way, the trend of activity levels (horizon graphs (B)) can be set directly in context with the normalized gene activity levels (pixel-column (C)). The *gene activity ratio* (GAR) patterns are shown as a matrix representation (D) next to the normalized gene activity levels. Thereby, biologists can easily set the GARs in context with the strength of the gene activity levels. Gene descriptions and functions are shown as plain text (E).

V-4.5 Interaction Design of *VisExpress*

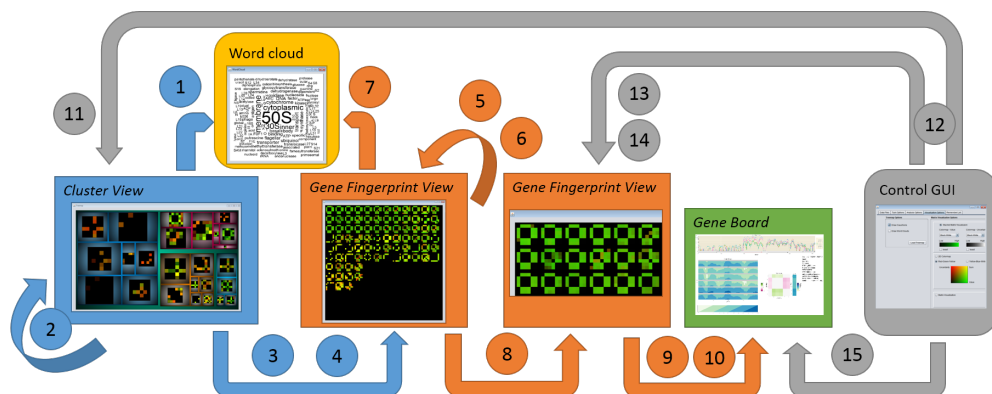
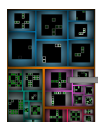


Figure V.10: This figure summarizes the interaction possibilities with the three different views *Cluster View* (blue), *Gene Fingerprint View* (orange), *Gene Board* (green) and the control GUI (grey), as well as the details-on-demand word cloud view (yellow). Interactions are indicated by arrows. Interactions are classified according to Brehmer and Munzner [Brehmer and Munzner, 2013]. See Section Interaction Design of *VisExpress* for explanations of the interactions (numbers are mentioned in the text).

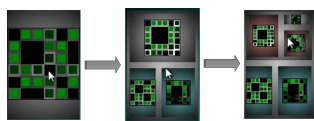
In this section, we explain *how* we have implemented the requirements with interactions, classified according to the multi-level task typology of Brehmer and Munzner [Brehmer and Munzner, 2013]. See Figure V.10 for an overview of interactions. The numbers in brackets, in the following sections, correspond to the interactions in the figure, interactions according to [Brehmer and Munzner, 2013] are set in *italics*.

V. VISUAL ANALYSIS OF DIFFERENTIAL GENE EXPRESSION

Interactions of the *Cluster View*



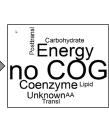
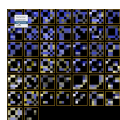
The *Cluster View* provides an overview of the data set by showing the **gene activity ratio (GAR) pattern** of the cluster representative per default (see Figure V.11 A). In order to *summarize* the gene functions (R2) within a cluster and to *compare* these with the GAR pattern representative of one cluster, the user can *navigate* (details-on-demand) by mouse over to the corresponding word cloud (1) (see Figure V.11).



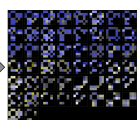
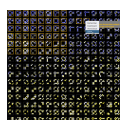
The quality of the cluster representative is encoded by the saturation of the colored surround to indicate if a cluster should be refined. For *identifying* the corresponding subclusters and, thereby, to explore the data set for interesting clusters (R3), *VisExpress* enables the user to drill-down (*navigate*) the cluster hierarchy by right clicking on the cluster representative (2). In order to support the exploration of genes (R3) and to *compare* or *identify* interesting genes users can *navigate* (zoom) to the *Gene Fingerprint View* showing all GAR patterns of genes by left-clicking on the cluster representative (3). Finally we allow the user to call up *Gene Fingerprint Views* of several clusters in order to support a *comparison* between clusters and GAR patterns (R1) by *arranging* the *Gene Fingerprint Views* next to each other (4).

Interactions of the *Gene Fingerprint View*

The *Gene Fingerprint View* visualizes all gene **gene activity ratio (GAR) patterns** of the selected cluster (see Figure V.11 C). See Figure V.10 for an overview, number in brackets are numbers from the figure. In order to *identify* a gene of interest and to relate the GAR pattern of the gene with its function, details-on-demand (*navigate*) showing the gene name and function in a tool-tip (R3) are provided by mouse over (5). Right clicking on the gene will *record* it on a remember list in the control GUI, where the gene fingerprint of the corresponding gene is saved with a thumbnail (6).



Users can also select a set of genes to *summarize* and relate the functions of the selected genes by *navigating* (details-on-demand) to the corresponding word cloud (see Figure V.12) (R2) (7).

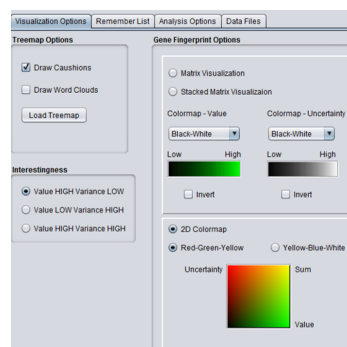


Furthermore, users can *filter* to a set of selected genes by opening a new *Gene Fingerprint View* to *compare* and *identify* interesting genes in the selection (R3) (8). Allowing the assessment of the trustworthiness (R5) users can *navigate* to the *Gene Board* showing details of the read coverage and further *summarized* information about the selected gene (9). Finally we allow the user to call up several *Gene Boards*. By *arranging* the windows next to each other a *comparison* between GAR patterns (R1) and the underlying data is supported (10).

Interactions of the *Gene Board*

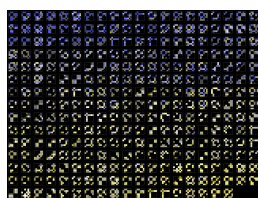
So far no interactions are implemented for the *Gene Board* which can be interpreted as a static Dash Board. However, the user evaluation revealed a set of useful interactions. This includes browsing and zooming in the line chart representation as well as the possibility to call up *Gene Boards* of neighboring genes, by clicking on the arrows indicating the gene locations. As neighboring genes are of special interest users also requested to show the location of clicked neighboring genes in the *Gene Fingerprint View*. Furthermore, the BIO team requested a direct link to the gene database entries at, e.g., NCBI [[Coordinators, 2013](#)].

Control GUI interface



Since the BIO team had no issues with the different designs and understood their advantages and disadvantages, we decided to let the user freely configure the system to the analyst's needs. All these adjustment possibilities give users the flexibility to adaptively test powerful combinations as they encounter different types of tasks. Additionally, visualizations can be further customized, for instance, by hiding specific conditions or enabling or disabling symmetric matrices (see Fig. V.12 D).

Allowing a *comparison* of gene functions between clusters, the *Cluster View* can be *changed* to a treemap showing word clouds (see Figure V.7) (R2) (11). In order to *identify* and *compare* interesting genes (R6, R1) users can *change* the visual design of the *Gene Fingerprint View* to best fit their current analysis task (12). This includes *changing* the color mapping as well as the design of the gene fingerprints (see Figure V.5). Additionally, the gene fingerprints can be *arranged* (ordered) by different interestingness functions to sort the layout of gene fingerprints for different analysis interests (13).



In the left Figure a 2D color map is used, the ordering is 'Value and Quality high'. The recursive pattern algorithm layouts the genes in a way that high value and high quality genes are shown at the top left and genes with low value and low quality are shown at the bottom right. The 2D color map is well suited to separate 'good' (blue) from 'bad' (yellow) genes (Notice: we also provide a 2D colormap for people with color vision deficiencies). In order to get a different perspective on the data, users can also add further measures to the *Gene Fingerprint View* (R4) (14). Users can *import* pre-calculated measures and add them to the *Gene Fingerprint View* (see Figure V.5 and V.8).

V. VISUAL ANALYSIS OF DIFFERENTIAL GENE EXPRESSION



Allowing the user to re-check genes, saved to the remember list, and to assess the trustworthiness (R5), users can *navigate* to the *Gene Board* showing details of the read coverage and further *summarized* information about the selected gene (15). The gene is always saved with the design that was active at the selection which allows the user to relate the gene to the reasons for the selection (see Figure at the left). The remember list allows the externalization of findings which supports the exploration and verification loop of the knowledge generation model of Sacha *et al.* [Sacha *et al.*, 2014].

V-4.6 Evaluation

User assessment

We conducted a qualitative evaluation with three professional molecular biologists. As *VisExpress* is intended to support a visual exploration of differential gene expression data, we decided to conduct an open-ended exploratory study and to evaluate *VisExpress* with a Pair Analytics [Arias-Hernandez *et al.*, 2011] study. Thus, a domain expert (biologist) and a visualization expert collaboratively explore a complex real-world data set and generate conversation about the domain experts' analytic activities.

For the whole study we captured screen activities and verbal reports using Camtasia Studio [Camtasia,] and also filmed the screen to capture when participants pointed on the screen. We performed the study with the three participants B1, B2, and B3 (domain experts; molecular biologists) and myself as the experimenter (visualization expert with a bioinformatical background).

Participants The three participants (B1, B2, and B3) were molecular biologists working with bacteria in the same institute. Richard Landstorfer (B1, 30, end of PhD studies), Klaus Neuhaus (B2, 45, PostDoc) and Anna Mühlig (B3, 28, end of PhD studies), working at the ZIEL institute, Technische Universität München. They have been working in the field of molecular biology for 5, 18 and 3 years, respectively. They analyzed NGS (RNAseq) data collected from their own experiments regularly for their research in the last 2 (B3) respectively 4 (B1 & B2) years, either for gene expression or differential gene expression between conditions. Both B1 and B2 are my cooperation partners in the [FOG-Project](#). In addition, the managing director of the institute (Prof. Siegfried Scherer), who is a professor for microbial ecology for over ten years, gave feedback about the *VisExpress* system (B4).

Data The data set consists of 6 different conditions and over 5000 genes are annotated for the used *Salmonella* Typhimurium strain. The data set was already analyzed by B3 but was unknown by B1 & B2. We have chosen this data set to evaluate how well *VisExpress* is suited for an exploration of an unknown real world data set (B1 & B2) as well as to evaluate if B3 could rediscover findings from her previous analyses.

Condition 1 is a standard condition, in condition 2 supplement A (a nutrition source) is added; in condition 3 supplements A and B (a food additive) are added, condition 4 is the same as condition 3 at a later point in time (stationary state), in condition 5 supplement C (an acid) is added, in condition 6 supplement B and C are added. The conditions 3-6 are different stress conditions.

Study procedure The study was conducted according to following procedure:

Instruction. Each participant entered the user study room separately which was reserved within experts' workplace. The participant sat down next to the experimenter with a notebook and one monitor (24" LCD). The experimenter provided detailed instructions on the system through a slideshow presentation. Details such as visual representations, underlying data, measures, and interaction capabilities were covered so that participants could use the functions later on.

