Detection and characterisation of RNA processing variation from deep RNA sequencing data

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
MSc. Bioinf., Dipl.-Math. Philipp Drewe
aus Berlin

Tübingen
2014

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	23.09.15
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Prof. Dr. Daniel H. Huson
2. Berichterstatter:	Prof. Dr. Gunnar Rätsch
3. Berichterstatter:	Prof. Dr. Mario Stanke

Abstract

The introduction of high-throughput sequencing technologies has opened unprecedented opportunities for research on the regulation of ribonucleic acid (RNA) processing, which is central to cellular information processing. By enabling accurate and extensive measurements of various properties of cellular RNAs, these techniques allow to systematically investigate the transcriptome and its regulation on a genome-wide scale. The development of computational methods to analyse the resulting data, however, is still lagging behind the advances in experimental data generation.

In this thesis, we present novel approaches to leverage the potential of high-throughput sequencing technologies for studying the regulation of RNA processing. More specifically, we focused on the following three research problems:

First, we investigated how to best extract information from RNA-sequencing (RNA-Seq) data and how to design RNA-Seq experiments in order to maximise their utility for answering the investigated question. For this purpose, we derived a probabilistic model to estimate the utility of RNA-Seq experiments as a function of the experimental parameters for typical analyses such as the identification of transcripts and the detection of differential splicing. Application of our models provided fundamental, experimentally supported insights into how particular experimental parameters influence the amount of information gained from an RNA-Seq experiment. Based on these insights, we suggest strategies for an improved experimental design of transcriptome analysis experiments.

The second investigated aspect was the detection of differential RNA processing based on high-throughput sequencing data. Here, we proposed novel statistical tests to detect changes in RNA processing for two distinct settings: When the gene annotation is complete (which is often the case for model organism) and for the case where the gene annotation is incomplete or unknown (as it is the case for non-model organism or pathological phenotypes). We showed that both on realistically simulated and on experimental data our newly developed tests out-competed state-of-the-art methods. Furthermore, we showed how our methods could be extended to detect differential RNA secondary structure and to associate changes in RNA processing with genetic variation. Finally, we successfully applied our methods to investigate the role of splicing in *human* cancer cells, to understand mechanisms of nonsense mediated decay in *A. thaliana* and to reveal regulatory structural motives of translation in *human*.

The third investigated aspect was the characterisation of changes in RNA processing. We showed that combining RNA-Seq data with information on genomic variation and transcription factor binding preferences explained causes of gene expression variation. For this, we first performed a comprehensive analysis of gene expression landscape in an *A. thaliana* population. Furthermore, we showed that there is a significant enrichment of genetic variants associated with gene expression in predicted transcription factor binding sites. Finally, we showed that alterations of transcription factor binding sites are a major driver of gene expression variation.

Overall, we addressed different aspects of the detection and characterisation of RNA processing. Using our new methods we have gained novel insights into the regulation of RNA processing. However, the work has also shown that there are still several open questions, which should be addressed in future studies.

Zusammenfassung

Die Regulierung der Ribonukleinsäure (RNS)-Prozessierung ist von zentraler Bedeutung für die zelluläre Informationsverarbeitung. Die Einführung von Technologien zur Hochdurchsatzsequenzierung (HTS) hat zur weiteren Erforschung dieses Gebietes neue Chancen eröffnet. Da diese Techniken umfangreiche und genaue Messungen verschiedener Eigenschaften der zellulären RNSs erlauben, ermöglichen sie die genomweite systematische Untersuchung des Transkriptoms und dessen Regulierung. Die Entwicklung von Methoden zur Analyse der resultierenden Daten ist jedoch nicht so fortgeschritten wie die experimentellen Datenerzeugung.

In unserer Arbeit präsentieren wir neue Ansätze, um das Potenzial der HTS zur Untersuchung der Regulation der RNS-Prozessierung auszuschöpfen. Hierbei konzentrierten wir uns auf die folgenden drei Aspekte:

Zum ersten, wie Informationen aus den RNS-Sequenzierungs (RNS-Seq)-Daten extrahiert werden können und wie RNS-Seq-Experimente konzipiert werden müssen, um einen maximalen Nutzen zu generieren. Zu diesem Zweck haben wir, abhängig von den Parametern des jeweiligen Experiments, probabilistische Modelle hergeleitet, um die Nützlichkeit der RNS-Seq-Experimente für gängige Analysen, wie beispielsweise die Identifizierung von Transkripten und die Erkennung von differentiellem Spleissen, zu bestimmen. Die Anwendung unserer Modelle ermöglicht es, grundsätztliche, durch experimentelle Daten bestätigte Einsichten zu erlangen, wie die experimentellen Parameter den Informationsgewinn von RNS-Seq-Experimenten beeinflussen. Auf diesen Erkenntnissen basierend, schlagen wir verbesserte Versuchspläne für Experimente zur Transkriptomanalyse vor.

Der zweite Aspekt war die Erkennung von Änderungen in der RNS-Prozessierung mit Hilfe von HTS-Daten. Hier präsentieren wir neuartige statistische Tests, um in zwei verschiedenen Anwendungsgebieten Änderungen in der RNS-Prozessierung zu detektieren: (a) für den Fall der vollständigen Genannotation, was oft bei Modellorganismen zutrifft, aber auch (b) für den Fall dass die Genannotation unvollständig oder unbekannt ist. Letzteres ist häufig bei Nicht-Modellorganismen oder pathologische Phänotypen der Fall. In dieser Arbeit konnten wir zeigen, dass unsere neu entwickelten Tests anderen modernen Methoden überlegen waren, sowohl bei Anwendung auf realistisch simulierten als auch auf experimentellen Daten. Darüber hinaus zeigten wir, wie unsere Methoden erweitert werden können, um Unterschiede in RNS-Sekundärstrukturen zu erkennen und auch um differentielle RNS-Prozessierung mit genetischer Variation zu assoziieren. Schliesslich konnten wir zeigen, wie unsere Methoden angewandt werden können, um erstens die Rolle des Spleissens in menschlichen Krebszellen zu untersuchen, zweitens die dem *Nonsense Mediated Decay* zugrunde liegenden Mechanismen zu verstehen und drittens regulatorische Strukturmotive der Translation im Menschen zu entdecken.

Der letzte Aspekt war die Charakterisierung von Änderungen der RNS-Prozessierung. Wir konnten zeigen, dass die gemeinsame Verwendung von RNS-Seq-Daten mit Informationen zur genomischen Variation und Transkriptionsfaktor (TF)-Bindungspräferenzen ermöglicht, den Mechanismus der Veränderung der Genexpression besser zu verstehen. Dazu haben wir zunächst eine umfassende Analyse der Genexpression in einer *A. thaliana* Population durchgeführt. Ausserdem haben wir demonstriert, dass eine signifikante Anreicherung von mit

Genexpression assoziierten genetischen Varianten in vorhergesagten TF-Bindestellen (TFBS) vorhanden war. Zuletzt haben wir gezeigt, dass Veränderungen in den TFBS in Promotoren eine bedeutende Ursache von Genexpressionsvariation waren.

Zusammenfassend haben wir unterschiedliche Aspekte der Detektion und Charakterisierung von RNS-Prozessierung untersucht. Mit Hilfe unserer neu entwickelten Methoden haben wir neue Einsichten in die Regulation von RNS-Prozessierung erhalten. Unsere Arbeit zeigte jedoch, dass es immer noch viele offene Fragestellungen gibt, welche in zukünftigen Untersuchungen behandelt werden sollten.

Acknowledgements

First and foremost my special thanks go to Gunnar Rätsch, who not only gave me the opportunity to pursue my research interests in his group, but was also an excellent scientific advisor. He created an inspiring and supportive work environment from which I have profited a great deal.

Furthermore, I would like to thank the members of my PhD advisory committee: Daniel Huson, Detlef Weigel and Karsten Borgwardt for providing me with constructive feedback during my PhD.

I would also like to thank Oliver Stegle for many helpful advices, for always being available for fruitful discussions and coffee breaks and who gave me orientation in academic science.

Thanks go also to my colleagues in Tübingen and New York, especially my fellow PhD students Jonas Behr, Fabio De Bona, Regina Bohnert, Elisabeth Georgii, Lisa Hartmann, Andre Kahles, Darya Karelina, David Kuo, Sebastian Schultheiss, Gabriele Schweikert, Vipin Sreedharan, Christian Widmer, Georg Zeller, Yi Zhong, but also the other members of our group Geraldine Jean, Marius Kloft, Kjong Lehmann, Xinghua Lou, Theofanis Karaletsos and Andre Noll, with whom I had countless great discussions and an amazing time.

I would also like to thank my collaborators and co-authors without whom much of the work that is presented in this thesis would not have been possible: Richard Clark, Timothy Hughes, Richard Mott, Dino Sejdinovic, Kamini Singh, Lisa Smith, Oliver Stegle, Heiko Strathmann, Andreas Wachter, Hans-Guido Wendel, Matthew Weirauch and Andrew Wolfe.

Finally, I would like to thank Jonas Behr, Cristina Boss, Jürgen Drewe, Andre Kahles, Kjong Lehmann and Nino Shervashidze for giving me critical feedback for my thesis.

This work was funded by the Max Planck Society and the Memorial Sloan Kettering Cancer Center.

In this work I will follow the scientific practice of using the pronoun "we" to indicate my collaborators, the reader and myself.

Contents

List of Abbreviations

A Adenine

AS Alternative splicing

auPRC Area under precision recall curve

auROC Area under ROC curve

bp Base pair

BRE B recognition element

C Cytosine

CDF Cumulative distribution function

cDNA Complementary DNA

ChIP-Seq Chromatin immunoprecipitation sequencing

circRNA Circular RNA

DNA Deoxyribonucleic acid

DPE Downstream promoter element

eQTL Expression QTL

FDR False discovery rate

FPR False positive rate

FWER Familywise error rate

G Guanine

Gb Giga base

GO Gene ontology

GQS G-quadruplex

HTS High-throughput sequencing

IRES Internal ribosome entry site

K-S Kolmogorov-Smirnov

lncRNA Long non-coding RNA

MAGIC Multiparent advanced generation inter-cross

Mb Mega base

miRNA Micro RNA

MMD Maximum mean discrepancy

mRNA Messenger RNA

NGS Next generation sequencing

NMD Nonsense mediated decay

PARS Parallel analysis of RNA structure

PCR Polymerase chain reaction

pdf Probability density function

PIC Preinitiation complex

poly(A) Polyadenylation

pre-mRNA Precursor mRNA

PSSM Position specific scoring matrix

PWM Position weight matrix

QTL Quantitative trait loci

RKHS Reproducing kernel Hilbert space

RNA Ribonucleic acid

RNA Ribonucleic acid

RNA-Seq RNA-sequencing

RNAP RNA polymerase

ROC Receiver operating characteristics

RPKM Reads per kilo base per million

rRNA Ribosomal RNA

RT-qPCR Real-time quantitative PCR

RV Random variable

siRNA Small interfering RNA

snRNP Small nuclear RNA

T Thymine

TF Transcription factor

TFBS Transcription factor binding site

TPR True positive rate

tRNA Transfer RNA

TSS Transcription start site

U Uracil

UMP Uniformly most powerful

UTR Untranslated region

wt wild type

1. Introduction

All living cells integrate genetic and external information to respond to the environment. The cellular information processing that performs this integration comprises numerous pathways, in which *ribonucleic acids* (RNA) play a central role. While acting as a carrier of genetic information these molecules also regulate a multitude of pathways. Thus understanding RNA processing and its regulation is a key to answering many fundamental questions in molecular biology such as how cellular information flow is orchestrated and to understand how the information that is encoded in the genome mediates the various phenotypes of the cells. However, despite advances in understanding RNA biology, many aspects of RNA processing and its regulation remain unclear.

Before the introduction of high-throughput sequencing technologies, detection and quantification of RNA molecules on a large scale has represented a major bottleneck in biomedical research. This was particularly hindering for systematically studying RNA regulation as it typically requires measuring RNA under various conditions to generate hypotheses on the cause of changes of RNA abundances, thus needing a substantial number of measurements. But also the identification and quantification of the cellular RNA molecules, which are a major determinant of the cellular identity, was challenging, thus making it difficult to map the transcriptome.

This bottleneck has vanished with the recent introduction of high-throughput sequencing technologies, which have radically transformed the research on RNA biology. These new technologies allow the simultaneous detection and quantification of many RNA molecules with unprecedented accuracy thus allowing to systematically detect variation of RNA abundances, thereby providing the means to detect and characterise the underlying causative variation in RNA processing. Therefore, these new transcriptome analysis technologies have enabled a shift from hypothesis-driven to data-driven research and thus provide new perspectives on common regulatory mechanisms.

The data resulting from these methods is challenging for several reasons: Firstly, due to the huge number of produced data humans cannot any more analyse the results without the help of computational methods. For example, if the sequences resulting from a typical RNA-sequencing (RNA-Seq) experiment would be printed out in a book format, the resulting text would require over 1.25×10^6 pages (and would weigh over 5 tons). Secondly, the data are typically complex and information extraction from them is non-trivial. Finally, the data are generally noisy and thus, its stochastic component has to be taken into account when performing inference.

These challenges in the analysis of high-throughput sequencing data have made *bioinformatics* methods that can efficiently process the resulting data and that can account for its stochastic nature indispensable for interpreting the data.

However, even though high-throughput sequencing of RNA is now feasible since several years, the robust detection of many aspects of differential RNA processing variation and its characterisation still remains an open problem. In this work, we will present new methods to address this open problem in data analysis of RNA-Seq experiments, the most popular high-throughput transcriptome analysis technique. In particular, we developed new methods and

approaches to detect and characterise changes in RNA processing upon various types of perturbations. These methods were tested on simulated data as well as in appropriate biological model organisms (the fly *Drosophila melanogaster* and the plant *Arabidopsis thaliana*). Furthermore, we applied them to *A. thaliana* and *human* to investigate regulatory mechanisms of RNA processing.