Introductory Tasks of the Paired Analytics Study After the introduction the participants were asked to perform a set of tasks. These tasks required them to conduct a series of analytic activities and system operations. We intended to demonstrate work patterns to explore data with our system showing interaction and adjustment possibilities of our design. The participants had to solve the following set of evaluation tasks:

ET1 Click on a cluster where only the conditions 1, 5 & 6 are different. (R0)

ET2 Which matrix shows a different pattern? What is different? (R1, R3)

ET3 Try out different designs to test if another one is better suited for finding matrices with high quality. (R6)

ET4 Look at the word cloud of the open cluster. (R2, T1)

ET5 Save some interesting genes to your remember list and call up the *Gene Board* for one.

ET6 Could the signal come from a neighboring gene? (R5)

ET7 Has the gene a good activity level in the conditions that show differential gene expression. (R5) Explain your procedure.

In this step of the study, the experimenter operated the system and participants were allowed to ask questions to clarify any uncertain areas.

V. VISUAL ANALYSIS OF DIFFERENTIAL GENE EXPRESSION

Table V.1 shows the time spent for each part of the study per participant.

	B1	B2	B3	B4
Instruction	0:22	0:22	0:22	0:00
Introduction to the system	0:21	0:28	0:14	0:00
Exploration	2:01	1:30	1:04	0:00
System demonstration	0:00	0:00	0:00	0:22
Informal Feedback	0:00	0:00	0:00	0:25
Sum	2:44	2:20	1:40	0:57

Table V.1: Total time spent of each participant for the different parts of the study. B4 is a senior researcher who gave informal feedback after an introduction and demonstration.

Open-ended exploratory part. After participants had completed all given tasks, we asked them to freely explore the data set which was the main part of the study. The participants were asked to verbally formulate, confirm, or reject hypotheses during the analysis process and to report interesting or unexpected findings along the way. As the experimenter, I encouraged the domain experts also to focus on patterns which appeared interesting to me as a bioinformatician to facilitate a more collaboratively exploration of the given data and to generate deeper conversation about the biologists' analytic activities, their reasons, and intentions. However, I made sure not to unduly influence the analysis by only suggesting a deeper look in a few cases and, otherwise, only acting as an active listener who did not initiate conversation unless I wanted to clarify uncertain motivation or action (e.g., 'why?' or 'how?'). As participants had no issues using *VisExpress* and since user interaction was quite high, I decided to let the domain experts operate the system themselves.

Coding procedure. We adopted a top-down and a bottom-up approach. Our goals were 1) to reveal the domain expert's workflows with the *VisExpress* system, 2) to clarify expert tasks, and 3) to specify areas for improvements. As the experimenter of the study, I first formulated findings from study impressions and verified them with corresponding clips of the video material. A second author checked against these findings with the corresponding clips. Second, I coded the whole video material. The video material was first annotated and split into clips according to the different used views (*Cluster View*, *Gene Fingerprint View*, *Gene Board*). For each clip, I coded the participants' analytic and visualization activities. In particular, the attempt was to reveal the reason behind new participant's actions and workflows that lead to findings. From this analysis, I formulated further findings. The findings were verified with the clips by a second author.

Results

Three domain experts (B1-B3) participated in this study. In addition, the managing director of the institute (professor for microbial ecology) gave feedback about the *VisExpress* system (B4). In total, 7 hours and 41 minutes were recorded (see Table V.1). We formulated the following findings from the study and verified them with video clips.

Study findings The used data set was new for B1 and B2. They remarked that they just got an overview during the study and would need more time to deeply analyze the whole data. Nevertheless, B1 and B2 and also B3 were impressed how fast they got an overview. B3 rediscovered several findings, regarding groups of genes and single genes as well. We concluded the following points which also distinguish *VisExpress* from state-of-the-art systems (all participants agreed on the quotes stated here):

- The system is in-line with the mental model of the biologists and easy to learn. Actually, we observed no learning curve at all for all participants. All participants answered the introductory tasks correctly and without much reflection. B2: *'The system is straightforward.'*; B4: *'I have not heard of these word clouds before but they are immediately comprehensible.'* (fts - free translation(s))
- *VisExpress* helps biologists to get a fast overview of the data. B2: *'I was astonished how fast I got an overview of this [bacteria] project'.* ; B1: *'It is a very nice tool since I got an overview of B3s data set very quickly.'* [The dataset was not known to B1 & B2] (fts)
- Biologists integrated data quality in their workflow. B1: *'I liked that I could skip many genes since their quality was low.'* (ft)
- *VisExpress* facilitates to generate hypotheses and to bring things into question. B2: *'Based on the patterns, it is easy to generate hypotheses and it is quite fast.'*; *'One can click on a certain [cluster] pattern and look which [genes] belong to that cluster and in no time one can generate a hypothesis.'* (fts) See also next Section *Biological findings*.

General workflow of participants We observed the same general workflow among the three participants. They started from the *Cluster View* and selected a cluster to analyze further. The cluster representatives were inspected as well as the corresponding word clouds to decide on a cluster. In the *Gene Fingerprint View*, participants selected genes to analyze in detail with the *Gene Board*. Genes were selected according to their gene activity ratio (GAR) patterns, their quality and their functional category provided by tooltips. With the *Gene Board*, participants

V. VISUAL ANALYSIS OF DIFFERENTIAL GENE EXPRESSION

assessed the trustworthiness of the GAR pattern. E.g., if the pattern is surprising for the function, a closer look can reveal that the strength of the gene activity levels is too low to trust the GAR pattern. After the inspection of all interesting genes in the *Gene Fingerprint View*, participants switched to the *Cluster View* and looked for the next cluster for further exploration. The outcome of an analysis session is a list of genes of interest which can be checked with literature research and database comparisons. B2 states about next steps: *'I would look up the genes at NCBI, Uniprot, perform a similarity search with BLAST and do a literature research.'*(ft). Consolidated hypotheses could then be verified by further experiments.

All participants used the quality to reduce the search space. They did not pay much attention to genes where all GARs had low quality after they were convinced that the quality is really an indicator for trustworthiness (checked with the *Gene Board*). However, they still inspected low quality genes later on if the pattern was of interest.

Individual analysis processes and findings The following paragraphs quote and describe different examples of the analysis processes and findings of each participant in details.

B1 said: *'I will successively look at all clusters.'* The word clouds were used to get an idea about the included functional categories in a cluster. E.g., B1 said: *'In this cluster should be [supplement A] depended genes.'* and for the corresponding word cloud: *'Energy production and conversion stands out. This is reasonable. [Supplement A] is an energy supplier.'* (fts). B1 also systematically checked gene functions by hovering over at least the first lines of gene fingerprints in each cluster (high quality ones). B1 explained: *'I am looking for the gene functions. It is striking that most genes have a functional annotation, this was not the case for some other clusters.'* (ft). Genes with interesting functions were inspected with the *Gene Board*. B1 tried to gather findings for each cluster and explained whether he had expected them. E.g., B1 said: *'Many genes are related to the cell membrane. I interpret this as extrinsic stress. I am surprised that condition 1 and condition 5 & 6 are similar.'* (ft, remark conditions 5 & 6 are stress conditions but condition 1 is not a stress condition).

B2 built a hypothesis about the data set at the beginning and looked for the respective patterns. However, B2 had to reject some hypotheses in the end. A hypothesis about, e.g., only small differences between condition 1 & 2 was rejected: *'It is a surprising finding, that [supplement A] has an effect on quite a number of genes. [...] I have not expected this.'* The word clouds were less frequently used by B2. After he had checked a few hypotheses, he checked random clusters with interesting patterns and or interesting word clouds. B2 also compared similar clusters, by arranging the *Gene Fingerprint Views* of two clusters next to

each other. In the *Gene Fingerprint View* B2 randomly hovered over genes to get the functional categories, he tended to focus more on varying patterns. E.g., B2 said: *'These are the acid genes. However, this gene stands out. This is obviously a gene reacting on acid and [supplement B] stress only'*. Genes with interesting patterns or functions were inspected with the *Gene Board*. B2 gathered findings for some inspected clusters and explained if they confirm or reject his hypothesis. E.g., B2 said: *'I have no explanation for this pattern. Standard condition and a condition in stationary state [1 & 4] behave similar. I have no idea what these genes should have in common.'* and about the corresponding word cloud: *'Ah...mostly no functional prediction. Thus, also others could not classify these genes.'*(ft; see Figure V.11).

B3 had analyzed the data set before. On the one hand, she tried to rediscover her findings and on the other hand, she inspected clusters with an interesting pattern or an interesting word cloud. E.g., B3 said for one cluster: *'Here we have no difference between conditions 4 & 5 but between most others. I also realized that in my former analysis.'* and for one gene in this cluster: *'I found exactly this gene in my own analysis. A database and literature analysis revealed that this function has not yet been experimentally verified for this organism. The annotation is only based on a low sequence similarity.'* B3 also looked more systematically at the genes in the *Gene Fingerprint View* and hovered over at least the first part of the genes in each cluster (high quality ones) to check the functional categories.

Biological findings - Use case In the following, we provide examples for some biological findings our BIO team made while using *VisExpress* in the Pair Analytics study with a real world data set:



B1 discovered that membrane proteins are disseminated between different clusters. Participant B1 observed many membrane proteins in cluster 'condition 4 high' and cluster 'condition 1, 5 & 6 high' (see graphic). Such patterns (relations of different conditions) are strikingly visible with our gene fingerprints which are easily overlooked in state-of-the-art representations since just (1:n) comparisons are shown. After looking for the gene product names, he concluded that in the cluster 'condition 4 high' more transporter genes are present which are located in the membrane to transport nutrients into the cell. This is reasonable since condition 4 is a stationary state condition and, thus, nutrients are reduced in the medium run and it would be important for the bacteria to increase membrane transporters to get a better yield. In the cluster 'condition 1, 5 & 6 high', B1 observed more membrane proteins related to stress. This is an unexpected finding since condition 1 is the control/reference

V. VISUAL ANALYSIS OF DIFFERENTIAL GENE EXPRESSION

condition. B1 had no explanation why these membrane proteins should react as in conditions 5 and 6 but mentioned that it would be interesting to analyze this surprising fact in detail.

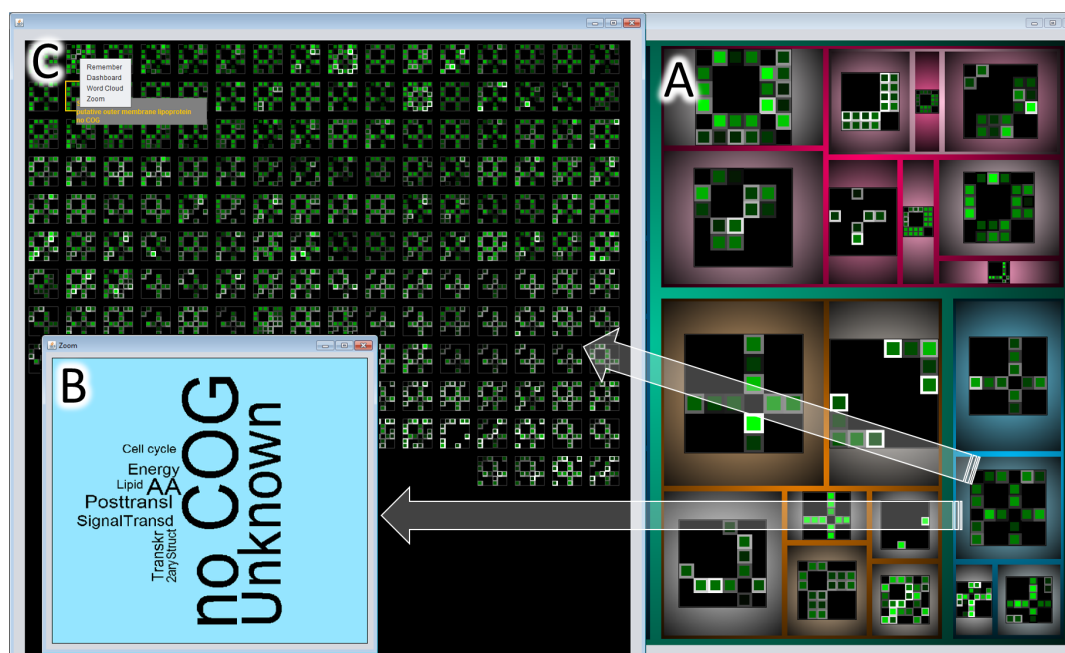


Figure V.11: Annotated screenshot of *VisExpress* on Level 1 (*Cluster View*) and Level 2 (*Gene Fingerprint View*). A) Treemap, showing all gene clusters with centroids represented by their fingerprints. B) Hovering over a cluster shows a word cloud with functional categories of the genes in the cluster. In this example, no functional annotation is given for most genes (no COG [Galperin et al., 2015] & unknown). C) A left click on the cluster in the treemap calls up the *Gene Fingerprint View*. In this cluster, condition 1 and 4 are prominent. Hovering over a gene fingerprints shows the gene product and the functional category in a tool tip (top-left). Multiple gene fingerprints can be selected (orange boarded). For selected genes the detailed *Gene Board* can be called up, users can also zoom to selected genes, create a word cloud for a selection, or add them to a remember list. See also Fig. V.12 for another screenshot and Fig. V.10 for interaction possibilities.

B2 quickly discovered low pH responding genes. Several genes were found to be similarly regulated in low pH (acidic) conditions but showed no or negligible differences between other conditions. By concentrating on this pattern, B2 discovered several genes annotated as ‘hypothetical’ which he would like to examine for their low pH (acidic) response. To B2 the regulation of the acid responsive genes appeared to be more significant than expected in today’s literature. The advantage of *VisExpress* was here that the world clouds allowed an easy and intuitive relation of the cluster to the gene functions. The word cloud allowed identifying that some genes in the cluster are annotated as ‘hypothetical’.

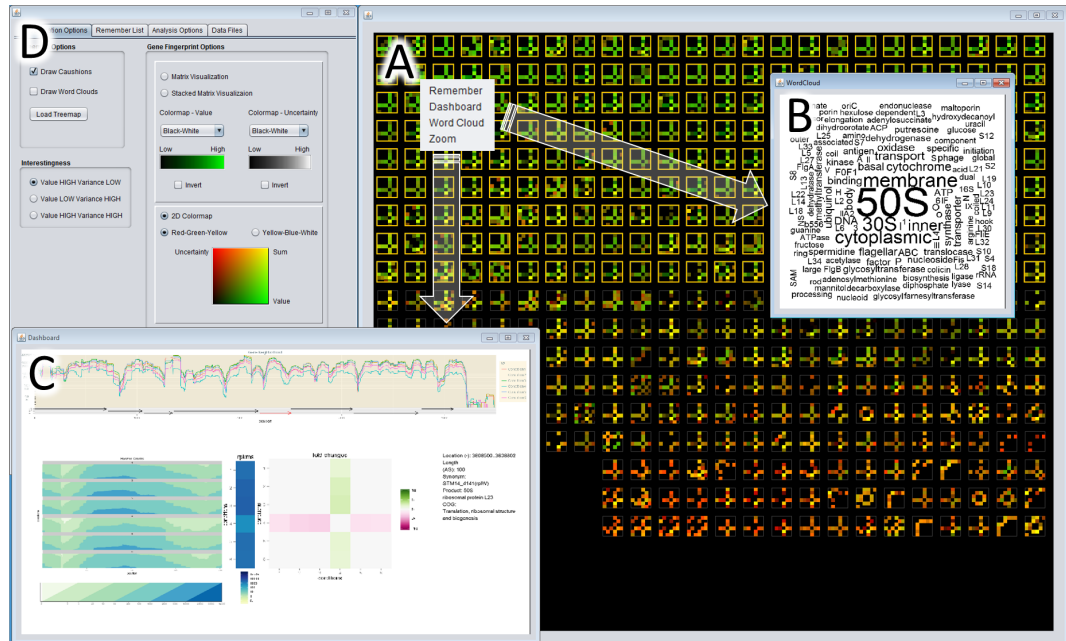



Figure V.12: Annotated screenshot of *VisExpress* on Level 2 (*Gene Fingerprint View*) and level 3 (*Gene Board*). A) *Gene Fingerprint View*, ordered according to high GAR value and high quality is shown in the overview with the 2D color map (green: high value, high quality; red: low quality, low value). Green genes are selected and a word cloud is called up for the selection. B) In the word cloud, 50S is the most prominent word. 50S and 30S are prefixes of ribosomal RNA which builds an important function of this cluster. C) displays the detailed *Gene Board* for one of the genes. It shows that this gene is down-regulated in condition 4. D) represents the control GUI. It is used to switch between the design of the gene fingerprints and the color maps, as well as interestingness functions. See Fig. V.10 for interaction possibilities.

B3 rediscovered that there is a relation between experimental conditions 5 & 6 and iron. Condition 6 is a stress condition which affects iron-sulfur cluster containing proteins. A relation to iron related genes is reasonable. B2 also discovered a specific gene that is responsible for iron-sulfur centers.

 **Ribosomal genes are enriched in a cluster with down-regulated GAR values in condition 4.** This enrichment was observed by all participants. This finding is not surprising because condition 4 is a stationary state (see Figure V.12). Bacteria move into stationary state if their habitat does not allow a further increase of the population size, due to space and low nutrient availability. In this state, bacteria slow their metabolisms to conserve energy. Consequently less ribosomes are needed which produce proteins (encoded by genes). This cluster of down regulated ribosomal genes could be excluded from now on, reducing the data set to more interesting and biologically relevant functions (other than growth speed).



Participants found several patterns they could not explain. B1 observed that several genes with the same function occurred in a cluster where condition 1, 2 and 5 stand out. Detailed analyses with the *Gene Board* revealed that condition 5 is up-regulated, conditions 1 and 6 are slightly up-regulated, and condition 2 is down-regulated. The gene is related with a substance which is added in condition 3, 4 and 6. The reaction pattern was, therefore, not explainable and surprising for B1. Such complex patterns were intuitively perceived by our experts due to the gene fingerprint design. Furthermore, *VisExpress* enables to inspect the functions of genes by demanding word clouds or the detailed *Gene Board*. The experts can query for a comprehensive view of such unexpected patterns more efficiently than in state-of-the-art tools which would require the analyst to perform additional workflows. Such findings are especially interesting in an open-ended/ hypotheses free data exploration because they are starting points for new hypotheses and further research.

V-5 Discussion & Lessons Learned

The problem driven nature of design studies with real domain users generates synergy effects as stated by Brooks [Brooks, 1996]: *'Hitching our research to someone else's driving problems, and solving those problems on the owners' terms, leads us to richer computer science research.'* In this section, we will share our lessons learned and discuss the limitations and future challenges that we identified during our design process of *VisExpress*.

V-5.1 Interrelations between BIO & VIS Experts

Synergy effects. The specific needs and requirements of our domain experts revealed an open gap and research challenges leading to this design study and our contributions to the information visualization domain. Further, I gained a deep understanding on the NGS data preparation process and was able to estimate and formulate sources of errors from the computer science point of view. This led to a new project proposal and is now funded¹. Furthermore, the VIS team questioned the common practice to calculate gene activity and gene expression values instead of analyzing the read coverage data directly (see Figure V.1(b)). Here, for example, methods for comparing time series could be applied which is again an interesting topic for VIS experts.

Do not underestimate biologists. Visualization experts often suggest fancy visualizations in the first place and have to realize in the end that a combination of state-of-the-art techniques is sufficient and gain a better acceptance by domain experts. However, a first refusal of

¹Funded in the third funding period of the priority programme "Information and Communication Theory in Molecular Biology" (InKoMBio SPP 1395) of the German Research Foundation (DFG).

sophisticated visualizations does not mean per se that everything should be simple. Domain experts can often surprise how well they also understand complex concepts. In a series of discussions, I realized, for example, that the BIO team had no issue with understanding the cluster hierarchy in treemaps; they even wanted to interactively drill-down in the cluster hierarchy and explicitly suggested splitting clusters on demand to *identify* interesting GAR patterns deeper in the cluster hierarchy (R3) (see section V-4.4). This was surprising since even VIS students have often problems to understand the hierarchy in treemaps in the beginning. Further, the experts demanded to sketch a GAR pattern to search for similar patterns. I had suggested a *search by sketch* functionality in another context. B2 remembered this and remarked he would like to look for patterns that match his (sketched) hypothesis. B1 & B3 agreed that this would be a helpful functionality.

V-5.2 *VisExpress* to identify new overlapping genes

VisExpress is designed as a proof-of-concept for genes only. Thus, requirement V-2 ('Highlight new gene candidates') is only partially fulfilled by *VisExpress* (see section V-2) in this current version. However, the necessary extensions to completely satisfy V-2 and to incorporate the detection of **overlapping genes (OLGs)** with *VisExpress*, requires only straightforward engineering efforts discussed below.

First, considering new gene candidates increases the number of gene fingerprints up to the 10 fold which requires a highly scalable solution. Our approach allows handling large volumes of data by the clustering and the interactive treemap visualization that create a scalable symmetrization of the data. However, the treemap visualization in combination with gene fingerprints has a limitation in the number of clusters and also in the number of hierarchy levels that are visualized. Therefore, a further interaction must be implemented that allows the user to call up a second treemap view for a selected cluster. Thereby, the full display can be used to visualize all sub-clusters and further levels of cluster hierarchies. In the case that the number of gene fingerprints per *Gene Fingerprint View* exceed the display space, the same approach can be applied. A second treemap is called up for the selected cluster before visualizing the *Gene Fingerprint View*, representing the lower cluster hierarchy and, thereby, splitting up the number of genes for per *Gene Fingerprint View*. Furthermore, genes with highly similar patterns and the same functional COG category can be aggregated in the *Gene Fingerprint View*. This is especially of interest for genes with no functional COG category [Galperin et al., 2015], as they do not contribute direct information. The number of aggregated genes could be indicated by size, colored bounding boxes (in combination with tool-tips). Switching views is also a possibility:

V. VISUAL ANALYSIS OF DIFFERENTIAL GENE EXPRESSION

E.g., showing a *Gene Fingerprint View* without indication of aggregations and switching with a smooth transition to a size-encoded view through user actions.

Second, new gene candidates need to be highlighted. Similar to the orange bounding box, indicating selected gene fingerprints, gene candidates would be highlighted with a bounding box. An alternative would be to split the *Gene Fingerprint View*, showing genes and gene candidates in the left and right part of the *Gene Fingerprint View*.