Thesis Structure and Contributions

Chapter 2 establishes the scientific framework of this thesis. We begin by presenting an overview on RNA processing. Next, we will discuss high-throughput sequencing methods and how the data they produce can be processed with existing bioinformatics approaches. The chapter will also introduce selected concepts from statistics and machine learning that will be used throughout this thesis.

In Chapter 3 we will address the information extraction from stochastic RNA-Seq data in the context of experimental design for RNA-Seq experiments. Here, we will analyse how different parameter choices of RNA-Seq experiments influence two types of transcriptome analyses that are commonly performed: The identification of expressed transcripts and the detection of differential transcript expression between two RNA samples. For this, we first will establish the extraction of relevant information from RNA-Seq data for these two tasks. Based on these insights we will derive probabilistic models for the information gain of these tasks and then we will apply these models in order to assess the influence of various parameter choices on the information gain. Finally, we will show that the insights from the modelling are supported by experimental evidence and we will derive general guidelines for RNA-Seq experiments. The probabilistic model for the transcript identification as well as the analysis of the experimental data is part of the following publication:

- L. M. Smith, L. Hartmann, **P. Drewe**, R. Bohnert, A. Kahles, C. Lanz, G. Rätsch, Multiple insert size paired-end sequencing for deconvolution of complex transcriptomes, *RNA Biology*, 9 (5), 596-609, 2012

The author's contributions to this work are stated in the publication. The probabilistic model and the results for the detection of differential transcript expression are unpublished work by Philipp Drewe and Gunnar Rätsch.

In Chapter 4, we will develop novel methods to detect changes in post-transcriptional RNA processing of genes from RNA-Seq data and will discuss the application of these methods to reveal modes of RNA processing regulation. Commonly, detecting differential RNA processing for a gene between two conditions requires first quantifying the transcript abundances, which represent a snapshot of the different products of the processing. Then a statistical test is applied to detect whether these transcript abundances are significantly different between the two conditions. The estimation of transcript abundances, however, is complex and inherently unstable and thus subsequent testing for differences in the estimates is problematic. This motivates the development of statistical tests that do not require prior quantification. Here, we will propose a series of robust statistical tests (rDiff) to address this need.

In the first part of the chapter, we will derive a parametric test (rDiff.poisson) that uses the gene annotation to test for differential regulation of transcripts. Next, we will show how the nonparametric *Maximum Mean Discrepancy* (MMD) test can be applied to detect changes when a gene annotation is not reliable or available. The elaboration of these two statistical tests was a joint work of Oliver Stegle, Philipp Drewe, Regina Bohnert, Karsten Borgwardt

and Gunnar Rätsch, which was published in:

- O. Stegle*, **P. Drewe***, R. Bohnert, K. Borgwardt, G. Rätsch,
  Statistical tests for detecting differential RNA-transcript expression from read counts,
  Nature Precedings, http://dx.doi.org/10.1038/npre.2010.4437.1, 2010

Next, we will show how rDiff.poisson and the MMD-test can be extended to account for biological variation while testing (rDiff.parametric and rDiff.mmd, respectively) and thus provide more reliable detection of differential RNA processing. Besides these extensions, we will present an approach to increase the sensitivity for detecting changes in alternative splicing of the nonparametric testing strategies (rDiff.nonparametric). This work and the author's contributions to it were published in:

- **P. Drewe**, O. Stegle, L. Hartmann, A. Kahles, R. Bohnert, A. Wachter, K. Borgwardt, G. Rätsch,
  Accurate detection of differential RNA processing,
  Nucleic Acids Research, 41 (10), 5189-5198, 2013

Beside the published contributions, we will also present unpublished alternative embeddings for the nonparametric tests, which was a joint work by Philipp Drewe and Gunnar Rätsch. In this chapter, we will furthermore derive a statistical test to detect changes in RNA secondary structure (sDiff), which is work of Philipp Drewe and we will present a reformulation of rDiff.mmd to allow association of RNA processing with genetic variation (rDiff.gmmd). The latter contribution was joint yet unpublished work by Philipp Drewe, Heiko Strathmann, Dino Sejdinovic and Gunnar Rätsch.

In the second part of Chapter 4, we will first assess the performance of our methods on realistic simulations as well as on experimental data. Furthermore, we will show how these methods can be applied to detect differential splicing in cancer, differential translation as well as to study regulation of nonsense mediated decay. The work on differential splicing in cancer was unpublished work by Philipp Drewe, Ahmet Zehir and Gunnar Rätsch. The other two applications have been published in:

- A. Wolfe*, K. Singh*, Y. Zhong, **P. Drewe**, V. Rajasekhar, V. Sanghvi, K. Mavrakis, J. Roderick, J. Van der Meulen, J. Schatz, C. Rodrigo, M. Jiang, C. Zhao, P. Rondou, E. de Stanchina, J. Teruya-Feldstein, M. Kelliher, F. Speleman, J Porco, J. Pelletier, G. Rätsch, G. Wendel,
  RNA G-quadruplexes cause eIF4A-dependent oncogene translation in cancer,
  Nature, 513 (7516), 65-70, 2014
- G. Drechsel*, A. Kahles*, A. Kesarwani, E. Stauffer, J. Behr, **P. Drewe**, G. Rätsch, A. Wachter,
  Nonsense-mediated decay of alternative precursor mRNA splicing variants is a major determinant of the Arabidopsis steady state transcriptome,
  The Plant Cell, 25 (10), 3726-3742, 2013

The author's contributions to these works were listed in the respective publications.

Finally, to provide easy access to selected methods of rDiff, we have created a Galaxy module and integrated rDiff into the Oqtans online transcriptome analysis toolbox. This work was published in:

- V. T. Sreedharan, S. J. Schultheiss, G. Jean, A. Kahles, R. Bohnert, **P. Drewe**, P. Mudrakarta, N. Görnitz, G. Zeller, G. Rätsch,

* contributed equally

Oqtans: The RNA-seq Workbench in the Cloud for Complete and Reproducible Quantitative Transcriptome Analysis,
Bioinformatics, 30 (9), 1300-1301, 2014

In Chapter 5, we will exemplify the characterisation of changes in RNA processing for gene expression in an *A. thaliana* population. To lay the foundation for a comparison of the gene expression between different *A. thaliana* strains, we will first establish how gene expression can be reliably estimated in presence of genetic variation, i.e. when the sequences and structures of genes can be different between strains. We will then quantify gene expression and perform an analysis of the expression patterns in the population. This part of the chapter and the author's contribution to it were published in:

- X. Gan*, O. Stegle*, J. Behr*, J. G. Steffen*, **P. Drewe***, K. L. Hildebrand, R. Lyngsoe, S. J. Schultheiss, E. J. Osborne, V. T. Sreedharan, A. Kahles, R. Bohnert, G. Jean, P. Derwent, P. Kersey, E. J. Belfield, N. P. Harberd, E. Kemen, C. Toomajian, P. X. Kover, R. M. Clark, G. Rätsch, R. Mott,
Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*,
Nature, 477 (7365), 419-423, 2011

Next, we will analyse the extent to which alterations of transcription factor binding sites underlie the observed changes in gene expression between different strains of the population. This analysis was performed by Philipp Drewe, Oliver Stegle and Gunnar Rätsch and was part of the following publication:

- M. T. Weirauch, A. Yang, M. Albu, A. Cote, A. Montenegro-Montero, **P. Drewe**, H. S. Najafabadi, S. A. Lambert, I. Mann, K. Cook, H. Zheng, A. Goity, H. van Bakel, J. Lozano, M. Galli, M. Lewsey, E. Huang, T. Mukherjee, X. Chen, J. S. Reece-Hoyes, S. Govindarajan, G. Shaulsky, A. J. M. Walhout, F. Bouget, G. Rätsch, L. F. Larrondo, J. R. Ecker, T. R. Hughes,
Determination and inference of eukaryotic transcription factor sequence specificity,
Cell, 158 (6), 1431-1443, 2014

Finally, we will present the fraction of the total variation in gene expression that can be explained by genetic variation in predicted transcription factor binding sites. This was unpublished work by Philipp Drewe, Oliver Stegle and Gunnar Rätsch.

Supplemental work not included into this thesis

In this work, we will not discuss the author's contributions to the following publications:

- S. Heinrich, E. Geissen, J. Kamenz, S. Trautmann, C. Widmer, **P. Drewe**, M. Knop, N. Radde, J. Hasenauer, S. Hauf,
Determinants of robustness in spindle assembly checkpoint signalling,
Nature Cell Biology, 15 (11), 1328-1339, 2013
- C. Widmer, **P. Drewe**, X. Lou, S. Umrania, S. Heinrich, G. Rätsch,
GRED: graph-regularized 3D shape reconstruction from highly anisotropic and noisy images,
arXiv preprint, arXiv:1309.4426, 2013,
- J. Behr, A. Kahles, Y. Zhong, V. T. Sreedharan, **P. Drewe**, G. Rätsch,
MITIE: Simultaneous RNA-Seq-based transcript identification and quantification in multiple samples,
Bioinformatics, 29 (20), 2529-2538, 2013

- M. J. Dubin, P. Zhang, D. Meng, M. Remigereau, E. J. Osborne, F. P. Casale, **P. Drewe**, A. Kahles, B. Vilhjalmsson, J. Jagoda, S. Irez, V. Voronin, Q. Song, Q. Long, G. Rätsch, O. Stegle, R. M. Clark, M. Nordborg,
  DNA methylation variation in Arabidopsis has a genetic basis and shows evidence of local adaptation,
  eLife, under revision, 2014

2. Background

2.1. Overview

In this chapter, we will present the general scientific framework upon that this thesis is built. In the first part of the chapter, we will introduce important concepts in RNA Biology. Next, we will discuss experimental high-throughput sequencing techniques that allow examining RNA processing on a system level. Following this, we will present existing bioinformatics approaches to process the data that are generated by the high-throughput sequencing techniques and present a small excursion on modelling of transcription factor binding affinities. In the penultimate part of this chapter, we will present statistical methods to model and analyse high throughput sequencing data. In the last part of this chapter, we will introduce general concepts in machine learning and how they can be used to model RNA processing.

2.2. RNA Biology

Cells can alter their appearances and properties (*phenotypes*) and respond to many environmental changes. The diversity of phenotypes that can be adopted by cells having the same hereditary information (*genetic information*) becomes apparent in multi-cellular organisms, where cells can adopt distinct roles. The basic building block of the genetic information is the *gene* [79], which we define in this work as "locatable regions of genomic sequence, corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions and/or other functional sequence regions" [131]. The genes are localised on large cellular molecules, the *chromosomes* [120].

The genetic information in genes is encoded in an alphabet of four different nucleotide bases [11]: *adenine* (A), *cytosine* (C), *guanine* (G) and *thymine* (T). These nucleotide bases are a building block of the nucleotides that have first been characterised in 1869 by Friedrich Miescher in Tübingen [118]. The nucleotides can be chained together and constitute the *deoxyribonucleic acid* (DNA). These chains can form a double helix that is maintained by *Watson-Crick base pairing* (hydrogen bonds between A and T as well as between C and G nucleotides) of nucleic acids between two chains (strands) [184]. Two chains that bind in this way are referred to as being complementary to each other.

The genetic information that is encoded in genes can be transferred, in a process called transcription, to another class of nucleotide chains, the ribonucleic acids (RNA). These molecules are long polymers composed of four nucleic acids: *adenine*, *cytosine*, *guanine*, and *uracil* (U). RNA is chemically similar to DNA but has a lower melting point, thus being less stable. Because RNA molecules are less stable they can also take distinct shapes (secondary structures) that play an important role in their function and regulation. During transcription, an enzyme called *RNA polymerase* (RNAP) synthesises an RNA chain. In this process, the RNAP moves along the DNA and determines (*reads*) the nucleotides in the sequence. It appends for each read DNA nucleotide a corresponding RNA nucleotide to the growing RNA chain, i.e. an U for an A, G for C, C for G, A for T. By having the one-to-one correspondence between the

nucleotides of the DNA and the produced RNA, the genetic information is copied from the gene to the RNA.

There are different classes of RNA molecules that perform diverse functions and regulate many cellular processes: The *messenger RNA* (mRNA) acts as an information carrier for protein synthesis (*translation*) by a molecule complex (*ribosome*). This role of mRNAs in the directed information flow from DNA over RNA to protein was first postulated by Francis Crick in 1958 [35] and is referred to as the *central dogma of molecular biology*. Besides their function as information carrier, other classes of RNAs play a major role in the translational machinery, such as *ribosomal RNAs* (rRNAs) that are part of the ribosome or *transfer RNAs* (tRNAs) that transport amino acids to the ribosome during translation. There also exist RNAs that play an active role in the regulation of other biological pathways, such as *micro RNA* (miRNA), the recently discovered *circular RNA* (circRNA), *long non-coding RNA* (lncRNA) and *small interfering RNA* (siRNA). In the remainder of this work we refer to the entirety of RNA molecules in a cell as the *transcriptome*.

Overall, RNAs are central in cellular information processing and an integral part of many pathways. In the following, we therefore discuss in more detail several aspects of the processing of mRNAs and their regulation. We first revisit the transcription, then introduce splicing and other post transcriptional modifications and finally discuss translation (see Fig. 2.1 for an illustration of cellular RNA processing).