Third, meta data of gene candidates need to be represented. The main meta information of gene candidates are BLAST hits. One way to represent this information would be to create word clouds of the functional description of the BLAST top hits (and/or COG category). The word clouds can be compared with the word cloud of the known genes in the cluster. Gene candidates with highly similar patterns and COG categories from the BLAST hits can be aggregated as described for genes.

V-6 Limitations & Future Work

Dimensionality. We have applied *VisExpress* on a data set with a maximum of six experimental conditions resulting in six rows and columns in the matrices. This is a reasonable number but also data sets of experiments with a higher number of conditions exist. We argue that the gene fingerprints scale with the number of dimensions since the most interesting patterns, such as crosses and blocks, remain visible. However, due to the high number of rows and columns, it will be harder for the analyst to determine which conditions form these patterns. Therefore, we plan to integrate details on demand revealing the involved conditions in the analyst's focus. Nonetheless, we expect that a number of more than 12 conditions is not feasible. On the one side due to visual scalability issues. Large gene fingerprints poses not only issues for the identification but also limit the number of gene fingerprints which can be overviewed on the screen. On the other hand users can not easily manage a high number of different conditions. 12 conditions would lead to 66 pairwise condition comparisons which is already hardly assessable.

Support for bottom-up analysis. Our design specifies a top-down analysis for exploration. Analysts start with a cluster hierarchy and narrow down the subject of analysis. Participants stated that they would also like to have the opportunity to start an analysis with a set of interesting genes, e.g., genes known to respond on acid. The system should import a list of genes with similar reactions (provided by the analysts) and expand this set of genes with new similar candidates. A similar approach is presented by Bertini *et al.* [Bertini *et al.*, 2011] to explore large chemical libraries and v.d. Elzen and v. Wijk [van den Elzen and van Wijk, 2014] to explore multivariate networks.

Chapter VI

Concluding Remarks and Perspectives

Note

This chapter is partly based the following publications and parts of this chapter appeared or will appear in the following publications [14, 13, 12]¹

[14]: Svenja Simon, Daniela Oelke, Richard Landstorfer, Klaus Neuhaus, and Daniel A. Keim. “*Visual Analysis of Next-Generation Sequencing Data to Detect Overlapping Genes in Bacterial Genomes.*” 2011 IEEE Symposium on Biological Data Visualization, October 23 - 24, Providence, Rhode Island, USA, 47-54, IEEE, DOI: [10.1109/BioVis.2011.6094047](https://doi.org/10.1109/BioVis.2011.6094047), 2011.¹

[13]: Svenja Simon, Sebastian Mittelstädt, BC Kwon, Andread Stoffel, Richard Landstorfer, Klaus Neuhaus, Anna Mühlig, Siegfried Scherer, and Daniel A. Keim. “*VisExpress - Visual Exploration of Differential Gene Expression Data.*” Information Visualization, 1-26, DOI: [10.1177/1473871615612883](https://doi.org/10.1177/1473871615612883), Published online before print December 14, 2015.²

[12]: Svenja Simon, Sebastian Mittelstädt, Daniel A. Keim, and Michael Sedlmair. “*Bridging the Gap of Domain and Visualization Experts with a Liaison.*” Eurographics Conference on Visualization (EuroVis) - Short Papers, Cagliari, Italy, 25 - 29 May 2015, 127-133, The Eurographics Association, DOI: [10.2312/eurovisshort.20151137](https://doi.org/10.2312/eurovisshort.20151137), 2015.³

¹For the division of responsibilities and work, as well as a statement of contributions in these publications, see [Visual Analysis of Next-Generation Sequencing Data to Detect Overlapping Genes in Bacterial Genomes](#) (p. 10), [VisExpress - Visual Exploration of Differential Gene Expression Data](#) (p. 11) and [Bridging the Gap of Domain and Visualization Experts with a Liaison](#) (p. 9).

¹The Institute of Electrical and Electronics Engineers (IEEE) is the copyright owner of this work [14] but, as an author, I am permitted to re-use the work of this publication (verbatim and derivative) for my personal use. Link to the published article in IEEE Xplore: <http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6094047>

²I own (with the co-authors) the copyright of this publication. The SAGE Publications Ltd holds the sole and exclusive right and license for publishing ([13]). The definitive version is available at <http://ivi.sagepub.com/> Direct link to the published article: <http://dx.doi.org/10.1177/1473871615612883>

³I own (with the co-authors) the copyright of this publication. EUROGRAPHICS holds the exclusive license for publishing ([12]). The definitive version is available at <http://diglib.eg.org/> Direct link to the published article: <http://diglib.eg.org/handle/10.2312/eurovisshort.20151137.127-131>

VI. CONCLUDING REMARKS AND PERSPECTIVES

This chapter concludes the thesis by setting this work in the broader context of interdisciplinary problem-driven visualization research. I will summarize the development process of this thesis as a comprehensive piece of interdisciplinary problem-driven visualization research and seize the idea to foster interdisciplinary communication with a *Liaison*, an idea that came up in the course of this thesis. Furthermore, I will summarize “*where we are today*” with respect to visual analysis of RNAseq data to discover and describe genes, and state open issues, as well as future lines of research¹. In this context, I will also name the biological publications and manuscripts the work of my thesis has contributed to. Finally I will close with some further challenges in biological data visualization.

VI-1 Experiences and Lessons Learned as an Interdisciplinary Problem-Driven Visualization Researcher

I started this thesis without major background knowledge in visualization and [next-generation-sequencing \(NGS\)](#). However, with a major in bioinformatics, I had grounded background knowledge in genetics and molecular biology. Therefore, when starting with my PhD, I had to learn the basics of visualization and the context of the [FOG-Project](#). Basically I had to learn:

“What is required, what is needed, what is feasible, and what is a Vis contribution?”

Find a good level of abstraction.

Chapter [III \(Requirement Analysis and Problem Abstraction\)](#) summarizes the problem understanding and abstraction I reached in this topic. However, in a complex application domain like molecular biology, the first hurdle is to identify a self-contained but still meaningful and essential problem, for which a visual solution is reasonable and a contribution in itself (see also [Awareness of the problem complexity contradicts with a practical solution](#) (p. 24)).

After the development of the *NGS Overlap Searcher* (see Chapter [IV](#)), the next step was not clear from the beginning, even though the direction was given by “[Comparison of gene activity levels between different experiment conditions](#)” (see Section [III-4](#)). A first prototype turned away from the common genomic coordinates of genome browsers. I discovered that the biologists needed the surround of genes and contextual information ([R-III](#) and [R-IV](#)) but not necessarily fixed on the genome or in the genomic order. Therefore, a solution was implemented, in which each open reading frame (ORF) was treated (with its surround) as one entity and its

¹Note that the future work which specially relates to the developed systems was already discussed in the respective sections ([IV-6](#) and [V-5](#)).

read coverage was visualized with a line chart. Further, all ORFs were visualized as small multiples in a grid (see Figure III.16 I (p.50)). Even though the design helped to concentrate on ORFs of interest, I encountered two issues: First, a visual comparison of **read coverage** of different experiments was not well supported by line charts alone. Secondly, the design did not help to correlate genes with a similar behavior.

Based on the experiences with the described prototype, the most important tasks became apparent (see Section III-4.3) which were first camouflaged by the pressing issue to deal with the massive data amount. One further important point was to identify **differential gene expression** analysis as a method to address these tasks.

This is a further important point, beside the challenge to identify the real and important tasks, abstractions are needed to tackle complex problems. This includes to select methods like gene expression analysis to generate derived data which is appropriate to support these tasks. Furthermore, data as well as tasks need to be further abstracted to a visualization language in order to derive requirements from them (see Figures III.18, III.19 and III.20 and the derived requirements in Section III-4.3).

However, it is often hard to find a good level of abstraction. In the *VisExpress* design study, for instance, **gene activity ratio (GAR)** values are visualized with a green color scale. However, one biologist mentioned in the evaluation study that a binary representation (green: any value, black: no value) would also be sufficient for some tasks. This would ease some analysis processes. However, after an explanation of the bias of automatic thresholding, the domain expert saw the danger of this approach and withdrew the request.

It is, therefore, important to work closely with domain experts to determine a good level of abstraction. A visualization researcher should scrutinize for all parts whether a further abstraction is reasonable. However, a visualization researcher should also scrutinize abstraction requests for their meaningfulness since domain experts sometimes tend to abstract too much. Furthermore, one should keep in mind that a reasonable abstraction level might also depend on tasks in mind.

A visualization expert with application domain knowledge helps to bridge the gap.

During my PhD, I encountered the issue that the visualization colleagues I collaborated with, had issues to understand the domain problems and the issues stated above. First, because the language between biology and VIS domain differs strongly. Secondly, because certain basic knowledge is required to understand biological problems. Therefore, misunderstandings occurred frequently. However, by applying my domain knowledge, I could reduce learning time by directly abstracting and translating application problems to visualization terms. Based on

VI. CONCLUDING REMARKS AND PERSPECTIVES

the insight that this approach was not only beneficial for the team-work with my Vis colleagues but also beneficial in general for problem-driven visualization research endeavors, I introduced the *Liaison* role [12]. This role is described in Chapter (II) and was utilized in the *VisExpress* project [13] where the advantages of the *Liaison* role were demonstrated (see section II-5.1 and II-5.2, as well as Chapter V).

Open challenges. In [12] the interdisciplinary communication issue was defined and introduced to the visualization community. Discussions and an exchange of ideas about currently applied and alternative approaches for this issue are promoted, as well as follow up research in the visualization community. In design study papers, a description of how visualization researchers bridged the knowledge gap and how they reached a problem characterization is often not available. However, such descriptions are valuable to foster discussions on how to address the *Bridging the Gap*-challenge (see also [Bridging the gap*](#) (p. 4)). I argue that the *Liaison* overcomes the issue sufficiently if the role is appropriately integrated and deployed. One important point here is to have an independent VIS team of several visualization experts to span a broad *Design Space* and to avoid that the VIS team is biased by detailed domain issues that may hamper the development of ideas. However, I also see a high potential in brainstorming sessions and/or other creativity techniques with the domain and visualization experts to reveal visualization experts insights in the domain problem which might lead to out-of-the-box ideas. An interesting research challenge is, thereby, to integrate all team and project members in the creativity discussion to take the advantage of the expertise in both fields mediated by the *Liaison* but, at the same time, preserving the independence and creativity of the VIS team.

VI-2 Were Are We Today?

Visual Analysis of RNAseq Data to Discover and Describe Genes

In this thesis, I contribute a problem characterization of the visual analysis of RNAseq data to discover and describe genes (see Chapter III). Thereby, I identify two research gaps that state-of-the-art systems do not support. First, to assess the trustworthiness of [RNAseq measurements](#) and, second, to discover and relate genes to identify their functions. For both research gaps I contribute a visual analysis solutions.

State-of-the-art systems to visualize RNAseq data are genome browsers. However, genome browsers do not follow visualization design guidelines. The stacked reads representation, for instance, introduces visual artifacts and the change of the scaling to the current view hampers a comparison of measurements. Furthermore, the high spatial distance between data

representations and gene annotations hampers the assessment of measurements. Finally, genome browsers do not support filter capabilities to reduce the search space. Systems related to the second research gap, are systems for gene expression data analysis. However, these systems do not incorporate quality issues and do not well support pair-wise (n:n) comparison of experiments.