2.2.1. Transcriptional Regulation

The transcription of a gene can be divided into three successive steps: *initiation*, *elongation* and *termination* [105]. Initiation of transcription starts at the *core promoter*, a region of the DNA that contains distinct nucleotide patterns (*sequence elements*). Specific proteins for transcription (*general transcription factors*) are recruited and bind to these sequence elements. This defines a site where the transcription of a gene can start (*transcription start site* (TSS)). The most frequent of these sequence elements for protein coding genes is the *TATA box*, a sequence element that contains the four nucleotides TATA. However, also other sequence elements and patterns, such as the *B recognition element* (BRE), *downstream promoter element* (DPE) or *CG-rich regions* (CpG islands) can substitute the TATA box for transcription initiation [105]. The binding of the general transcription factors is regulated mainly by the accessibility of the DNA through methylation and the chromatin structure of the DNA [105].

After binding to the DNA, the general transcription factors then guide the RNAP II to the TSS, where they together form the DNA binding *preinitiation complex* (PIC). The general transcription factors are always required to initiate transcription but their presence alone leads only to a low (*basal*) level of transcription. In contrast to the general transcription factors, other transcription factors (TF) are not sufficient to initiate transcription on their own, but in presence of general transcription factors they can drastically alter the transcriptional efficiency. These TFs have in general two or more domains, one that recognises specific sequences (*DNA binding domain*), the *transcription factor binding site* (TFBS) and at least one other domain (*activation domain*) that interact with other proteins [105].

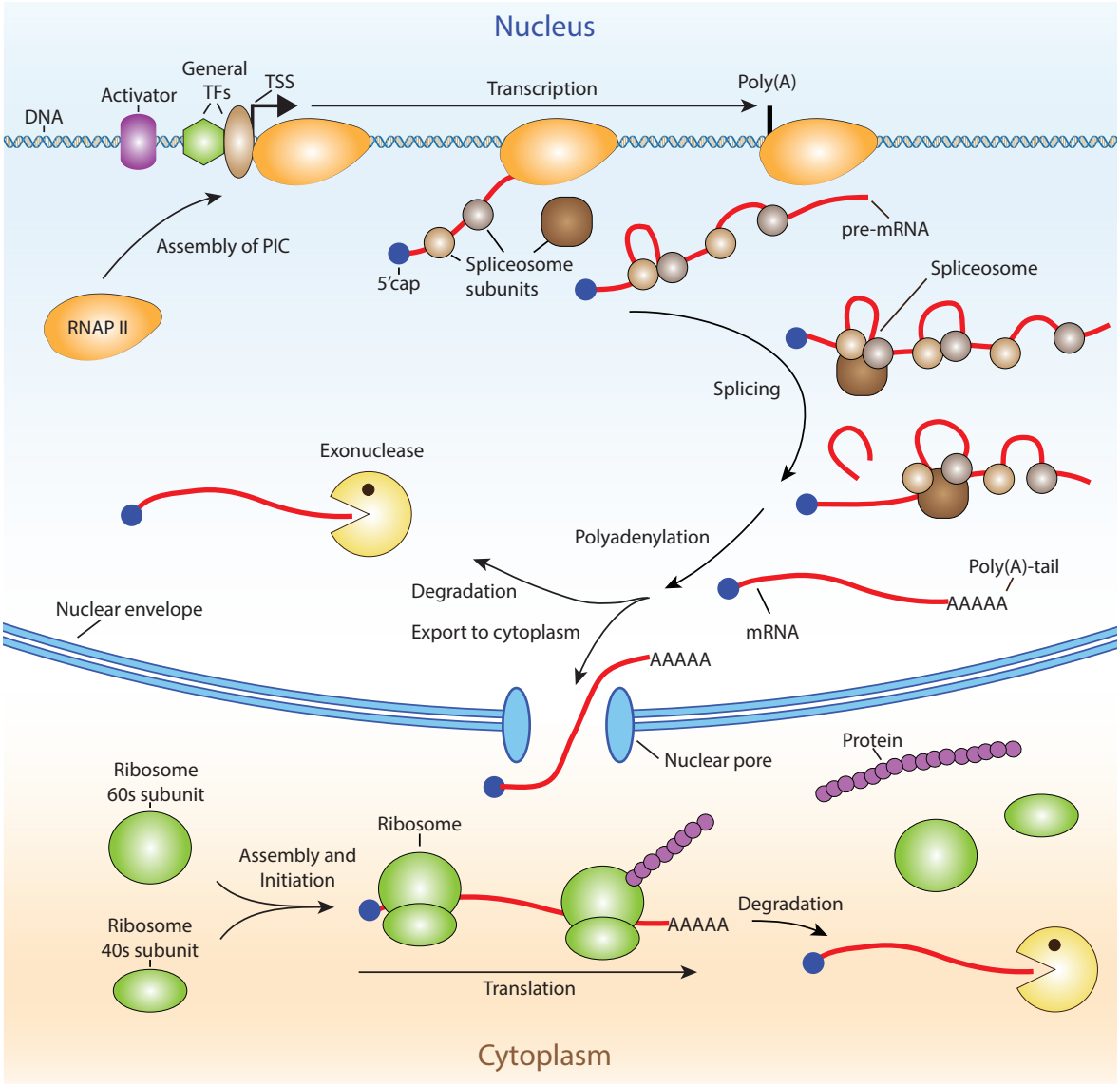

Figure 2.1.: Cellular processing of mRNA. The mRNA synthesis starts in the nucleus with the recruitment of transcription factors and assembly of the preinitiation complex (PIC) on the DNA. The RNA polymerase (RNAP) II then starts transcription to produce the pre-mRNA. During transcription, the 5'cap is ligated to the nascent RNA molecule and the spliceosome subunits start to bind to it. Transcription ends when the RNAP II reaches the Poly(A)-site. Then, the pre-mRNA is released from the RNAP II. Following this, the spliceosome removes parts of the pre-mRNA (introns) and covalently binds the remaining parts (exons) together. Subsequently, the poly(A)-tail is added to pre-mRNA resulting in the mRNA. Molecules that have not correctly been spliced or polyadenylated are degraded by exonucleases. The other mRNAs are exported to the cytoplasm. There they are translated. For this, first the ribosome assembles at the 5' end of the mRNA. Subsequently the ribosome moves along the mRNA and appends the amino acids to the growing amino acid chain. Finally, the amino acid chain is released from the ribosome and the ribosome disassembles. Translation can then repeated or the mRNA can be degraded by exonucleases.

The TFs can be classified by their effect on transcription of a gene, i.e. whether they increase transcription (activators) or decrease is (*repressors*). Activators and repressors can work synergistically, meaning that the change in transcription caused by two TFs can be more than the sum of the changes that would be caused by each of the two TFs separately. The combination of activators and repressors therefore allows differentiated regulation of transcription. This makes well-defined responses to environmental changes possible.

Usually, the TFs bind in within 200 base pairs (bp) of the TSS. Together with the core promoter, their binding sites constitute the *promoter*. The promoter, however is not the only location that determines transcriptional regulation. TFBS can also be located in *enhancers*. Enhancers are, similarly to promoters, bound by transcription factors that interact with the PIC. In contrast to promoters, however, enhancers can act independently of their direction, even when they are distant from the TSS. Thereby, enhancers can influence the transcriptional activity of many genes. These enhancers further increase the richness of transcription regulation and allow well-orchestrated cellular transcription, e.g. to establish tissue specific regulation of gene expression [160].

After initiation of transcription, the elongation starts with the separation of the RNAP II from the PIC. Following this, the RNAP II then moves along the DNA and produces the precursor of the mRNA (Precursor mRNA (pre-mRNA)). During the elongation of the nascent RNA, a terminal methylation cap (5' cap) is linked to the first transcribed nucleotide. It prevents the nascent transcript from degradation through *exonucleases* and is also critical for the ribosome to recognise the transcript and subsequently initiate translation.

The transcription terminates at the *polyadenylation* (poly(A)) *site*, a nucleotide pattern containing a long stretch of adenines. Finally, the pre-mRNA is released. Termination can also take place at alternative locations, leading to nascent transcripts of different lengths. This leads to distinct regulatory regimes for the transcripts, as the 3'-ends of them often harbour binding sites for miRNAs that can induce degradation of the transcripts. However, the mechanisms that determine the alternative poly(A) usage and the implications of their different usages are still not fully understood.

For a more detailed excerpt of transcription and its regulation we refer to [32, 105, 111, 147].

2.2.2. Post-transcriptional RNA Processing

After transcription the synthesised pre-mRNAs are further processed. Complexes of *small nuclear RNAs* (snRNAs) and proteins, the *spliceosomes*, form at specific sites (*splice sites*) of the nascent transcript. Subsequently, the spliceosomes remove parts of the transcript (*introns*) and covalently bind the remaining parts (*exons*) together in a process that is referred to as *splicing*. Recent findings suggest that splicing already starts during transcription [192].

After the splicing process, the nascent transcript is cleaved at the poly(A) site, which lies at the 3' end of the transcript. Following cleavage, a stretch of adenines is added to the 3' end of the transcript. Similar to the 5' cap, this polyadenylation prevents the transcript from being targeted by certain cellular degradation mechanisms. The resulting transcript is called mRNA transcript. The newly synthesised mRNA is then subject to quality control mechanisms that degrade wrongly processed transcripts, e.g. those where the 5'cap or the poly(A)-tail is missing. After the poly(A)-tail is added, the mRNA transcript is exported from the nucleus to the cytoplasm, where it is translated to synthesise proteins.

In 1977, it was first discovered in Adenoviruses and shortly thereafter in eukaryotes that different RNAs can be derived from a single gene [17, 31]. This can result from splicing of

different exon combinations (*alternative splicing* (AS)). The different types of transcripts that are produced from a gene are called *isoforms*. The numbers of isoforms that can be produced from a gene can be high. For example, the *Drosophila melanogaster* DSCAM gene up to 38,016 isoforms can be generated [151]. It was recently revealed that alternative splicing is involved in the processing of a large fraction of genes; up to 92-94 % of all *Homo sapiens* genes are alternatively spliced [181]. It was furthermore shown that alternative splicing is tightly regulated and can be tissue specific (e.g. [188]), suggesting that alternative splicing is central for controlled RNA processing.

Alternative splicing leads, together with alternative TSS and poly(a) sites, to an increase in the number of isoforms that can be generated from a gene. This increases the diversity of the proteins that can be generated, without substantially increasing the size of the genome. Additionally to the regulation of transcription, alternative splicing therefore constitutes another layer of RNA processing regulation. The importance of this layer of regulation is reflected by the abundance of diseases that are caused by disruption of splicing (e.g. see [47]).

Common alternative splicing patterns are shown in Fig. 2.2. These include the skipping or inclusion of an exon (*exon skip*), the skipping of an exon that depends on the inclusion of another exon (*mutually exclusive exons*), the inclusion of an intron (*intron retention*) or the use of alternative 5' and 3' ends (*alternative 5' splice site* and *alternative 3' splice site*, respectively). Although the use of alternative TSS and poly(A) sites is not splicing, in the sense that are mediated by the spliceosome, they still lead to different transcripts. In the following, we therefore refer to them for simplicity also as splice events.

For an extensive review of the post-transcriptional processing of the pre-mRNA and mRNA we refer to [21, 30, 90, 133]

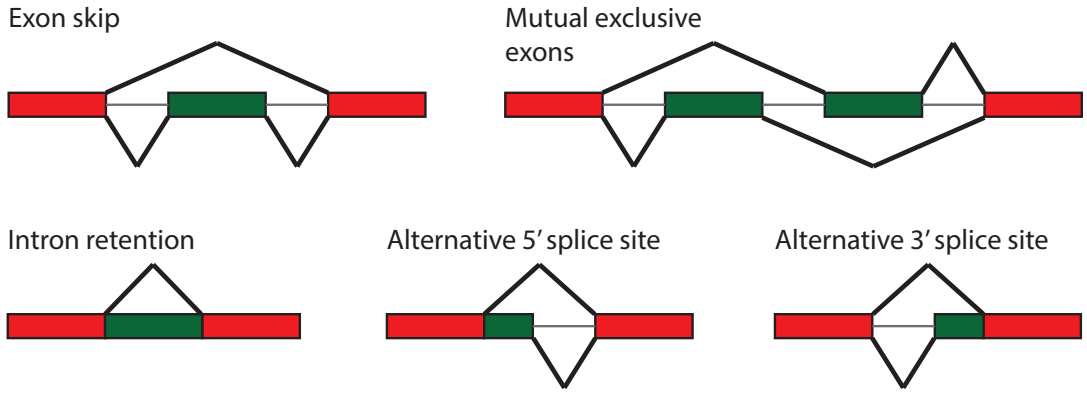

Figure 2.2.: Shown are different types of alternative splicing events. The parts of the pre-mRNA that are always contained in the mRNA are shown in red and the transcript parts that can be spliced in are shown in green. The parts that are never contained are indicated by a thin black line. Pairs of exon-ends that can be joined together during splicing are linked by thick black lines.

2.2.3. Mechanisms and Regulation of Translation

The mature mRNA transcripts are exported, after being spliced and polyadenylated, from the nucleus to the cytoplasm. There they serve as a template for the synthesis of proteins in a process called *translation*. The translation of the mRNA is mediated by ribosomes,

which consist of two subunits: the small ribosomal subunit and the large ribosomal subunit. Ribosomes are either free in the cytosol or bound to the membrane of the rough endoplasmatic reticulum.

The process of translation can be, similar to transcription, divided into three steps: *initiation*, *elongation* and termination. Initiation begins with the recognition of the 5'cap by the small ribosomal subunit. In some rare cases the ribosome also recognises an RNA structural motif that is called *internal ribosome entry site* (IRES) instead of the 5'cap. After the recognition, the small ribosomal subunit binds to the mRNA. From there it scans the mRNA sequence until it reaches the nucleotide triplet (*codon*) A-U-G, termed *start codon*. The region 5' of the start codon is not translated and is therefore referred to as the *5' untranslated region* (UTR). After the small ribosomal subunit is positioned at the start codon, the large ribosomal subunit binds it and the translation is initiated.