The *NGS overlap searcher* (IV) supports the assessment of the trustworthiness of RNAseq measurements to address the issues of genome browsers. In order to tackle the large volume of data, the *NGS overlap searcher* provides filter capabilities to define [open reading frames \(ORFs\)](#) of interest. For visualizing the effects of parameter changes and to adjust them the *NGS overlap searcher* comprises a *Genome Overview Bar* which summarizes the status of genes with respect to the parameter settings. In order to allow an effective and efficient assessment of RNAseq measures, the *NGS overlap searcher* resembles the design of genome browsers but overcomes their weaknesses. The RNAseq measures are represented as bar charts between the genome strands to reduce the mental load to map measurements to ORF representations. Furthermore, RNAseq measurements are mapped to the ORF representation itself with two-tone pseudo coloring to visually fit the measurements to the exact ORF position. The *NGS overlap searcher* was already applied in several studies to inspect the read coverage of genes (see [Applications](#) (p. 118)).

The *VisExpress*-system (V) addresses the second research gap and supports the exploration of RNAseq data to discover and relate genes to identify their functions. In order to structure the large data volume, genes are clustered in *VisExpress* based on similarity of [gene activity ratio \(GAR\) patterns](#). Cluster representatives are represented in a treemap to provide an overview. Gene function word clouds can be called up on demand to relate GAR patterns with gene functions. For a further exploration of the data set, users can switch to the *Gene Fingerprint View* and analyze genes in detail with the *Gene Board*. Gene fingerprints are used to represent the GAR patterns of genes, as well as their quality. The applicability of *VisExpress* was demonstrated in a pair analytics study (see Section [V-4.6](#)).

Open Issues and Challenges

Analysis of RNAseq bias sources. In Section [III-2.2](#) the bias sources in the sequencing protocol are discussed. The awareness of the uncertainty issues led, in the course of [FOG-Project](#), to the idea to adapt an experimental protocol to label each sequence fragment with an individual so-called index before PCR amplification. Sequencing data from these experiments is now available and opens up the possibility to study the reasons and strengths of biases in more detail, as PCR bias can be subtracted. This analysis might also lead to the development of automatic methods to compensate for biases.

VI. CONCLUDING REMARKS AND PERSPECTIVES

Analysis of gene function enrichments. The *VisExpress* design study focuses on the visual exploration of differential gene expression patterns. The relation of gene activity ratio (GAR) patterns to the functions of genes is revealed in *VisExpress* by word clouds. For further enhancements the analysis of gene functions, functional and gene set enrichment analysis could be integrated (see [Hung et al., 2012] for an overview). Beside, statistical analysis of “unexpectedness” of gene functions, this also requires a tightly integrated expert for justification with visual analysis systems since “expectedness” depends also on implicit domain expert knowledge and is, therefore, ill-defined. A similar problem and solution was presented by Mittelstädt *et al.* [Mittelstädt et al., 2014] that requires a tightly integrated physician for adverse drug event detection.

Applications

The *NGS Overlap Searcher* [14] was applied in several studies to inspect read coverage of genes to exclude false positives:

Richard Landstorfer, Svenja Simon, Steffen Schober, Daniel Keim, Siegfried Scherer, and Klaus Neuhaus. “*Comparison of Strand-Specific Transcriptomes of Enterohemorrhagic Escherichia Coli O157:H7 EDL933 (EHEC) under Eleven Different Environmental Conditions Including Radish Sprouts and Cattle Feces.*” *BMC Genomics* 15, no. 1 (May 9, 2014): 353. DOI: [10.1186/1471-2164-15-353](https://doi.org/10.1186/1471-2164-15-353). [9]

Richard Landstorfer, Svenja Simon, Steffen Schober, Daniel A. Keim, Siegfried Scherer, and Klaus Neuhaus. “*Differentiation of true ncRNAs from translated ‘non-coding’ RNAs in Escherichia coli O157:H7 EDL933 by ribosomal footprinting.*” *Nucleic Acids Res* under review:NAR-02924-Y-02014.

Klaus Neuhaus, Richard Landstorfer, Lea Fellner, Svenja Simon, Harald Marx, Olga N. Ozoline, Andrea Schafferhans, Bernhard Küster, Daniel A. Keim, and Siegfried Scherer. “*Translatomics reveals novel, evolutionarily young orphan genes in Escherichia coli O157:H7 (EHEC).*” (in preperation)

Klaus Neuhaus, Richard Landstorfer, Svenja Simon, Harald Marx, Bernhard Küster, Daniel A. Keim, and Siegfried Scherer. “*Translatome data derived by ribosomal footprinting are more sensitive and can substitute proteome data obtained by mass spectrometry, indicating gene*

VI-3 Further Challenges in Biological Data Visualization

expression of about 2/3 of the genes in EHEC in one condition.” (in preparation)

Klaus Neuhaus, Richard Landstorfer, Katharina Mir, Svenja Simon, Steffen Schober, Daniela Oelke, Daniel A. Keim, Martin Bossert, Siegfried Scherer. “*Hundreds of novel overlapping protein-coding genes in enterohemorrhagic Escherichia coli O157:H7 revealed by ribosomal footprinting and strand-specific transcriptomes.*” (in preparation)

The *VisExpress*-system has been developed recently, therefore, no studies using *VisExpress* have been applied yet. However, *VisExpress* was used in a pair analytic study on a real data set. See section [Biological findings - Use case](#) (p. 107) for the biological findings revealed in this study.

VI-3 Further Challenges in Biological Data Visualization

Based on my experiences during my PhD, I identified further specific biological characteristics for some visualization and visual analytics challenges which I have not covered in Section [I-4](#) and which are not addressed in this thesis.

Leveraging Interactions

Interactions are the means to analyze and explore large data sets, for instance, by switching between views, different levels of detail or by calling up details-on-demand. However, seamless and intuitive interactions are needed to support the user, instead of overwhelming the user. Furthermore, interactions need to be designed and incorporated to support the generation of knowledge [[Sacha et al., 2014](#)]. However, how to support the exploration, verification and knowledge generation loop, is still an open research challenge.

Specific biological interaction characteristics. Biological tasks are often ill-defined and data sets are complex. This makes it hard for visualization experts to abstract tasks and map the mental model with the visual and interaction design of a visual analysis system. However, especially interactions are essential to allow an exploration of complex data set and to support knowledge generation.

As biologists normally have no training in visualizations, interactions need to be absolutely intuitive for the target users. However, intuitive does not necessarily mean “easy”, it rather means that the interactive analysis must be in-line with the mental model of biologists. Therefore, visualization experts should not underestimate biologists (see [Do not underestimate biologists](#) (p. 110) for an example). Furthermore, dedicated trainings could also be helpful to teach biologists new interaction concepts.

VI. CONCLUDING REMARKS AND PERSPECTIVES

Exploiting Collaboration

Collaboration refers to the collaboration between several domain experts. I see two sub-directions here. First, to support direct collaborative analyses of domain experts. Thus, two analysts work directly together. For some applications, a distinction of independent sub-tasks might be possible. In other cases, the work with different devices is an interesting direction. The second direction is a successive data analysis, e.g., if one analyst wants to double check the results of a colleague or wants to continue the analysis of a colleague. In my opinion, methods to support collaborative analyses have the potential to support knowledge generation with visual analytics systems. Annotated insights, for instance, could have the potential to be starting points for new insights for a subsequent analyst. See also “The Science of Analytical Reasoning (Chapter 2)” in [Thomas and Cook, 2005].

Specific biological collaboration characteristics. Biological data sets are often very complex and information dense and the gaining of insights is often dependent on background knowledge and the point of view. In the pair analytics study of the *VisExpress*-system, I observed that some findings were interpreted and judged differently by the domain experts. A possibility to facilitate the individual differences is to capture, present, and communicate analysis results among the colleagues. This would also support the verification loop of the knowledge generation model for visual analytics [Sacha et al., 2014]. Notes, for instance, could save time to rediscover findings already known and also provide starting points for new insights.

Minimizing Hardware constraints and exploiting Hardware possibilities

Real-time interactions with visualizations and data are an important aspect for visualization systems. This includes the pure data rendering but also underlining computational methods. A comparison of the influences of different parameter settings is cumbersome, if each calculation needs hours of runtime. On the other side, new visualization and computing devices like touch-tablets, tablets and large scale high-resolution displays are emerging. How to effectively use these, also in combination, is an open challenge.

Specific biological hardware characteristics. High-throughput technologies like next-generation-sequencing produce massive amounts of data which need to be processed before the data analysis and visualization. Pre-processings like read mapping (see [Mapping](#) (p. 38)) can last hours, making a comparison between different mapping algorithms ineffective. Furthermore, I see a high potential for using different devices in collaborative analysis of biologists.

Building an Infrastructure

Many visual analytics solutions use their own infrastructures. Often very specific problems are addressed which are not abstracted to a higher level. Therefore, it can often not be assessed whether a system could be applied to another use case in another domain. Furthermore, visual analytics infrastructures and common components are often developed over and over again which hampers rapid prototyping and the development of reusable and adaptable systems.

Specific biological infrastructure characteristics. For biological applications many bioinformatics systems have been developed. However, a consideration of these as related work is often cumbersome, as the visual and interaction design is not documented efficiently and tasks that can be addressed with the systems are not defined. Thus, often each system needs to be installed and tested to assess its relevance. A better documentation and benchmark data sets would help to improve this.

VI. CONCLUDING REMARKS AND PERSPECTIVES

My own Publications

- [1] Jan Aerts, Jean-Fred Fontaine, Michael Lappe, Raghu Machiraju, Cydney Nielsen, Andrea Schafferhans, Svenja Simon, Matthew O. Ward, and Jarke J. van Wijk. Sequence Data Visualization, 2013. Chapter in Biological Data Visualization (Dagstuhl Seminar 12372). Dagstuhl Reports, Volume 2, Issue 9, Chapter 4.2, pages 143-148. Editors: Carsten Görg and Lawrence Hunter and Jessie Kennedy and Sean O’Donoghue and Jarke J. van Wijk.
- [2] Feeras Al-Masoudi, Daniel Seebacher, Mario Schreiner, Manuel Stein, Christian Rohrdantz, Fabian Fischer, Svenja Simon, Tobias Schreck, and Daniel A. Keim. Similarity-Driven Visual-Interactive Prediction of Movie Ratings and Box Office Results. In *VAST Challenge 2013 - Award for Effective Visualization*, 2013.
- [3] Michael Behrisch, James Davey, Svenja Simon, Tobias Schreck, Daniel Keim, and Jörn Kohlhammer. Visual Comparison of Orderings and Rankings. In M. Pohl and H. Schumann, editors, *EuroVis Workshop on Visual Analytics*, pages 7–11. The Eurographics Association, 2013.
- [4] Fernando Benites, Svenja Simon, and Elena Sapozhnikova. Mining Rare Associations between Biological Ontologies. *PLoS ONE*, 9(1):e84475, January 2014.
- [5] Min Chen, Julian Heinrich, Jessie Kennedy, Andreas Kerren, Falk Schreiber, Svenja Simon, Christian Stolte, Corinna Vehlow, Michel Westenberg, and Bang Wong. Uncertainty Visualization, 2013. Chapter in Biological Data Visualization (Dagstuhl Seminar 12372). Dagstuhl Reports, Volume 2, Issue 9, Chapter 4.6, pages 154-155. Editors: Carsten Görg and Lawrence Hunter and Jessie Kennedy and Sean O’Donoghue and Jarke J. van Wijk.
- [6] Mennatallah El Assady, Daniel Hafner, Michael Hund, Alexander Jäger, Wolfgang Jentner, Christian Rohrdantz, Fabian Fischer, Svenja Simon, Tobias Schreck, and Daniel A. Keim. Visual Analytics for the Prediction of Movie Rating and Box Office Performance. In *VAST Challenge 2013 - Award for Effective Analytics*, 2013.

MY OWN PUBLICATIONS

- [7] Lea Fellner, Niklas Bechtel, Michael A. Witting, Svenja Simon, Philippe Schmitt-Kopplin, Daniel Keim, Siegfried Scherer, and Klaus Neuhaus. Phenotype of *htgA* (*mbiA*), a recently evolved orphan gene of *Escherichia coli* and *Shigella*, completely overlapping in antisense to *yaaW*. *FEMS Microbiology Letters*, 350(1):57–64, 2014.
- [8] Daniel A. Keim, Leishi Zhang, Miloš Krstajić, and Svenja Simon. Solving Problems with Visual Analytics: Challenges and Applications. *Journal of Multimedia Processing and Technologies, Special Issue on Theory and Application of Visual Analytics*, 3(1):1–11, 2012.
- [9] Richard Landstorfer, Svenja Simon, Steffen Schober, Daniel Keim, Siegfried Scherer, and Klaus Neuhaus. Comparison of strand-specific transcriptomes of enterohemorrhagic *Escherichia coli* O157:H7 EDL933 (EHEC) under eleven different environmental conditions including radish sprouts and cattle feces. *BMC Genomics*, 15(1):353, May 2014.
- [10] Daniela Oelke, Halldór Janetzko, Svenja Simon, Klaus Neuhaus, and Daniel A. Keim. Visual Boosting in Pixel-based Visualizations. *Computer Graphics Forum*, 30:871–880, 2011.
- [11] Svenja Simon, Reinhard Guthke, Thomas Kamradt, and Oliver Frey. Multivariate analysis of flow cytometric data using decision trees. *Frontiers in Microbiology*, 3(00114), April 2012.
- [12] Svenja Simon, Sebastian Mittelstädt, Daniel A. Keim, and Michael Sedlmair. Bridging the gap of domain and visualization experts with a Liaison. In E. Bertini, J. Kennedy, and E. Puppo, editors, *Eurographics Conference on Visualization (EuroVis) - Short Papers, Cagliari, Italy, 25 - 29 May 2015*, pages 127–133. The Eurographics Association, 2015.
- [13] Svenja Simon, Sebastian Mittelstädt, Bum Chul Kwon, Andreas Stoffel, Richard Landstorfer, Klaus Neuhaus, Anna Mühlig, Siegfried Scherer, and Daniel A. Keim. VisExpress - Visual exploration of differential gene expression data. *Information Visualization*, 2015.
- [14] Svenja Simon, Daniela Oelke, Richard Landstorfer, Klaus Neuhaus, and Daniel A. Keim. Visual Analysis of Next-Generation Sequencing Data to Detect Overlapping Genes in Bacterial Genomes. In *2011 IEEE Symposium on Biological Data Visualization, October 23 - 24, Providence, Rhode Island, USA*, volume 1, pages 47–54, October 2011.
- [15] Svenja Simon, Daniela Oelke, Richard Landstorfer, Klaus Neuhaus, and Daniel A. Keim. Visual Analysis of RNAseq Data to Detect Overlapping Genes in Bacterial Genomes.

MY OWN PUBLICATIONS

Poster at VIZBI 2012 - VISUALIZING BIOLOGICAL DATA 2015, Heidelberg, Germany (Poster), 2012.

- [16] Svenja Simon, Daniela Oelke, Klaus Neuhaus, and Daniel A. Keim. Visualization of the sensitivity of BLAST to changes in the parameter settings. Poster at GCB 2012 - German Conference on Bioinformatics 2012, Jena, Germany (Poster), September 2012.

MY OWN PUBLICATIONS

References

- [Aflitos et al., 2015] Aflitos, S. A., Sanchez-Perez, G., de Ridder, D., Fransz, P., Schranz, M. E., de Jong, H., and Peters, S. A. (2015). Introgression browser: high-throughput whole-genome SNP visualization. *The Plant Journal*, 82(1):174–182.
- [Albinsson et al., 2007] Albinsson, L., Lind, M., and Forsgren, O. (2007). Co-Design: An Approach to Border Crossing, Network Innovation. In Cunningham, P. and Cunningham, M., editors, *Expanding the Knowledge Economy: Issues, Applications, Case Studies*. IOS Press, Amsterdam.
- [Altschul et al., 1990] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410.
- [Altschul et al., 1997] Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402.
- [Angiuoli et al., 2008] Angiuoli, S. V., Gussman, A., Klimke, W., Cochrane, G., Field, D., Garrity, G. M., Kodira, C. D., Kyrpides, N., Madupu, R., Markowitz, V., Tatusova, T., Thomson, N., and White, O. (2008). Toward an Online Repository of Standard Operating Procedures (SOPs) for (Meta)genomic Annotation. *OMICS A Journal of Integrative Biology*, 12:137–141.
- [Arias-Hernandez et al., 2011] Arias-Hernandez, R., Kaastra, L., Green, T., and Fisher, B. (2011). Pair Analytics: Capturing Reasoning Processes in Collaborative Visual Analytics. In *System Sciences (HICSS), 2011 44th Hawaii International Conference on*, pages 1–10.
- [Baroukh et al., 2011] Baroukh, C., Jenkins, S. L., Dannenfels, R., and Ma’ayan, A. (2011). Genes2wordcloud: a quick way to identify biological themes from gene lists and free text. *Source Code for Biology and Medicine*, 6(1):15.

REFERENCES

- [Bateman et al., 2008] Bateman, S., Gutwin, C., and Nacenta, M. (2008). Seeing things in the clouds: The effect of visual features on tag cloud selections. In *Proceedings of the Nineteenth ACM Conference on Hypertext and Hypermedia*, HT '08, pages 193–202, New York, NY, USA. ACM.
- [Battke et al., 2010] Battke, F., Symons, S., and Nieselt, K. (2010). Mayday - integrative analytics for expression data. *BMC Bioinformatics*, 11(1):121.
- [Behrens et al., 2002] Behrens, M., Sheikh, J., and Nataro, J. P. (2002). Regulation of the overlapping *pic/set* locus in *Shigella flexneri* and enteroaggregative *Escherichia coli*. *Infect Immun*, 70(6):2915–2925.
- [Berndt and Clifford, 1994] Berndt, D. J. and Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In *AAAI-94 workshop on knowledge discovery in databases*, pages 229–248.
- [Bertini et al., 2011] Bertini, E., Strobel, H., Braun, J., Deussen, O., Groth, U., Mayer, T., and Merhof, D. (2011). HiTSEE: A visualization tool for hit selection and analysis in high-throughput screening experiments. In *Biological Data Visualization (BioVis), 2011 IEEE Symposium on*, pages 95–102.
- [Besemer and Borodovsky, 2005] Besemer, J. and Borodovsky, M. (2005). GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Research*, 33(suppl 2):W451–W454.
- [Besemer et al., 2001] Besemer, J., Lomsadze, A., and Borodovsky, M. (2001). 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.
- [Beyer and Holtzblatt, 1997] Beyer, H. and Holtzblatt, K. (1997). *Contextual Design: Defining Customer-Centered Systems*. Elsevier.
- [Borodovsky and McIninch, 1993] Borodovsky, M. and McIninch, J. (1993). GenMark: parallel gene recognition for both DNA strands. *Computers & Chemistry*, 17(2):123–133.
- [Bratteteig, 1997] Bratteteig, T. (1997). Mutual learning - Enabling cooperation on systems design. *Proceedings of IRIS'20*, pages 1–20.
- [Brehmer and Munzner, 2013] Brehmer, M. and Munzner, T. (2013). A multi-level typology of abstract visualization tasks. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2376–2385.

-
- [Brooks, 1996] Brooks, Jr., F. P. (1996). The computer scientist as toolsmith II. *Commun. ACM*, 39(3):61–68.
- [Bruls et al., 2000] Bruls, M., Huizing, K., and van Wijk, J. (2000). Squarified treemaps. In *Proceedings of Joint Eurographics and IEEE TCVG Symposium on Visualization*, pages 33–42.
- [Bruno et al., 2010] Bruno, V. M., Wang, Z., Marjani, S. L., Euskirchen, G. M., Martin, J., Sherlock, G., and Snyder, M. (2010). Comprehensive annotation of the transcriptome of the human fungal pathogen *Candida albicans* using RNA-seq. *Genome Research*, 20(10):1451–1458.
- [Camtasia,] Camtasia. <http://www.techsmith.de/camtasia.html>. Accessed at 23.04.2015 18:00 EST.
- [Carver et al., 2012] Carver, T., Harris, S. R., Berriman, M., Parkhill, J., and McQuillan, J. A. (2012). Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics*, 28(4):464–469. 00106.
- [Chirico et al., 2010] Chirico, N., Vianelli, A., and Belshaw, R. (2010). Why genes overlap in viruses. *Proceedings of the Royal Society B: Biological Sciences*, 277(1701):3809–3817.
- [Clark et al., 2011] Clark, M. B., Amaral, P. P., Schlesinger, F. J., Dinger, M. E., Taft, R. J., Rinn, J. L., Ponting, C. P., Stadler, P. F., Morris, K. V., Morillon, A., Rozowsky, J. S., Gerstein, M. B., Wahlestedt, C., Hayashizaki, Y., Carninci, P., Gingeras, T. R., and Mattick, J. S. (2011). The Reality of Pervasive Transcription. *PLoS Biol*, 9(7):e1000625.
- [Coordinators, 2013] Coordinators, N. R. (2013). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 41(D1):D8–D20.
- [Creative Commons, 4] Creative Commons (4). <http://creativecommons.org/licenses/by/4.0>. Accessed at 23.04.2015 18:00 EST.
- [de Hoon et al., 2005] de Hoon, M. J., Makita, Y., Nakai, K., and Miyano, S. (2005). Prediction of transcriptional terminators in *Bacillus subtilis* and related species. *PLoS Computational Biology*, 1:e25.
- [Delcher et al., 2007] Delcher, A. L., Bratke, K. A., Powers, E. C., and Salzberg, S. L. (2007). Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics*, 23(6):673–679.