During the elongation phase, the ribosome reads one codon at a time and appends a specific amino acid per codon to the nascent amino acid chain and subsequently moves to the next codon where the same procedure is repeated. The ribosome elongates the nascent amino acid chain until it reads a *stop codon* (TAG, TAA, TGA). It then *terminates* the translation and releases the protein from the ribosome. Thereafter, the ribosome dissociates from the mRNA and disassembles. As the region after the stop codon is not translated, it is called the *3'UTR*.

Translation regulation occurs mainly in the initiation step. Known factors that regulate the translation initiation are the availability of the translational cofactors as well as miRNAs that bind the 5'UTR and secondary structures of the 5'UTR [74]. At the level of translation there exist also various quality control mechanisms. If, for example, the ribosome encounters a *premature stop codon*, the ribosome marks the mRNA for degradation by nucleases in a process called *nonsense mediated decay* (NMD). But also the *stalling* of the ribosome during elongation or the missing of a stop codon triggers decay pathways that mediate degradation of the template mRNA. However, compared to the factors that regulate transcription and splicing, the mechanisms that underlie translation regulation are still only partially understood.

For a detailed exposition of the subject we refer to [19, 74, 88, 108, 156].

2.3. High-Throughput Sequencing

2.3.1. DNA Sequencing

A major breakthrough in the analysis of genomes was the development of methods that allow identifying the nucleotides in DNA sequences. This identification, which is commonly referred to as *sequencing*, allowed to analyse how the genetic information is encoded in the genome. The first two methods to sequence DNA were proposed in 1977 by Maxam and Gilbert [114] and by Frederick Sanger [148]. Of these two methods, the chain-termination method developed by Sanger became the predominantly used method, due to the easier protocol. This method is nowadays often referred to as *Sanger sequencing*. The principle behind Sanger sequencing is that the DNA of interest is replicated nucleotide by nucleotide. By subsequently adding nucleotides that have distinct *radioactive labels*, the nucleotide that was incorporated in the growing DNA chain can be identified

In the following twenty-five years after Sanger sequencing was proposed the protocol was further refined. Advances, such as *fluorescent labelling*, *capillary sequencing* and *automation of the sequencing* lead to a higher throughput of sequencing techniques [61], i.e. an increase

in the number of nucleotides that can be identified in a day. These developments paved the way for the sequencing of the genomes of many model organisms such as *Saccharomyces cerevisiae* [57] and *Arabidopsis thaliana* [7]. Finally, in 2001, the first two reconstruction of the *H. sapiens* genome were presented [95, 179]. One major drawback of the Sanger capillary sequencing is however the relatively low throughput of maximally 6 Mega bases (Mb) per day and the high price of 500$ per 1 Mb [86].

In the last decade several new DNA sequencing methods have been introduced that allowed a higher throughput in sequencing at much lower cost per sequenced base than Sanger capillary sequencing, the most popular of these new methods, being *Illumina sequencing* [15], *454 sequencing* [109] and *SOLiD sequencing* [176]. Using the Illumina HiSeq sequencer [1] it is now possible to sequence up to 22 giga bases (Gb) in 7 hours. These methods are sometimes referred to as next generation sequencing (NGS) methods, but given the fact that they are already almost a decade old we refer to them in this work as *high-throughput sequencing* (HTS) methods. The HTS methods all have in common that the amplification and the sequencing is greatly parallelised compared to the first generation sequencing methods. In the following, we illustrate the Illumina sequencing protocol as an example for high throughout sequencing protocols. This sequencing approach is the most widely used one and the results presented in this work are all based on data obtained using this protocol. For an extensive review of other protocols, we refer for example to [115].

The general protocol for Illumina sequencing consists of two main steps, the preparation of a *library* of DNA fragments from the DNA sample (*library preparation*) and the sequencing of these fragments (see Fig. 2.3 for illustration) [15].

In the library preparation the first step is the fragmentation of the sample DNA using for example sonication. This provides a *random fragmentation* of the DNA molecules. Then a gel-based size selection of the fragments is performed to filter for DNA fragments of a certain length (*insert size*). Subsequently, short DNA fragments (*adapters*) are ligated to the double stranded fragments and then the double stranded DNA fragments are denatured. Finally, the single stranded fragments are applied onto a surface (*flow cell*), where they bind via their adapters.

The sequencing itself consists of two steps. In the first step, the fragments that are bound to the flow cell are amplified. This is done by repeatedly adding nucleotides and DNA polymerases, to initiate *polymerase chain reaction* (PCR) bridge amplification and then denaturing the resulting double stranded DNA fragments. This procedure leads to formation of *clusters* of identical fragments on the flow cell. In the second step of sequencing, the fragments in the cluster are sequenced. This is achieved by first adding *labelled nucleotides*, where each nucleotide is linked to a distinct fluorescent dye. The fluorescent labels not only allow identifying the incorporated nucleotide, but also serves as a *terminator* of the polymerization. This guarantees that only a single nucleotide is incorporated. Then the fluorescent dye is excited with laser light and the fluorescence emitted from all clusters is imaged. From the resulting signal intensities of each cluster in each colour channel the incorporated nucleotide can then be inferred. Finally, the fluorescence label is removed from the nucleotide. The second step of the sequencing step can then be repeated to reveal the identity of the fragments base-by-base from one end. The resulting sequences that are obtained from this procedure are called reads. When the fragments are sequenced only from one end the reads are called *single-end reads*. A popular modification of the sequencing protocol also allows sequencing the fragments from both ends. The resulting read pairs are called *paired-end reads*.

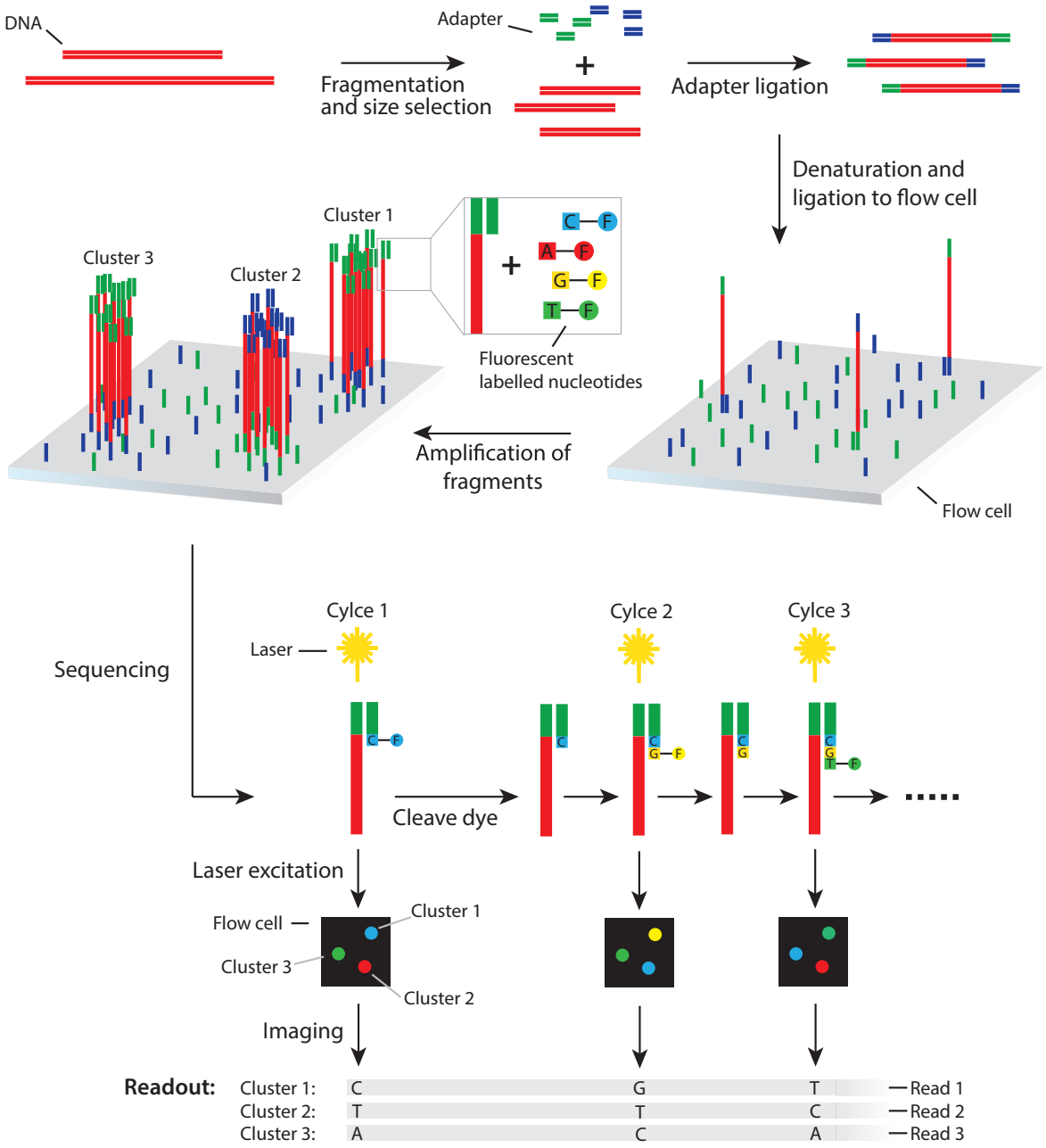

Figure 2.3.: Illustration of Illumina DNA sequencing protocol. First the DNA is fragmented and the fragments are size selected. Then adapters are added ligated to the double stranded DNA fragments and the fragments are denatured. The resulting single stranded DNA fragments are then bound to the flow cell and subsequently amplified until they form clusters. In the last step, the clusters are sequenced. For this, first fluorescent-labelled nucleotides are added. A DNA polymerase then adds the nucleotides to the complementary strand of the fragments. The flow cell is then exited by a laser and fluorescence emitted from all clusters is imaged. Lastly, the incorporated nucleotides are inferred from the image. The cluster sequencing step is then repeated.

After sequencing, the reads can be used to reconstruct the sequence of the DNA (*sequence assembly*) that was in the sample. For this task, many bioinformatics software methods have been proposed (e.g. [101, 125, 162]).

2.3.2. RNA Sequencing

Besides sequencing DNA, high-throughput sequencing methods can also be applied to sequence RNA. This is achieved by first converting the RNA into DNA using a reverse transcriptase [121], a retroviral enzyme that generates a *complementary DNA sequence* (cDNA) from a RNA transcript. The generated cDNA library can, subsequently, be sequenced using standard HTS approaches. The resulting reads allow identification of the RNA molecules in a sample and also allows to infer the relative quantity of these RNA in the sample. Therefore, RNA-Seq allows to get a detailed picture of the transcriptome.

Compared to other methods such as microarrays [72] or *real-time quantitative PCR* (RT-qPCR) [69], RNA-Seq has some desirable properties. Firstly, RNA-Seq has a similar throughput as microarrays but the gene expression measurements obtained with RNA-Seq are far more accurate. This has been shown in [126], where it was observed that RT-qPCR measurements, the gold standard for measuring gene expression, and RNA-Seq gene expression measurements had a Pearson correlation coefficient of 0.98. In contrast microarrays only had a Pearson correlation coefficient of 0.72. Secondly, it was shown in [126] that the *dynamic range*, defined as the highest measured value divided by the lowest measured value, was as high as $8,000$. This is much higher than the dynamic range of microarrays, which was only 60. Therefore, RNA-Seq allows to detect differences in gene expression more accurately than microarrays. Moreover, RNA-Seq has the advantages that it has a *single nucleotide resolution* and can also be used when the genome is not known. Lastly, RNA-Seq typically needs less starting material than microarrays [183] and is therefore more efficient than microarrays. For these reasons, RNA-Seq is very suitable for genome-wide quantification of gene expression.

For practical applications another advantage is that different RNA libraries can be sequenced together. This can be achieved by first ligating an identifier sequence (*barcode*) that is unique for each library to every RNA in the library (*multiplexing*) and then pooling all libraries together. After sequencing, the library of origin for each read can be recovered by the barcode.

A challenge with RNA-Seq is that typically the different classes of RNAs have different abundances. Typically, rRNAs comprise a large fraction of the transcriptome. When studying mRNAs, it is therefore beneficial to filter RNAs having a poly(A) tail using oligo(dT) beads [121] and deplete rRNA, e.g. using the RiboMinus kit (Invitrogen) [186]. This increases the number of mRNA reads when sequencing.

Following the publication of the RNA-Seq protocol, various extensions and modifications have been proposed to gain an even more detailed view of the transcriptome. These extensions and modifications focus on different aspects of the protocol. Some focus on the enrichment or depletion of certain classes of RNA, such as developments to enrich for small RNAs (e.g. [96]) or specific mRNAs (e.g. [155]). Other extensions allow determining the parts of the RNA molecules that are bound by other molecules such as ribosomes or splicing factors (e.g. [64, 73]). Another example for a modification is the use of enzymes that cleave RNA (*RNase*) with different cleave preferences to map the locations where the RNA structure was single- and double-stranded [84]. Overall, these extensions and modifications allow studying many aspects of the transcriptome and laid the foundation for many insights.

2.3.3. Impact of High Throughput Sequencing on Biomedical Research

The impact of HTS methods on research in biology can hardly be underestimated. This is because these methods allow to get an unbiased view on the genome and transcriptome at a low cost and with high throughput. Therefore, HTS methods allow determining the state of cells on a systems-level. They enable researchers for example to genome-wide analyse genomic variation [2, 52, 150], gene regulation [167] and transcriptome plasticity [38, 52], to name only a few applications. Therefore, HTS methods help to understand cellular regulation and thus can provide novel perspectives on mechanisms of diseases [29]. Consequently, HTS methods have been established as standard tools in biology and increasingly also in medicine.

2.4. Bioinformatics

2.4.1. High-Throughput Sequencing Data Analysis

With the increasing popularity of RNA-Seq experiments, the analysis of the data that are generated by these experiments has emerged as an important branch of bioinformatics. The challenges in the analysis lie in the large amount of data to be processed and the stochastic nature of the data. This requires efficient algorithms to extract the relevant information from the read sequences. The typical analysis consists of several steps. The first step of the analysis is the control of the quality of the reads in order to identify potential mistakes in the library preparation and to filter out reads that have low quality. After quality control, a *de novo* transcript prediction, using tools such as [59, 138], can be performed in order to reconstruct the transcriptome. The reconstruction can be useful when working for example with incomplete genomes or annotations. A more common approach however is to first determine the genomic loci from which the sequenced read stem (*read mapping*). Since in most cases both the number of reads and the size of the genome are large, efficient mapping is non-trivial. The mapping is further complicated by the fact that non-continuous mapping positions must be considered due to splicing. Another difficulty in the mapping stems from the fact that reads sometimes do not perfectly map to the genome as they can have sequencing errors. Established tools to perform the mapping of RNA-Seq reads include [39, 75, 172]. After mapping, reads can than be used for example to identify expressed transcripts (see Sec.3 for details), to quantify gene expression and transcript expression or to detect differential gene expression and differential transcript expression (see Sec.4 for details).

For a comprehensive overview of tools to analyse RNA-Seq data, we refer to [4].

2.4.2. Models for Transcription Factor Binding

Many of the fundamental processes that control transcription are regulated by transcription factors. These transcription factors can bind the DNA near the transcription start site and can thereby be recruited for the transcription initiation process. According to the *Michaelis-Menten* model [116], the probability of binding and thus the efficiency of recruitment is determined by two factors: The *concentration* of the transcription factor and the *Gibbs free energy* of its binding to the DNA (*binding affinity*). Both, increases in protein concentration of a transcription factor and an increase in its binding affinity, lead to an increased probability of binding. Depending on the sequence specificity of the binding affinity of transcription factors, two types of binding can be distinguished: Binding that requires a specific nucleotide pattern (*specific binding*) and general binding to DNA independent of the nucleotide pattern of the

DNA (*unspecific binding*). Especially the specific DNA-binding of transcription factors has been studied intensively as it underlies the distinct gene expression profiles in different cellular contexts. In the following, we discuss the approaches that have been developed to study specific binding. For this, we first present models that have been designed to characterise the sequences that a transcription factor binds to and discuss bioinformatics approaches to fit these models. Next, we present experimental approaches to measure transcription factor binding.

Assume that a set of sequences $S_1, \ldots, S_n$, each having length k, is given that are bound by a transcription factor. Assume furthermore that the transcription factor binds at the same position in each sequence. Then, we denote by s_i^j the j-th nucleotide of the sequence S_i and by f_i^j the frequency of nucleotide j at position i in all sequences. Under these assumptions, the most simple model describes the binding sites of a transcription factor by a sequence $M = m_1 \ldots m_k$ of length k, where $m_i, i \in \{1, \ldots, k\}$ is the nucleotide that occurs most often in the sequences $S_1, \ldots, S_n$ at position i, i.e. $m_i := \arg\max_{j \in \{A,C,G,T\}} f_i^j$. With this model, potential transcription factor binding sites in a DNA sequence D can be predicted by determining the location where M is a subsequence of D. A drawback of this model, however, is that it represents each position by only one nucleotide. Therefore, the model does not account for the fact that different positions in a transcription factor binding site contribute not equally to the binding affinity, i.e. positions in the binding site where the identity of the nucleotide is not important are considered as being equally important as positions where a certain nucleotide is required for binding. Furthermore, the model does not provide stable predictions as subtle changes in the binding frequency can lead to substantially different predictions, which is biochemically not plausible.

A more elaborate model of transcription factor binding sites describes these sites as the model before but uses an extended alphabet, the *IUPAC alphabet* [34]. This alphabet has a letter for all combinations of nucleotides (e.g. the letter M for nucleotides A or C) and thus allows a more fine-grained description of the binding site. But also this model suffers to some extent from the same shortcomings as the model described above.

A third model to describe transcription factor binding sites, that does not suffer from these shortcomings, is the *position weight matrix* (PWM). This model describes a binding site by the matrix P of nucleotide frequencies at each position as shown below:

$$P = \begin{pmatrix} f_1^A & f_2^A & \cdots & f_k^A \\ f_1^C & f_2^C & \cdots & f_k^C \\ f_1^G & f_2^G & \cdots & f_k^G \\ f_1^T & f_2^T & \cdots & f_k^T \end{pmatrix}$$

The PWM model of transcription factor binding preferences is also referred to as a motif. A score $P(S)$ for a putative binding site $S = s_1 \ldots s_k$ can then be obtained by $P(S) = \prod_{i=1}^{k} f_i^{s^i}$. This model has the appealing property that its score is directly related to the binding affinity as under the assumption that the contributions to the binding energy of all positions are additive, the binding energy E is proportional to:

$$E \propto \sum_i \log \frac{P_{i,j}}{b_j},$$

where b_j is the background frequency of nucleotide j in the genome [166]. This relation between the binding energy and the normalised log transform of the frequencies $(\log \frac{P_{i,j}}{b_j})_{i,j}$ has motivated the use of $P(S) = \sum_{i=1}^{k} \log \frac{P_{i,s^i}}{b_{s^i}}$ as an alternative scoring function for a putative

binding site S. The matrix $(\log \frac{P_{i,j}}{b_j})_{i,j}$ is then referred to as the *position specific scoring matrix* (PSSM). The PWM and PSSM models of TFBSs are often visualised using sequence logos [152] that show the frequencies of each nucleotides and the information content at every position (for see Fig. 2.4 an example).

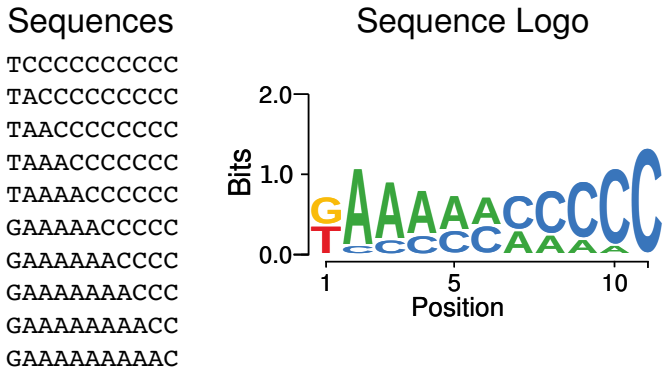

Figure 2.4.: Illustration of a sequence logo. Shown on the left are binding sites of a transcription factor. Shown on the right is the logo that is inferred when assuming an uniform background distribution of the nucleotide frequencies.

For learning of the models that were presented above, it was assumed that a set of known binding sites is available. However, in many applications this is not the case. Such situations arise for example if only the genes that are regulated by a TF are known. To determine putative binding sites in these cases, several bioinformatics approaches exist. These approaches are usually based on distinct characteristics of transcription factor binding sites. The first of these characteristics of transcription factor binding sites is that they are typically enriched in the surrounding sequences of genes that are regulated by this transcription factor compared to the surrounding sequences of other genes. A second characteristic of transcription factor binding sites, which can be used to detect them, is that they are typically better conserved than the surrounding sequence. The third characteristic that can be used is that binding sites tend to cluster with binding sites of other transcription factors. Together these characteristics can be used in order to determine putative transcription factor binding sites for a transcription factor (e.g. [12, 51, 168]).

Besides bioinformatics approaches to identify transcription factor binding sites, also experimental approaches have been proposed. An example for such an approach are *protein binding microarrays* [16]. These arrays allow to investigate the binding specificity of a transcription factor *in vitro*. In this approach the TF of interest is added to a microarray that contains a large array of DNA fragments of a fixed length and different nucleotide patterns. Then it is recorded to which DNA fragments the TF is bound, yielding a set of bound sequence fragments. Another experimental procedure that allows investigating binding *in vivo* is *Chromatin Immunoprecipitation Sequencing* (ChIP-Seq) [80]. In this assay the protein of interest is cross-linked to the DNA it binds using UV light or formaldehyde. Subsequently, the DNA that is not bound by proteins is fragmented and the protein of interest with the cross-linked DNA is purified using antibodies. Finally, the cross-link between the protein and the DNA is broken and the DNA fragments that were bound are sequenced. From these fragments the *in vivo* transcription factor binding sites can then be inferred.

For many transcription factors the binding specificity has already been experimentally determined. Most of these binding profiles are available from the JASPAR [146] or the TRANSFAC [113] databases.

2.5. Statistics

One of the challenges in HTS data analysis is dealing with the noise in the data that is caused by random processes in the library preparation and the sequencing. Moreover, these random processes are not yet fully understood, which adds another layer of complexity to the analysis. Therefore, it is non-trivial to draw conclusions from the data (i.e. to do *inference*), as it is not clear which part of the data is due to random noise and which part is due to the underlying signal of interest. Statistical methods are a well-established way to address this problem. These methods allow to draw inference when noise is present in the data. In the following, we discuss the different types of noise that are commonly encountered in HTS data and present statistical methods that can be used for data analysis.

2.5.1. Fundamental Definitions

We begin by introducing fundamental concepts and basic definitions. The definitions and derivations are adapted from [53], unless otherwise stated. In this work, we denote the set of natural numbers $\{0, 1, 2, \dots\}$ by $\mathbb{N}$ and the real numbers by $\mathbb{R}$. For simplicity, we assume that the space in which observations are represented is either a subset of $\mathbb{N}$ or $\mathbb{R}^n$. We further require that the set of observable random events for discrete spaces $\mathcal{X} \subseteq \mathbb{N}$ is the power set of $\mathcal{X}$ and for sets $\mathcal{X} \subseteq \mathbb{R}^n$ it is the Borel σ-algebra of $\mathcal{X}$. Therefore, we omit the set of random events in the definitions for the remainder of this thesis.

In statistics, the data generation process is typically assumed to be a *random experiment* with (*random events*) as possible outcomes. Formally, this is done by first defining the set $\mathcal{X}$ of random events and then establishing a measure P on this space that assigns a probability of occurring to a set of outcomes (*probability measure*). A probability measure P on a set of events $\mathcal{X}$ can be defined in the following way:

Definition 2.1. *(Probability distribution)*: Let $P : \mathcal{X} \to [0, 1]$ be a function with the following properties:

1. $P(\mathcal{X}) = 1$,
2. If $A_1, A_2, \dots \subset \mathcal{X}$ are pairwise disjoint random events, then

$$P\Big(\bigcup_{i \geq 1} A_i\Big) = \sum_{i \geq 1} P(A_i).$$

Then P is called a *probability measure* or *probability distribution* and the tuple $(\mathcal{X}, P)$ is called a *probability space*.

In order to describe transformations of the events, functions on the set of random events $\mathcal{X}$ (*random variables*) can be defined as follows:

Definition 2.2. *(Random variable)*: Let $\mathcal{X}$ and $\mathcal{Y}$ be two sets and $X : \mathcal{X} \to \mathcal{Y}$ be a function such that for each random event $A \subseteq \mathcal{Y}$, $X^{-1}(A)$ is a random event on $\mathcal{X}$. Then X is called a *random variable* (RV).

These random variables $X : \mathcal{X} \to \mathcal{Y}$ induce a probability distribution on the target space $\mathcal{Y}$ by assigning to each random event $A \subseteq \mathcal{Y}$ the probability $P(X^{-1}(A))$, where P is a probability measure on $\mathcal{X}$. Thereby, they provide a *functional description* of the probability distribution under the assumptions explained below. This description of the probability space by a random variable can be, depending on the application, more convenient to work with than the description by a measure.

Definition 2.3. *(Probability density function)*: Let $(\mathcal{X}, P)$ be a probability space and $p : \mathcal{X} \to \mathbb{R}^+$ be a random variable. Depending on $\mathcal{X}$ being discrete ($\mathcal{X} \subseteq \mathbb{N}$) or continuous ($\mathcal{X} \subseteq \mathbb{R}^n$) there are two cases:

Discrete case If $\sum_{x\in\mathcal{X}} p(x) = 1$ and $P(A) = \sum_{x\in A} p(x)$ for all $A \in \mathcal{X}$, then p is called a *probability density function* (pdf) of P on $\mathcal{X}$.

Continuous case If $\int_{\mathcal{X}} p(x)dx = 1$ and $P(A) = \int_A p(x)dx$ for all $A \in \mathcal{X}$, then p is called a *probability density function* (pdf) of P on $\mathcal{X}$.

Probability density functions on a Borel space that satisfy the conditions in Def. 2.3 can also induce a unique probability measure on $\mathcal{X}$. Moreover, for each such pdf p there is one and only one probability measure P such that p is the pdf of P [53].

2.5.2. Probability Distributions for High-Throughput Sequencing

Probability distributions allow to describe random processes. In the following, we therefore introduce probability distributions that describe various aspects of HTS data.

Random Distributions

We begin by introducing distributions that describe the number of reads from an RNA transcript that are obtained by sequencing an RNA sample. Assume for this, that in the sample under investigation there are T different transcripts with N_t copies of each transcript $t \in \{1, \dots, T\}$. If we assume furthermore, that after fragmentation and size selection there are in total N fragments in the library and that from these K fragments contain a position P of a transcript t. Then, if $n, n \leq N$ reads are sequenced, the resulting reads covering position P can be modelled by a *hypergeometric distribution* $\mathcal{H}_{N,n,K}$. In statistics this distribution is used to model sampling from a set without *putting back* the drawn sample to the set after drawing it. The hypergeometric distribution can be defined as shown below.