REFERENCES

- [Delcher et al., 1999] Delcher, A. L., Harmon, D., Kasif, S., White, O., and Salzberg, S. L. (1999). Improved microbial gene identification with GLIMMER. *Nucleic Acids Research*, 27(23):4636–4641.
- [Dorigo et al., 2006] Dorigo, M., Birattari, M., and Stutzle, T. (2006). Ant colony optimization. *IEEE Computational Intelligence Magazine*, 1(4):28–39.
- [Fellows, 2013] Fellows, I. (2013). *wordcloud: Word Clouds*. R package version 2.4.
- [Fisher et al., 2011] Fisher, S., Barry, A., Abreu, J., Minie, B., Nolan, J., Delorey, T. M., Young, G., Fennell, T. J., Allen, A., Ambrogio, L., Berlin, A. M., Blumenstiel, B., Cibulskis, K., Friedrich, D., Johnson, R., Juhn, F., Reilly, B., Shammass, R., Stalker, J., Sykes, S. M., Thompson, J., Walsh, J., Zimmer, A., Zwirko, Z., Gabriel, S., Nicol, R., and Nusbaum, C. (2011). A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. *Genome Biol*, 12(1).
- [Flaherty et al., 2011] Flaherty, B. L., Van Nieuwerburgh, F., Head, S. R., and Golden, J. W. (2011). Directional rna deep sequencing sheds new light on the transcriptional response of anabaena sp. strain pcc 7120 to combined-nitrogen deprivation. *BMC Genomics*, 12:332–332.
- [Galperin et al., 2015] Galperin, M. Y., Makarova, K. S., Wolf, Y. I., and Koonin, E. V. (2015). Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Research*, 43(Database issue):D261–D269.
- [Gehlenborg et al., 2005] Gehlenborg, N., Dietzsch, J., and Nieselt, K. (2005). A Framework for Visualization of Microarray Data and Integrated Meta Information. *Information Visualization*, 4(3):164–175.
- [Gehlenborg et al., 2010] Gehlenborg, N., O’Donoghue, S. I., Baliga, N. S., Goesmann, A., Hibbs, M. A., Kitano, H., Kohlbacher, O., Neuweger, H., Schneider, R., Tenenbaum, D., and Gavin, A.-C. (2010). Visualization of Omics Data for Systems Biology. *Nature Methods*, 7(3 Suppl):S56–S68.
- [Gentleman et al., 2004] Gentleman, R. C., Carey, V. J., Bates, D. M., and others (2004). Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 5:R80.
- [Gonçalves et al., 2009] Gonçalves, J. P., Madeira, S. C., and Oliveira, A. L. (2009). BiGGES: integrated environment for biclustering analysis of time series gene expression data. *BMC Research Notes*, 2(124).

-
- [Grady, 2013] Grady, J. O. (2013). *System Requirements Analysis*. Elsevier, second edition edition.
- [Harrower and Brewer, 2003] Harrower, M. and Brewer, C. A. (2003). ColorBrewer.org: An Online Tool for Selecting Colour Schemes for Maps. *The Cartographic Journal*, 40(1):27–37.
- [Healey, 1996] Healey, C. G. (1996). Choosing effective colours for data visualization. In *Proceedings of Visualization'96.*, pages 263–270. IEEE.
- [Heer et al., 2009] Heer, J., Kong, N., and Agrawala, M. (2009). Sizing the Horizon: The Effects of Chart Size and Layering on the Graphical Perception of Time Series Visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '09*, pages 1303–1312, New York, NY, USA. ACM.
- [Hilbert, 1891] Hilbert, D. (1891). Über die stetige Abbildung einer Line auf ein Flächenstück. *Mathematische Annalen*, 38(3):459–460.
- [Hu et al., 2009] Hu, G.-Q., Zheng, X., Zhu, H.-Q., and She, Z.-S. (2009). Prediction of translation initiation site for microbial genomes with TriTISA. *Bioinformatics*, 25(1):123–125.
- [Hung et al., 2012] Hung, J.-H., Yang, T.-H., Hu, Z., Weng, Z., and DeLisi, C. (2012). Gene set enrichment analysis: performance evaluation and usage guidelines. *Briefings in Bioinformatics*, 13(3):281–291.
- [Hyatt et al., 2010] Hyatt, D., Chen, G.-L., LoCascio, P., Land, M., Larimer, F., and Hauser, L. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11(1):119.
- [Jacques et al., 2006] Jacques, P. E., Rodrigue, S., Gaudreau, L., Goulet, J., and Brzezinski, R. (2006). Detection of prokaryotic promoters from the genomic distribution of hexanucleotide pairs. *BMC Bioinformatics*, 7:423.
- [Karlen et al., 2007] Karlen, Y., McNair, A., Perseguers, S., Mazza, C., and Mermod, N. (2007). Statistical significance of quantitative PCR. *BMC bioinformatics*, 8(1):131+.
- [Karolchik et al., 2014] Karolchik, D., Barber, G. P., Casper, J., Clawson, H., Cline, M. S., Diekhans, M., Dreszer, T. R., Fujita, P. A., Guruvadoo, L., Haeussler, M., Harte, R. A., Heitner, S., Hinrichs, A. S., Learned, K., Lee, B. T., Li, C. H., Raney, B. J., Rhead, B., Rosenbloom, K. R., Sloan, C. A., Speir, M. L., Zweig, A. S., Haussler, D., Kuhn, R. M.,

REFERENCES

- and Kent, W. J. (2014). The UCSC Genome Browser database: 2014 update. *Nucleic Acids Research*, 42(Database issue):D764–D770.
- [Keim et al., 2001] Keim, D., Hao, M., Dayal, U., Hsu, M., and Ladisch, J. (2001). Pixel bar charts: A new technique for visualizing large multi-attribute data sets without aggregation. In *Information Visualization, IEEE Symposium on*, pages 113–113. IEEE Computer Society.
- [Keim and Oelke, 2007] Keim, D. and Oelke, D. (2007). Literature fingerprinting: A new method for visual literary analysis. In *IEEE Symposium on Visual Analytics Science and Technology, 2007. VAST 2007*, pages 115–122.
- [Keim et al., 1995] Keim, D. A., Ankerst, M., and Kriegel, H.-P. (1995). Recursive pattern: A technique for visualizing very large amounts of data. In *Proceedings of the 6th conference on Visualization '95*, page 279. IEEE.
- [Keim et al., 2010] Keim, D. A., Kohlhammer, J., Ellis, G., and Mansmann, F. (2010). *Mastering the information age - solving problems with visual analytics*. Eurographics.
- [Keim et al., 2009] Keim, D. A., Kohlhammer, J., Santucci, G., Mansmann, F., Wanner, F., and Schaefer, M. (2009). Visual Analytics Challenges. In *Proceedings of eChallenges 2009*.
- [Keim and Zhang, 2011] Keim, D. A. and Zhang, L. (2011). Solving problems with visual analytics: challenges and applications. In ACM, editor, *Proceedings of I-KNOW 2011, 11th International Conference on Knowledge Management and Knowledge Technologies*. ACM.
- [Keim et al., 2012] Keim, D. A., Zhang, L., Krstajić, M., and Simon, S. (2012). Solving Problems with Visual Analytics: Challenges and Applications. *Journal of Multimedia Processing and Technologies, Special Issue on Theory and Application of Visual Analytics*, 3(1):1–11.
- [Kingsford et al., 2007] Kingsford, C., Ayanbule, K., and Salzberg, S. (2007). Rapid, accurate, computational discovery of rho-independent transcription terminators illuminates their relationship to DNA uptake. *Genome Biology*, 8(2):R22.
- [Kirby and Meyer, 2013] Kirby, R. and Meyer, M. (2013). Visualization Collaborations: What Works and Why. *IEEE Computer Graphics and Applications*, 33(6):82–88.
- [Lesnik et al., 2001] Lesnik, E. A., Sampath, R., Levene, H. B., Henderson, T. J., McNeil, J. A., and Ecker, D. J. (2001). Prediction of rho-independent transcriptional terminators in *Escherichia coli*. *Nucleic Acids Res*, 29:3583–3594.

- [Li et al., 2010] Li, J., Jiang, H., and Wong, W. (2010). Modeling non-uniformity in short-read rates in RNA-seq data. *Genome Biology*, 11(5):R50+.
- [Lloyd and Dykes, 2011] Lloyd, D. and Dykes, J. (2011). Human-Centered Approaches in Geovisualization Design: Investigating Multiple Methods Through a Long-Term Case Study. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2498–2507.
- [Lukashin and Borodovsky, 1998] Lukashin, A. V. and Borodovsky, M. (1998). GeneMark.hmm: New solutions for gene finding. *Nucleic Acids Research*, 26(4):1107–1115.
- [Madupu et al., 2010] Madupu, R., Brinkac, L. M., Harrow, J., Wilming, L. G., Böhme, U., Lamesch, P., and Hannick, L. I. (2010). Meeting report: a workshop on Best Practices in Genome Annotation. *Database : the Journal of Biological Databases and Curation*, 2010:baq001+.
- [Manske and Kwiatkowski, 2009] Manske, H. M. and Kwiatkowski, D. P. (2009). Lookseq: A browser-based viewer for deep sequencing data. *Genome Research*, 19(11):2125–2132.
- [Marioni et al., 2008] Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., and Gilad, Y. (2008). Rna-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, 18:1509–1517.
- [McKenna et al., 2014] McKenna, S., Mazur, D., Agutter, J., and Meyer, M. (2014). Design activity framework for visualization design. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2191–2200.
- [McVeigh et al., 2000] McVeigh, A., Fasano, A., Scott, D. A., Jelacic, S., Moseley, S. L., Robertson, D. C., and Savarino, S. J. (2000). Is1414, an *Escherichia coli* insertion sequence with a heat-stable enterotoxin gene embedded in a transposase-like gene. *Infect Immun*, 68(10):5710–5715.
- [Meyer et al., 2010a] Meyer, M., Munzner, T., DePace, A., and Pfister, H. (2010a). MulteeSum: A Tool for Comparative Spatial and Temporal Gene Expression Data. *Visualization and Computer Graphics, IEEE Transactions on*, 16(6):908–917.
- [Meyer et al., 2009] Meyer, M., Munzner, T., and Pfister, H. (2009). MizBee: A Multiscale Synteny Browser. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):897–904.
- [Meyer et al., 2013] Meyer, M., Sedlmair, M., Quinan, P. S., and Munzner, T. (2013). The nested blocks and guidelines model. *Information Visualization*, page 1473871613510429.