Definition 2.4. *(Hypergeometric distribution)*: Let $\mathcal{X} = \{0, \dots, N\}$ be a finite set and $n, K \in \{0, \dots, N\}$. Then the distribution $\mathcal{H}_{N,n,K}$, with

$$\mathcal{H}_{N,n,K}(i) = \frac{\binom{K}{i}\binom{N-K}{n-i}}{\binom{N}{n}}, \text{ for all } i \in \{0, \dots, n\}$$

is called hypergeometric distribution with parameters N, n and K.

One limitation of the hypergeometric distribution is that it has many parameters. This is because for the probability of drawing a certain sample, the history of previous draws has to be considered. However, there exist approximations of the hypergeometric distribution if the number of drawn samples n is considerably smaller than the total number of samples N. In this case, the probability of drawing a certain sample depends only little on the previously

drawn samples. Therefore, the sampling without putting back the drawn sample can be described by sampling with putting back the sample after drawing it [53]. A distribution describing the drawing with putting back the samples that can be used to approximate the hypergeometric distribution is the *binomial distribution* $\mathcal{B}_{n,p}$ with parameters n and $p = \frac{K}{N}$. This binomial distribution can be defined in the following way:

Definition 2.5. *(Binomial distribution)*: Let $\mathcal{X} = \{0, \ldots, n\}, n \in \mathbb{N}$ be a finite set and $p \in [0, 1]$. Then the distribution $\mathcal{B}_{n,p}$,with

$$\mathcal{B}_{n,p}(i) = \binom{n}{i} p^i (1-p)^{n-i}, \text{ for all } i \in \{0, \ldots, n\}$$

is called binomial distribution with parameters n and p.

However, this formulation is numerically hard to compute accurately in cases where n is large and p is small. In this case, the binomial distribution can be approximated as shown in Theorem 2.1.

Theorem 2.1. *(Limit of binomial distribution)*: For a $\lambda > 0$ and a series $(p_n)_{n \geq 1}$ with $np_n \to \lambda$ as $n \to \infty$, the limit $\lim_{n\to\infty} \mathcal{B}_{n,p_n}(i)$ exists and is given by:

$$\lim_{n\to\infty} \mathcal{B}_{n,p_n}(i) = \frac{e^{-\lambda}\lambda^i}{i!}, \text{ for all } i \in \mathbb{N}.$$

Proof. See [53]. □

The distribution resulting from this approximation is known as the *Poisson distribution* $\mathcal{P}_\lambda$ and can be defined as shown below.

Definition 2.6. *(Poisson distribution)*: Let $\mathcal{X} = \mathbb{N}$ be the natural numbers and $\lambda > 0$. Then the distribution $\mathcal{P}_\lambda$ with

$$\mathcal{P}_\lambda(i) = \frac{e^{-\lambda}\lambda^i}{i!}, \text{ for all } i \in \mathbb{N},$$

is called Poisson distribution with *intensity* λ.

The Poisson distribution is commonly used to model the reads that are observed at a certain position of a transcript or genome, due to the fact that it is easy to compute and handle.

The distributions that we have presented so far can be used in order to describe the noise that arises as a consequence of the sequencing. However, they do not model that the RNA sample that is used for sequencing is itself a random sample from a biological system, with its own inherent variance. Specifically, it is not accounted for the fact that the abundance of a transcript is not the same in each cell of a homogeneous cell population, but varies from cell to cell. Therefore, the noise in the read counts is underestimated by the distributions presented above [140]. In order to account for this variation that is induced by the sampling of the RNA sample in modelling the read distribution, the distribution of transcript abundances in a cell population has to be included in the model. This distribution can be well approximated by a *Gamma distribution* $\Gamma_{\alpha,r}$ (see e.g. [50]), which can be defined in the following manner:

Definition 2.7. *(Gamma distribution)*: Let $\mathcal{X} = \mathbb{R}^+$ be the positive real numbers and $r, \alpha > 0$. Then the distribution $\Gamma_{\alpha,r}$, with

$$\Gamma_{\alpha,r}(i) = \frac{\alpha^i}{\Gamma(i)} r^{i-1} e^{-\alpha r} \text{ for all } i \in \mathbb{R}^+$$

is called the Gamma distribution. The special case when $r = 1$, is called the exponential distribution with parameter α.

Incorporating the variation of transcript abundances in methods that model the sequencing can be achieved by assuming that the intensity λ of $\mathcal{P}_\lambda$ follows a Gamma distribution $\Gamma_{\alpha,r}$, i.e. $\lambda \sim \Gamma_{\alpha,r}$ [81]. The resulting distribution of this is the *negative binomial distribution* $\mathcal{NB}_{r,\frac{\alpha}{1+\alpha}}$, which is defined below. This distribution has a long history in modelling count data in biology [23].

Definition 2.8. *(Negative binomial distribution)*: Let $\mathcal{X} = \mathbb{N}$ be the positive numbers, $p \in]0, 1[$ and $r > 0$. Then the distribution $\mathcal{NB}_{r,p}$, with

$$\mathcal{NB}_{r,p}(i) = \binom{-r}{i} p^r (p-1)^i \text{ for all } i \in \mathbb{N}$$

is called negative binomial distribution, where the *general binomial coefficient* $\binom{-r}{i}$ is defined as:

$$\binom{-r}{i} := \frac{(-r)(-r-1)\cdots(-r-i+1)}{i!}$$

The negative binomial distribution can also be parametrised by its mean and variance [81], which we use in this work when convenient.

Beside the negative binomial distribution, another distribution that can be used in order to model read generation is the *Gaussian distribution*. This distribution can be defined as shown below. Using this distribution has the advantage that there exists a large theoretical framework around it and it has many theoretical properties that are appealing to work with. In the setting of HTS data its disadvantage is, however, that it is not capable of capturing the discrete nature of the reads.

Definition 2.9. *(Gaussian distribution)*: Let $\mathcal{X} = \mathbb{R}$ be the real numbers $m \in \mathbb{R}$ and $v > 0$. Then the distribution $\mathcal{N}_{m,v}$, with

$$\mathcal{N}_{m,v}(x) = \frac{1}{\sqrt{2\pi v}} e^{-(x-m)^2/2v} \text{ for all } x \in \mathbb{R}$$

is called the Gaussian or *normal distribution*.

As already mentioned above, the choice of the distribution is influenced by its convenience to work with. Examples of other distributions to model HTS data include the beta-binomial [134] or beta negative binomial [171] distribution, which will not be discussed in detail in this work.

Characterisations of Probability Distributions

The probability distributions that we have introduced above can be characterised by properties of the samples drawn from them, e.g. where the expected average of their samples lies and how far they are spread around this average. These properties can be formally summarised by the *moments*. For example, the first moment (*expected value*) of a probability distribution determines where the mean of infinite samples is located and can be defined as described below.

Definition 2.10. *(Expected value)*:

Discrete random variable Let $(\mathcal{X}, P)$ be a discrete probability space and $X : \mathcal{X} \to \mathbb{R}$ be a real valued random variable. Then, if $\sum_{x \in X(\mathcal{X})} |x| P(X = x)$ exists, we call

$$\mathbb{E}(X) := \sum_{x \in X(\mathcal{X})} x P(X = x)$$

the expected value of X.

Continuous random variable Let $(\mathcal{X}, P), \mathcal{X} \subseteq \mathbb{R}^n$ be a probability space and $X : \mathcal{X} \to \mathbb{R}$ be a real valued random variable. Let furthermore p be the pdf of X. Then, if $\int_{\mathcal{X}} |X(x)| p(x) dx$ exists, we call

$$\mathbb{E}(X) := \int_{\mathcal{X}} X(x) p(x) dx$$

the expected value of X.

Furthermore, we denote by $\mathcal{L}^1(\mathcal{X})$ the space of all random variables for which the expected value is defined. The first moment $\mathbb{E}(X)$ is also called the mean of X and often abbreviated by μ_X or simply μ.

As can be seen from the definition, there exist distributions for which the expected value is not defined, such as for the Cauchy distribution [81]. This is because they have many values that have a high value and a high probability such that $\sum_{x \in X(\mathcal{X})} |x| P(X = x) = \infty$. For the distributions that have been presented before, however, the expected value exists.

Besides the mean, also higher order moments can be defined. This can be done in a similar manner as for the first moment:

Definition 2.11. *(r-th Moment of a random variable)*: Let $(\mathcal{X}, P)$ be a probability space, $X : \mathcal{X} \to \mathbb{R}$ be a real valued random variable and $r \in \mathbb{N}$ such that $r \geq 2$. If $X^r \in \mathcal{L}^1(\mathcal{X})$ then $\mathbb{E}((X - \mu_X)^r)$ is called the *r-th moment* of X. The second moment $\mathbb{E}((X - \mu_X)^2)$ is called the variance of X and is abbreviated by σ_X^2 or σ^2.

A list of the first two moments of the probability distributions that we have presented above is shown in Tab. 2.1.

Table 2.1.: First two moments of probability distributions

Probability distribution	Mean μ	Variance σ^2
Binomial distribution $\mathcal{B}_{n,p}$	np	$np(p-1)$
Hypergeometric distribution $\mathcal{H}_{N,n,K}$	$n\frac{K}{N}$	$n\frac{K}{N}\frac{N-K}{N}\frac{N-n}{N-1}$
Poisson distribution $\mathcal{P}_\lambda$	λ	λ
Negative binomial distribution $\mathcal{NB}_{r,p}$	$\frac{pr}{1-p}$	$\frac{pr}{(1-p)^2}$
Normal distribution $\mathcal{N}_{m,v}$	m	v
Gamma distribution $\Gamma_{\alpha,r}$	$\frac{r}{\alpha}$	$\frac{r}{\alpha^2}$

2.5.3. Statistical Hypothesis Testing

One of the main motivations for using statistical methods in the analysis of HTS data is to make decisions based on the observed data. One statistical technique for doing this is *hypothesis testing*. In this approach, first a hypothesis is stated and then it is decided whether this hypothesis can be supported by the observed data. For example, if it has to be decided whether the expression of a gene is the same in two conditions, the hypothesis that has to be tested is whether the probability distribution of the two gene's expression is the same in both conditions. Given the data, the probability for all random events that support this hypothesis can be computed and used to decide whether the expression was the same in both conditions.

Formally, this can be done as follows. First the set of all potential probability distributions $\mathcal{M} = \{P_\omega | \omega \in \Omega\}$, parametrised by Ω is defined. This set $\mathcal{M}$ is called a *statistical model*. The set of probability distributions is then partitioned into two disjoint subsets. Into the subset $\mathcal{N}$ of probability distributions that support the hypothesis and the subset $\mathcal{A}$ of those probability distributions that do not support the hypothesis. The set $\mathcal{N}$ is called the *Null hypothesis* and the set $\mathcal{A}$ *Alternative hypothesis*. Then, a statistical test that assigns a class to each sample can be defined as shown below.

Definition 2.12. *(Statistical test)*: Let $\mathcal{M} = \mathcal{N} \cup \mathcal{A}$ be a statistical model where $\mathcal{N}$ and $\mathcal{A}$ represent the Null and the Alternative hypothesis, respectively. Let furthermore $P_\omega \in \mathcal{M}$ be a probability distribution on $\mathcal{X}$ and x be a sample of size $n, n \in \mathbb{N}^+$ drawn from P_ω. Then, a *statistical test* ϕ is a random variable $\phi : \mathcal{X}^n \to [0, 1]$ that assigns to each sample x from $\mathcal{X}$ the probability $1 - \phi(x)$ that it is an element of $\mathcal{N}$ and the probability $\phi(x)$ that it is an element of $\mathcal{A}$. The set $\phi^{-1}(0)$ is called the *acceptance region* and the $\phi^{-1}(1)$ the *rejection region*.

For simplicity we focus in the following part of this chapter on a subclass of statistical tests, the *non-randomised test*. These are tests for which ϕ takes only the values zero or one. We would also like to mention that in the context of hypothesis testing the random variable ϕ is called a *statistic*. This term is introduced to emphasise the different interpretations of the function ϕ. A random variable is thought of being the outcome of a random experiment, whereas a statistic is a constructed function to measure aspects of observations [53].

Statistical tests can be categorised into two classes, based on the parameter space Ω: If the parameter space Ω is finite dimensional, a test is called a *parametric test* and otherwise it is called a *nonparametric test*. These two subclasses of tests differ by the assumptions they make on the probability distributions that are tested. Parametric tests typically assume a specific class of distributions and test for properties of these distributions, e.g. whether two Gaussian distributions have the same mean. In contrast, non-parametric tests, in general, pose far fewer assumptions on the distributions and avoid to use any particular property of them [53].

We next discuss the errors a statistical test makes. Statistical tests, as defined above, decide for a given sample x, whether it stems from the Null or the Alternative hypothesis. In this decisions there can be two types of errors. The first is that the Null hypothesis is rejected when it is true, that is $P_\omega \in \mathcal{N}$ and $\phi(x) = 1$. This error is called *error of the first kind* or *Type I error*. The second error is when the Null hypothesis is accepted even though it is wrong, i.e. that $P_\omega \in \mathcal{A}$ and $\phi(x) = 0$. This error is called *error of the second kind* or *Type II error* [98].

In the design of a statistical test, the test statistic is usually chosen such that the probability of a Type I error is less than a certain value (*level of significance*) α. In science, customary choices are 0.05, 0.01 or 0.001 for the level of significance, although the choice is arbitrary.

Additionally, how strong a sample x contradicts the Null hypothesis can be of interests, that is the smallest significance level $\alpha(X)$ for which $\mathcal{N}$ is still rejected [98]. This value is called the *p-value* of the sample x. If the p-value of a sample is less than the level of significance we say that the test for this sample or the sample is *significant*. The p-value for a sample can be obtained by integrating the probability of all test statistics that have a more extreme value than the observed statistic of the sample.

A common problem, especially in the setting of nonparametric statistics, is however, that the distribution of a statistic under the Null hypothesis is very hard to derive or sometimes even unknown. Therefore, computing the p-values for even simple statistics, such as for the identity of two distributions can be challenging, as it can involve integrating over exponentially many examples. In those cases *resampling* strategies have proven themselves useful. The idea of these resampling strategies is that the observed samples (*empirical distribution*) are used to approximate the underlying Null distribution for computing the p-value. This is commonly done by subsampling, with or without replacing, from the observed samples and then computing test statistics from these subsamples. From these a p-value can be estimated by counting the fraction of subsampled statistics that have a more extreme value than the observed statistics. Popular resampling strategies are *jackknife* [174] and *bootstrapping* [44]. An advantage of these methods is that they tend to give accurate estimations of the p-value. This, however, comes at the expense of high computational cost as typically the number of iterations needs to be large in order to approximate the empirical Null distribution and thus being able to compute small p-values.

We now discuss an issue that is important when multiple tests are performed, e.g. testing whether the expression of all genes is the same between two conditions. In this case, one has to be careful when interpreting the results as the fraction of significant tests in all tests is, by definition of the level of significance, expected to be at least as high as the significant level α of the employed test. This can be problematic if the number of true alternatives is not very high because the fraction of true positives in the set of all tests that for which the Null hypothesis has been rejected is low. Therefore, the positives are not representative for the true positives. This situation often occurs in the analysis of HTS data where the number of measurements is very high, e.g. for analyses that try to correlate genetic variants with gene expression changes (*association studies*). In association studies for example, a test is performed for each variant and the expression of each gene, leading to billions of tests. As often the number of true causal variants is small, this means that they are hard to detect in many associations that are significant only by chance and have no underlying biological cause.

A solution for this problem is to introduce alternative significance measures that account for the large number of tests that are performed (*multiple testing*). One of these alternative significance measure are the *familywise error rates* (FWER) that are determining the probability of making one or more Type I errors in the family of test. Examples of such controls are the *Bonferroni correction* [25] or *Holm's step-down procedure* [71]. These methods are in general very conservative. A less conservative significance measure is the *false discovery rate* (FDR) [165]. This is the expected fraction of false discoveries in all tests, i.e. the expected fraction of cases where the Null hypothesis was wrongly rejected. The FDR is the significance measure that is most commonly used in cases where the number of test is large, such as in association studies.

In order to quantify how well a statistical test can decide whether the Null hypothesis is true or not the *power* of a test can be computed. The power of a test is defined to be the probability of correctly rejecting the Null hypothesis for a given significance level α. This measure can also be used to compare how well tests can discriminate between the Null and

the Alternative hypothesis. A test that has a bigger power for a given α than all other tests is called a *uniformly most powerful* (UMP) test for the significance level α.

An example of a UMP test is the *likelihood ratio test* that can be applied in cases where there are only two alternative distributions, P_0 for the Null hypothesis and P_1 for the Alternative hypothesis. This test is constructed by comparing the quotient of the likelihood $P(x \,|\, \mathcal{N})$ of P_0 and the likelihood $P(x \,|\, \mathcal{A})$ of P_1 given the data x using the so called likelihood ratio $r(x)$:

$$r(x) := \begin{cases} \infty & \text{if } P(x \,|\, \mathcal{N}) = 0 \\ \frac{P(x \,|\, \mathcal{A})}{P(x \,|\, \mathcal{N})} & \text{otherwise} \end{cases}$$

The likelihood-ratio measures how much more likely it is for an observation x to come from the alternative distribution than from the Null distribution. Based on this test statistic the test ϕ for significance level α can then by constructed in the following manner [127]:

$$\phi(x) := \begin{cases} 0 & \text{if } r(x) \leq c \\ 1 & \text{if } r(x) > c \end{cases},$$

where c is chosen such that $P(r(x) < c \,|\, \mathcal{N}) = \alpha$.

2.5.4. Homogeneity Tests

A fundamental question that often arises is whether two distributions are identical. To answer this question based on samples from these two distributions only, statistical homogeneity tests can be applied. If the class of the underlying distributions is known, (e.g. to be Gaussian), tests like Welch's t-test can be applied. For this setting it is often straight forward to construct a powerful test. However, in many applications the class of distributions is unknown. For these cases nonparametric homogeneity tests can be applied. The most prominent of these tests is the two-sample *Kolmogorov-Smirnov* (K-S) test [87]. This test can be applied to test for identity of two continuous univariate probability distributions. It determines based on the maximal distance between the two empirical *cumulative distribution functions* whether the distributions are identical. More formally, if P and Q are two distributions and C_P and C_Q are their respective empirical cumulative distribution functions, then the test statistic Δ of the K-S-test is given by:

$$\Delta(P, Q) := \sup_{x \in \mathbb{R}} |C_P(x) - C_Q(x)|$$

The test statistic Δ measures the L^∞-norm, that is the maximal distance between the two empirical cumulative distribution functions. For two samples from the same distribution this test statistic converges to zero almost surely as the sample size increases, whereas for samples from different distributions the test statistics converges to a positive value almost surely. It has been shown that the distribution of Δ is asymptotically given by the Kolmogorov-Smirnov distribution, which can be used to compute p-values [87].

A shortcoming of the K-S test is that it can only be applied in the case when the probability distributions are one dimensional. An alternative to the K-S test for testing the identity of probability distributions that can also be applied to multivariate distributions is the *Maximum Mean Discrepancy* (MMD) test [26].

This test exploits the fact that differences of the expected values of a random variable on two distributions indicate differences of the distributions. Specifically, for two distributions P and Q on a metric space $\mathcal{X}$ and a function $f : \mathcal{X} \to \mathbb{R}$ with $\|f\|_2 < \infty$, the difference between the expected values of the function on P and Q

$$d_f(P, Q) := |\mathbf{E}_{x \sim P}[f(x)] - \mathbf{E}_{y \sim Q}[f(y)]|$$

induces a pseudometric on the set of probability measures on $\mathcal{X}$, that is a metric except that $d(P,Q) = 0$ does not imply $P = Q$.

However, this pseudometric depends strongly on the choice of the function f: For every two distributions P and Q that are not identical there exists a function f with $\|f\|_2 < \infty$, such that $d_f(P,Q) > 0$. On the other hand, for every given function f there do exist two distributions that are not the same but yield the same expected value for that function. Therefore, the pseudometric $d_f(\cdot,\cdot)$ is only of limited use as a test statistic for a statistical test.

An extension of this pseudometric is the maximum mean discrepancy. For two distributions P and Q and a set of functions $\mathcal{F}$ the MMD is defined by:

$$\mathrm{MMD}(\mathcal{F},P,Q) := \sup_{f\in\mathcal{F}} |\mathbf{E}_{x\sim P}[f(x)] - \mathbf{E}_{y\sim Q}[f(y)]|$$

As shown in [62], a biased empirical estimate $\widehat{\mathrm{MMD}}_b$ of the MMD for two finite samples $X = \{x_1,\ldots,x_n\}$ drawn from P and $Y = \{y_1,\ldots,y_m\}$ drawn from Q is given by:

$$\widehat{\mathrm{MMD}}_b(\mathcal{F},P,Q) := \sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} f(x_i) - \frac{1}{m}\sum_{j=1}^{m} f(y_j) \right|$$

It should be noted that the MMD critically depends on the richness of $\mathcal{F}$. It has been shown, for example, in [43] that if $\mathcal{F}$ is the set of all continuous functions on $\mathcal{X}$ then P and Q are identical if and only if $\mathrm{MMD}(\mathcal{F},P,Q) = 0$. A stronger categorisation has shown that in a compact metric space $\mathcal{X}$, it suffices for $\mathcal{F}$ to be a set of functions that is dense in the set of bounded continuous functions on $\mathcal{X}$ with respect to the L^∞-norm, for $\mathrm{MMD}(\mathcal{F},\cdot,\cdot)$ to be a metric [62]. For the empirical estimate $\widehat{\mathrm{MMD}}$, however, the class of all real valued functions is too large, as the MMD for all non-identical samples is non-zero [26]. This shows that for construction of a test the class of functions $\mathcal{F}$ must be rich enough in order to detect all differences but must not be too rich as otherwise $\widehat{\mathrm{MMD}}$ overestimated MMD.

A class of functions $\mathcal{F}$ that has this property can be constructed as follows: Let $\mathcal{H}$ be the complete inner product space (*Hilbert space*) of the set of functions $\{f \mid f : \mathcal{X} \to \mathbb{R}\}$ on a compact metric space $\mathcal{X}$. Then if the point evaluation $f \mapsto f(x)$ for a point $x \in \mathcal{X}$ is a linear continuous function then $\mathcal{H}$ is called a *reproducing Kernel Hilbert Space* (RKHS). In this case, by the Riesz's representation theorem (see e.g. [145]), there exists a function, termed *feature map*, $\phi : f : \mathcal{X} \to \mathbb{R}$ such that for each function $f \in \mathcal{H}$, $f(x) =< \phi(x), f >_{\mathcal{H}}$, where $< \cdot,\cdot >_{\mathcal{H}}$ is the *inner product* of $\mathcal{H}$. The feature map, then induces the *kernel* $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ by $k(x,y) =< \phi(x), \phi(y) >_{\mathcal{H}}$. If $k(x,\cdot)$ is continuous and $\mathcal{H}$ is dense in the set of bounded continuous functions on $\mathcal{X}$ with respect to the L^∞-norm, $\mathcal{H}$ is called a *universal* RKHS. If this is the case, then the unit ball $\mathcal{F} \subset \mathcal{H}$ has the property [62]:

$$\mathrm{MMD}(\mathcal{F},P,Q) = 0 \Leftrightarrow P = Q$$

Moreover, if $\mathcal{F}$ is the unit ball of a RKHS $\mathcal{H}$ with a kernel k then an unbiased estimator $\widehat{\mathrm{MMD}}$ of $\mathrm{MMD}(\mathcal{F},P,Q)$ can be obtained by:

$$\widehat{\mathrm{MMD}}^2(\mathcal{F},P,Q) = \frac{1}{n(n-1)}\sum_{i\neq j}^{n} k(x_i,x_j) + \frac{1}{m(m-1)}\sum_{i\neq j}^{m} k(y_i,y_j) - \frac{2}{nm}\sum_{i,j=1}^{m,n} k(x_i,y_j)$$

With this test statistic a p-value can then be computed using a bootstrapping approach. This can be done by repeatedly sampling without replacement, sets of size n and m from

the combined samples from both distributions $\{x_1, \ldots, x_n, y_1, \ldots, y_m\}$. Under the Null hypothesis ($P = Q$), this yields two samples from the Null distribution of the maximum mean discrepancy. Therefore, the empirical Null distribution of the MMD can be estimated by repeatedly sampling two samples from the Null distribution and computing their MMD. From the empirical Null distribution the p-value can then be obtained as explained before. A recent development showed that when the sample size is large an analytic approximation of the maximum mean discrepancy Null distribution exists [62], which allows a substantial speed-up in computation.

2.6. Machine Learning

In recent years, *machine learning* has become an important field in data analysis. It is a field that is closely related to statistics but with a slightly different focus. In statistics, the focus lies on modelling the underlying processes to then perform inference from observations. According to Murphy [123], in contrast, the aim of machine learning is to develop methods to automatically *learn* structures in data and use the uncovered pattern to predict future data or other outcomes of interest.

In cases where complex systems with many unknown factors are studied, focusing on *generalisation* of data rather than on understanding the data generation process typically leads to better prediction for unseen data and identification of the important factors. When studying biological systems for example, this allows better identification of important regulators.

In machine learning there are several distinct settings [20]. Two important of these that are the *supervised learning* and the *unsupervised learning* setting. In *supervised learning* the goal is to learn from a labelled set of observations the labels for observations where the label is unknown. If the set of labels is finite, supervised learning is called *classification*. If the set of labels is infinite, supervised learning is typically called *regression*. In *unsupervised learning* the aim is to uncover the structure of observations. In the case when the aim is to discover groups of similar observations in all observations, unsupervised learning is called *clustering*. In the case when the aim is to estimate the distribution that generates the data, unsupervised learning is called *density estimation*.

In this work, we focus on the supervised learning setting. For an overview of unsupervised and other machine learning settings we refer to [20, 67, 124]. In the following, we first introduce general principles of supervised learning and then discuss methods for regression.

2.6.1. General Principles of Supervised Machine Learning

In the supervised learning setting typically a set of points $(x_1, \ldots, x_n) \in \mathcal{X}$ from a space $\mathcal{X}$ (*feature space*) together with their respective labels $(y_1, \ldots, y_n) \in \mathcal{Y}$ from the space of labels $\mathcal{Y}$ are given. The objective is then to learn a function (*predictor*) $f : \mathcal{X} \to \mathcal{Y}$ that predicts for a data point $x \in \mathcal{X}$ its label $y \in \mathcal{Y}$. For the remainder of this chapter we assume that the labels for classification are $\mathcal{Y} = \{0, \ldots, N\}$ and for regression $\mathcal{Y} = \mathbb{R}$.