REFERENCES

- [Meyer et al., 2010b] Meyer, M., Wong, B., Styczynski, M., Munzner, T., and Pfister, H. (2010b). Pathline: A Tool For Comparative Functional Genomics. *Computer Graphics Forum*, 29(3):1043–1052.
- [Mittelstädt et al., 2014] Mittelstädt, S., Hao, M. C., Dayal, U., Hsu, M.-C., Terdiman, J., and Keim, D. A. (2014). Advanced visual analytics interfaces for adverse drug event detection. In *Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces*, pages 237–244. ACM.
- [Mittelstädt et al., 2015a] Mittelstädt, S., Jäckle, D., Stoffel, F., and Keim, D. A. (2015a). ColorCAT: Guided Design of Colormaps for Combined Analysis Tasks. In *Proc. of the Eurographics Conference on Visualization (EuroVis 2015: Short Papers)*.
- [Mittelstädt et al., 2014] Mittelstädt, S., Stoffel, A., and Keim, D. A. (2014). Methods for Compensating Contrast Effects in Information Visualization. *Computer Graphics Forum*, 33(3):231–240.
- [Mittelstädt et al., 2015b] Mittelstädt, S., Wang, X., Eaglin, T., Thom, D., Keim, D. A., Tolone, W., and Ribarsky, W. (2015b). An Integrated In-Situ Approach to Impacts from Natural Disasters on Critical Infrastructures. In *IEEE 48th Annual Hawaii International Conference on System Sciences (nominated for Best Paper Award)*.
- [Mortazavi et al., 2008] Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nature Methods*, 5(7):621–628.
- [Munzner, 2009] Munzner, T. (2009). A nested model for visualization design and validation. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):921–928.
- [Munzner, 2014] Munzner, T. (2014). *Visualization Analysis and Design*. A K Peters Visualization Series. CRC Press.
- [Nicol et al., 2009] Nicol, J. W., Helt, G. A., Blanchard, S. G., Raja, A., and Loraine, A. E. (2009). The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets. *Bioinformatics*, 25(20):2730–2731.
- [Nielsen et al., 2010] Nielsen, C. B., Cantor, M., Dubchak, I., Gordon, D., and Wang, T. (2010). Visualizing genomes: techniques and challenges. *Nature methods*, 7(3 Suppl):S5–S15.
- [Oelke et al., 2009] Oelke, D., Hao, M., Rohrdantz, C., Keim, D., Dayal, U., Haug, L., and Janetzko, H. (2009). Visual opinion analysis of customer feedback data. pages 187–194.

-
- [Ozoline and Deev, 2006] Ozoline, O. N. and Deev, A. A. (2006). Predicting antisense RNAs in the genomes of *Escherichia coli* and *Salmonella typhimurium* using promoter-search algorithm PlatProm. *Journal of Bioinformatics and Computational Biology*, 4:443–454.
- [Piringer et al., 2010] Piringer, H., Berger, W., and Krasser, J. (2010). HyperMoVal: Interactive Visual Validation of Regression Models for Real-Time Simulation. *Computer Graphics Forum*, 29(3):983–992.
- [Quail et al., 2008] Quail, M. A., Kozarewa, I., Smith, F., Scally, A., Stephens, P. J., Durbin, R., Swerdlow, H., and Turner, D. J. (2008). A large genome center’s improvements to the illumina sequencing system. *Nat Methods*, 5(12):1005–1010.
- [R Core Team, 2013] R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- [Rangannan and Bansal, 2007] Rangannan, V. and Bansal, M. (2007). Identification and annotation of promoter regions in microbial genome sequences on the basis of DNA stability. *Journal of Bioscience*, 32:851–862.
- [Rapaport et al., 2013] Rapaport, F., Khanin, R., Liang, Y., Pirun, M., Krek, A., Zumbo, P., Mason, C. E., Succi, N. D., and Betel, D. (2013). Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol*, 14(9).
- [Robinson et al., 2010] Robinson, M., McCarthy, D., and Smyth, G. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26:139–40.
- [Rutherford et al., 2000] Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M. A., and Barrell, B. (2000). Artemis: sequence visualization and annotation. *Bioinformatics*, 16(10):944–945.
- [Sacha et al., 2014] Sacha, D., Stoffel, A., Stoffel, F., Kwon, B. C., Ellis, G., and Keim, D. (2014). Knowledge generation model for visual analytics. *Visualization and Computer Graphics, IEEE Transactions on*, 20(12):1604–1613.
- [Saeed et al., 2003] Saeed, A. I., Sharov, V., White, J., Li, J., Liang, W., Bhagabati, N., Braisted, J., Klapa, M., Currier, T., Thiagarajan, M., Sturn, A., Snuffin, M., Rezantsev, A., Popov, D., Ryltsov, A., Kostukovich, E., Borisovsky, I., Liu, Z., Vinsavich, A., Trush, V., and Quackenbush, J. (2003). Tm4: a free, open-source system for microarray data management and analysis. *Biotechniques*, 34(2):374–378.

REFERENCES

- [Saeys et al., 2007] Saeys, Y., Abeel, T., Degroeve, S., and Van de Peer, Y. (2007). Translation initiation site prediction on a genomic scale: beauty in simplicity. *Bioinformatics*, 23(13):i418–i423.
- [Saito et al., 2005] Saito, T., Miyamura, H. N., Yamamoto, M., Saito, H., Hoshiya, Y., and Kaseda, T. (2005). Two-tone pseudo coloring: Compact visualization for one-dimensional data. In *IEEE Symposium on Information Visualization*, pages 173–180. IEEE.
- [Salzberg et al., 1998] Salzberg, S. L., Delcher, A. L., Kasif, S., and White, O. (1998). Microbial gene identification using interpolated Markov models. *Nucleic Acids Research*, 26(2):544–548.
- [Santamaría et al., 2008] Santamaría, R., Therón, R., and Quintales, L. (2008). BicOverlapper: A tool for bicluster visualization. *Bioinformatics*, 24(9):1212–1213.
- [Santamaría et al., 2014] Santamaría, R., Therón, R., and Quintales, L. (2014). BicOverlapper 2.0: visual analysis for gene expression. *Bioinformatics*.
- [Schbath et al., 2012] Schbath, S., Martin, V., Zytynicki, M., Fayolle, J., Loux, V., and Gibrat, J. F. (2012). Mapping reads on a genomic sequence: An algorithmic overview and a practical comparative analysis. *J Comput Biol*.
- [Sedlmair et al., 2012a] Sedlmair, M., Frank, A., Munzner, T., and Butz, A. (2012a). RelEx: Visualization for Actively Changing Overlay Network Specifications. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2729–2738.
- [Sedlmair et al., 2011] Sedlmair, M., Isenberg, P., Baur, D., Mauerer, M., Pigorsch, C., and Butz, A. (2011). Cardiogram: Visual Analytics for Automotive Engineers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 1727–1736, New York, NY, USA. ACM.
- [Sedlmair et al., 2012b] Sedlmair, M., Meyer, M., and Munzner, T. (2012b). Design Study Methodology: Reflections from the Trenches and the Stacks. *IEEE Trans. Visualization and Computer Graphics (Proc. InfoVis)*, 18(12):2431–2440.
- [Shavkunov et al., 2009] Shavkunov, K. S., Masulis, I. S., Tutukina, M. N., Deev, A. A., and Ozoline, O. N. (2009). Gains and unexpected lessons from genome-scale promoter mapping. *Nucleic Acids Research*, 37(15):4919–4931.
- [Shiroguchi et al., 2012] Shiroguchi, K., Jia, T. Z., Sims, P. A., and Xie, S. S. (2012). Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized

- single-molecule barcodes. *Proceedings of the National Academy of Sciences*, 109(4):1347–1352.
- [Shneiderman, 1996] Shneiderman, B. (1996). The eyes have it: a task by data type taxonomy for information visualizations. In *IEEE Symposium on Visual Languages, 1996. Proceedings*, pages 336–343.
- [Silby et al., 2004] Silby, M. W., Rainey, P. B., and Levy, S. B. (2004). IVET experiments in *Pseudomonas fluorescens* reveal cryptic promoters at loci associated with recognizable overlapping genes. *Microbiology*, 150:518–520.
- [Spinuzzi, 2005] Spinuzzi, C. (2005). The Methodology of Participatory Design. *Technical Communication*, 52(2):163–174.
- [Thomas and Cook, 2005] Thomas, J. and Cook, K., editors (2005). *Illuminating the Path: Research and Development Agenda for Visual Analytics*. IEEE Press.
- [Thorvaldsdóttir et al., 2013] Thorvaldsdóttir, H., Robinson, J. T., and Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, 14(2):178–192.
- [van den Elzen and van Wijk, 2014] van den Elzen, S. and van Wijk, J. J. (2014). Multivariate network exploration and presentation: From detail to overview via selections and aggregations. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2310–2319.
- [van Wijk, 2006] van Wijk, J. (2006). Bridging the gaps. *IEEE Computer Graphics and Applications*, 26(6):6–9.
- [van Wijk and van de Wetering, 1999] van Wijk, J. and van de Wetering, H. (1999). Cushion treemaps: visualization of hierarchical information. In *1999 IEEE Symposium on Information Visualization, 1999. (Info Vis '99) Proceedings*, pages 73–78, 147.
- [Viegas et al., 2009] Viegas, F., Wattenberg, M., and Feinberg, J. (2009). Participatory visualization with wordle. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1137–1144.
- [Vivancos et al., 2010] Vivancos, A. P., GÅijell, M., Dohm, J. C., Serrano, L., and Himmelbauer, H. (2010). Strand-specific deep sequencing of the transcriptome. *Genome Research*, 20(7):989–999.

REFERENCES

- [Vredenburg et al., 2002] Vredenburg, K., Mao, J.-Y., Smith, P. W., and Carey, T. (2002). A survey of user-centered design practice. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '02, pages 471–478, New York, NY, USA. ACM.
- [Wainer and Francolini, 1980] Wainer, H. and Francolini, C. M. (1980). An empirical inquiry concerning human understanding of two-variable color maps. *The American Statistician*, 34(2):81–93.
- [Wang and Benham, 2006] Wang, H. and Benham, C. J. (2006). Promoter prediction and annotation of microbial genomes based on DNA sequence and structural responses to superhelical stress. *BMC Bioinformatics*, 7:248.
- [Wang et al., 2008] Wang, L., Giesen, J., McDonnell, K. T., Zolliker, P., and Mueller, K. (2008). Color design for illustrative visualization. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1739–1754.
- [Wang et al., 1999] Wang, L. F., Park, S. S., and Doi, R. H. (1999). A novel *Bacillus subtilis* gene, *antE*, temporally regulated and convergent to and overlapping *dnaE*. *J Bacteriol*, 181(1):353–6.
- [Ward et al., 2010] Ward, M. O., Grinstein, G., and Keim, D. A. (2010). *Interactive Data Visualization: Foundations, Techniques, and Application*. A. K. Peters, Ltd.
- [Ware, 2004] Ware, C. (2004). *Information Visualization: Perception for Design*. Morgan Kaufmann, 2nd edition edition.
- [Warnes et al., 2014] Warnes, G. R., Bolker, B., Bonebakker, L., Gentleman, R., Liaw, W. H. A., Lumley, T., Maechler, M., Magnusson, A., Moeller, S., Schwartz, M., and Venables, B. (2014). *gplots: Various R programming tools for plotting data*. R package version 2.14.1, <http://CRAN.R-project.org/package=gplots>.
- [Westenberg et al., 2008] Westenberg, M. A., Van Hijum, S. a. F. T., Kuipers, O. P., and Roerdink, J. B. T. M. (2008). Visualizing genome expression and regulatory network dynamics in genomic and metabolic context. *Computer Graphics Forum*, 27(3):887–894.
- [Xia et al., 2013] Xia, J., Lyle, N. H., Mayer, M. L., Pena, O. M., and Hancock, R. E. W. (2013). INVEX - a web-based tool for integrative visualization of expression data. *Bioinformatics*, 29(24):3232–3234.
- [Yi et al., 2008] Yi, J. S., Kang, Y.-a., Stasko, J. T., and Jacko, J. A. (2008). Understanding and characterizing insights: How do people gain insights using information visualization? In

REFERENCES

Proceedings of the 2008 Workshop on BEyond Time and Errors: Novel evaLuation Methods for Information Visualization, BELIV '08, pages 4:1–4:6, New York, NY, USA. ACM.

[Zuker, 2003] Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res*, 31(13):3406–3415.