Intuitively, a predictor f is a good predictor if its predicted label $f(x)$ for a point x is close to the true label y. This can be formalised by first introducing of *loss function* $L(f(x), y)$ that assigns a cost to the discrepancy between the predicted label $f(x)$ and the true label y. This loss allows judging the quality of a predictor for a given data point. The loss function is usually chosen such that the cost is a monotone increasing function of the deviation

discrepancy between the prediction and the ground truth. An example for a loss function that is commonly used for classification is the $0-1$ loss $\mathbf{I}_{f(x)=y}$. For regression a common losses are the squared loss $\|f(x)-y\|_2^2$ or the more general L^d losses $\|f(x)-y\|_d^d, d \in [0,\infty]$ that are based on the L^d norms $\|\cdot\|_d$. Another common loss function that can be used when there are many outliers is the epsilon insensitive loss $\max(0, \|f(x)-y\|_1 - \epsilon)$. For a given loss function the overall quality can then be measured by the expected loss of f, termed the *risk* $\mathcal{R}(f)$. This risk $\mathcal{R}(f)$ of f can be defined as follows:

$$\mathcal{R}(f) := \int_{\mathcal{X}\times\mathcal{Y}} L(f(x), y) p(x, y) \,\mathrm{d}x \,\mathrm{d}y,$$

where $p(x, y)$ is the probability of seeing x with label y. An optimal predictor f^* from a set of potential predictors $\mathcal{F}$ is then characterised by

$$f^* = \arg\min_{f\in\mathcal{F}} \mathcal{R}(f)$$

The predictor f^* then has the property that there exists no other predictor that has a smaller risk. In practical applications, however, the risk of a predictor can often not be computed as the distribution $p(x, y)$ is usually unknown and, therefore, the optimal predictor cannot be determined. In this case, where $p(x, y)$ is unknown, the risk of the predictor f on the observations $\mathcal{R}_n(f)$ (*empirical risk*) can be used as a substitute. This empirical risk can be defined as follows:

$$\mathcal{R}_n(f) := \frac{1}{n}\sum_{i=1}^{n} L(f(x_i), y_i).$$

For the empirical risk the minimiser f_n^* is then given by $f_n^* := \arg\min_{f\in\mathcal{F}} \mathcal{R}_n(f)$. Determining the predictor f_n^* is commonly performed using optimisation methods.

It is important to note, however, that the predictor f_n^* with the smallest empirical risk $\mathcal{R}_n$ does not necessarily need to also have a minimal risk $\mathcal{R}(f_n^*)$. This is especially important to consider for a function class $\mathcal{F}$ that is rich enough such that it is always possible to find a predictor that would perfectly fit the data, i.e. that has empirical risk 0. This is because such a predictor would fit on nosy data the noise (*overfitting*) and would therefore have, in general, a suboptimal performance on previously unseen data [27].

To prevent overfitting, different strategies have been proposed. These strategies all have in common that they aim to find a simple function in order to explain the data (for a detailed motivation of this approach we refer to [27]). The first strategy is to choose a reasonably small class of functions $\mathcal{F}$ that is still big enough to approximate the minimiser of the risk. This strategy is known as *empirical risk minimisation* [177, 178]. Another strategy is *regularisation*. Here, the idea is to minimise the *regularised empirical risk* $\mathcal{R}_n^r$ instead of the risk $\mathcal{R}$, which can be defined in the following way:

$$\mathcal{R}_n^r(f) := \mathcal{R}_n(f) + \lambda\Omega(f),$$

where $\lambda > 0$ is the strength of the regularisation and $\Omega(f)$ is a *regulariser* of f that gives a high value to complex functions, e.g. the norm of f or of the parameters of f. For regularisation, the class of functions $\mathcal{F}$ can be larger (e.g. the class of continuous functions) than for the empirical risk minimisation as the complexity of the function is implicitly reduced by the regularisation [27]. The parameter λ determines the degree of regularisation of the solution. Determining the optimal parameter λ can be performed using *cross validation*. In cross validation the best choice of λ from a set of candidates $(\lambda_i)_i$ is determined in the following manner: First, the data are randomly split up in two sets, the *training set* and the *test set*.

Next, for each of a set of candidates $(\lambda_i)_i$ the best predictor is determined based on the data in the training set and its performance is subsequently assessed on the test set. This procedure is repeated for different splits of the data into training and test sets and the aggregated performance over all splits is finally used to determine the best λ_i.

In the next section, we discuss how regularisation can be applied for regression.

2.6.2. Linear Models for Regression

In the regression setting the aim is to learn a function that can predict the labels of unseen data, i.e. has minimal risk. In the following, we assume that the observations $x \in \mathbb{R}^N$ are real valued N-dimensional vectors and that their labels $y \in \mathbb{R}$ are real valued scalars. In this situation, learning a predictor can be achieved by minimising the regularised empirical risk $\mathcal{R}^r_n(f)$. In the following, we exemplify how this can be done for one of the best established cases of regression, the *linear models*.

In case of linear models, the set of potential predictors $\mathcal{F}$ is the set of functions that are linear in their parameters, i.e. every function $f \in \mathcal{F}$ is given by:

$$f(x) = \alpha_0 + \sum_{i=1}^{m} \alpha_i \phi_i(x),$$

where $\phi_i, i \in \{1, \dots, m\}$ are fixed functions of x (*basis functions*) and $\alpha_0 \dots \alpha_m \in \mathbb{R}$ are the parameters of f. It should be noted that the functions f in linear models are in general not linear; they are only linear when the all the basis functions ϕ_i are linear in x. An advantage of linear models is that they include many commonly known regression problems and are widely applicable. If, for example, ϕ_i are the functions that map a vector x to its i-th coordinate $x_{(i)}$, then the linear model reduces to *linear regression*. In the case when ϕ_i are the functions that map a vector x to its i-th coordinate $x_{(i)}$, the resulting functions f are polynomial and thus linear models can be used for *polynomial regression*. Linear models can be used with different loss functions and regularisers. In the following, we present three of the most commonly used combinations of loss functions and regularisers: The *least square regression*, *ridge regression* and *lasso regression*. To simplify the notation, we define $\phi_0 = 1$ and we introduce a vector notation where $\mathbf{Y} = (y_1, \dots, y_n)$ is the vector of labels, $\mathbf{\Phi} = (\phi_i(x_j))_{j,i,i\in\{0,\dots,m\},j\in\{1,\dots,n\}}$ is called the *design matrix* and $\mathbf{A} = (\alpha_0, \dots \alpha_m)$ is the vector of parameters.

Least square regression The case when L is the squared loss and $\lambda = 0$ (i.e. there is no regularisation), is called least square regression. In this case the minimiser of regularised empirical risk can be obtained by solving the following optimisation problem for the parameter $\mathbf{A} \in \mathbb{R}^{m+1}$:

$$\mathbf{A}^* = \arg\min_{\mathbf{A}\in\mathbb{R}^{m+1}} \|\mathbf{Y} - \mathbf{\Phi}\mathbf{A}\|_2^2$$

For this optimisation problem an analytic solution exists:

$$\mathbf{A}^* = (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{Y},$$

where $(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T$ is the *Moore-Penrose pseudoinverse* of the design matrix $\mathbf{\Phi}$ [20]. It is worthwhile noting that this solution is also the maximum likelihood solution for $\mathbf{A}^*$ under the assumption that there is a Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$ on the observations, i.e. that $y = f(x) + \epsilon$ [20].

Ridge regression The case where L is the squared loss, $\lambda > 0$ and the regulariser is given by the L^2 norm of the parameter vector $\mathbf{A}$ is called ridge regression [170]. In this case, the

optimal parameter for the regularised regression $\mathbf{A}^* \in \mathbb{R}^{m+1}$ can be obtained by solving the following optimisation problem :

$$\mathbf{A}^* = \arg\min_{\mathbf{A}\in\mathbb{R}^{m+1}} \|\mathbf{Y} - \mathbf{\Phi}\mathbf{A}\|_2^2 + \lambda\|\mathbf{A}\|_2^2$$

For this there also exists a closed form solution. This closed form solution is given by:

$$\mathbf{A}^* = (\lambda\mathbf{I} + \mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{Y},$$

where $\mathbf{I}$ is the identity matrix [20]. This solution can be interpreted as the maximum likelihood solution of $\mathbf{A}^*$ when there is a Gaussian noise on the observations and additionally a Gaussian prior $\mathcal{N}(0, \frac{1}{\lambda}\mathbf{I})$ on the parameters $\mathbf{A}$ [20]. The ridge regression is an example of an estimator that minimises the parameters that are not important for the regression (*shrinkage*), as the regulariser forces the parameters to become small if there is no evidence in the data that they are relevant for prediction.

Lasso regression The case where L is the quadratic loss, $\lambda > 0$ and the regulariser is given by the L^1 norm of the parameter vector $\mathbf{A}$ is called the lasso regression [169]. Here, the optimal parameter of the lasso regression $\mathbf{A}^*$ can be obtained by minimizing:

$$\|\mathbf{Y} - \mathbf{\Phi}\mathbf{A}\|_2^2 + \lambda\|\mathbf{A}\|_1$$

In contrast to the previous two methods however, there does not exist a general closed form solution for this problem. To solve this optimisation problem nonetheless, algorithms such as *least angle regression* [45] have been proposed. Compared to the L^2 regularisation, the L^1 regularisation of the lasso regression leads to a sparser estimation of $\mathbf{A}$, meaning that more of the parameters α_i are zero. In cases where the true model is sparse and there is little data, lasso regression tends to perform better than ridge regression. The lasso regression also has a probabilistic interpretation, namely that it is the maximum likelihood solution of $\mathbf{A}^*$ when there is a Laplace distribution prior on the parameters and Gaussian noise on the observations [67].

Besides the three cases that have been presented above, linear models allow for various other choices of loss functions, regularisations and basis functions. For an overview of established alternative choices we refer to [20, 67, 124, 159].

2.6.3. Mixed Models

Regression models have been successfully applied in genetics where it is of interest to model the effect of genetic variants on a phenotype in a group of individuals. This idea has first been proposed in 1919 by Fisher [48] and since then has been further developed. From these developments, a special class of regression models, the mixed models, have emerged as method of choice [190].

One reason for this is that they allow modelling the effects of genetic variants and also unobservable factors on the phenotype, e.g. unwanted batch effects or different environmental conditions. These models also allow accounting for the correlation of phenotypes that is caused by individuals being related to some degree (*population structure*), which reduces the false positive rate [190].

A mixed model consists of three components [77]. The first component models the effects that are assumed to be non-random (*fixed effects*). The second component models the effects that are assumed to be random (*random effects*). This component can be used to model the

correlation structure between the traits of the individuals, e.g. to model population structure or batch effects. The last component is a noise model, where the noise is not correlated between individuals. Formally, the mixed model for a quantitative phenotype $\mathbf{Y}$ can be stated as:

$$\mathbf{Y} = \mathbf{X}\alpha + \mathbf{Z}\beta + \epsilon,$$

where $\mathbf{X}$ is the design matrix for the fixed effects, α is a vector with unknown regression coefficients, $\mathbf{Z}$ is the design matrix of random effect, β is a vector of random effects and ϵ is a vector that models the noise [77].

In order to determine the regression coefficients α and β the model can be fitted to the data. This can be done, for example, using algorithms that iteratively fit the different components of the data such as the EM algorithm [102].

A special case of mixed models is the Gaussian mixed model. In this case the assumption is that both, the random effects and the noise are normally distributed, with $\beta \sim \mathcal{N}(0, K)$ and with $\epsilon \sim \mathcal{N}(0, D)$, where K is the covariance matrix of the random effects and D is a diagonal covariance matrix of the noise. An advantage of the Gaussian mixed models is that modelling the random components as a Gaussian distribution provides confidence intervals for the predictions. For fitting the Gaussian mixed models, efficient strategies have been proposed (see, e.g. [103]).

3. Experimental Design for RNA-Seq experiments

3.1. Motivation

RNA-Seq experiments have non-negligible costs. Therefore, to optimally invest the resources for sequencing, it is advisable to clearly define the objective of the investigation and design the experiment accordingly. Several factors have emerged to be important to be taken into account when designing experiments. Some of them affect the overall layout of the experiment (*design layout*) and others the sequencing itself (*sequencing parameters*). In the following, we first introduce several factors related to the design layout and then continue by discussing the most important sequencing parameters.

An important question that affects the overall experimental design is whether the purpose of the transcriptome sequencing is to obtain qualitative or quantitative information. The answer to this question determines which of the different aspects are to be considered and, consequently, different design layouts apply. When only qualitative information is needed, replicated independent measurements (replicates) are not necessary in general. In contrast, when information of quantitative nature is needed (e.g., to quantify gene expression), stable estimates are important; these typically require replicates. These replicates can be created from the same RNA sample (*technical replicates*) or from different samples (*biological replicates*). Replicates also provide the means to estimate the variance of the quantifications, thus providing the basis for further robust statistical analysis.

There are many known and unknown biases during sequencing that systematically alter the measurements [3, 121, 143]. These typically affect the quantitative measurements to a far larger extent than qualitative ones. Some authors (e.g. [9]) suggest categorizing these biases into two classes, those that are induced between the random fragmentation of the RNA and its insertion into the flow cell (*batch effects*) and the others that arise after this (*lane effects* [110]). Factors that can induce these distortions are different experimental conditions during sequencing, differences in the chemicals that are used, barcodes for multiplexing, but also lane differences of the sequencers itself [65, 121, 130]. If a comparison between multiple samples is intended, it is important to design the experiment so that the effect of these biases on the measurements is minimised and does not in turn bias the comparison. This can be achieved by preparing the samples in parallel and by using the same reagents. An additional option is to use a *balanced block design* [9, 142]. In this design the samples are prepared such that the samples are distributed equally across different batches and lanes to avoid systematically different sequencing for different groups of samples. In cases, where the number of samples is small, this can be done by splitting up libraries followed by multiplexing and pooling libraries from different samples [9]. Finally, it is advisable to quantify the remaining distortion and estimate the variance induced by these effects using the information from the dispersion of the replicated measurements. As an experimental means to estimate the distortions that are induced in the sequencing, spike-ins have been proposed [78]. These are synthetic RNAs that can be added at a known concentration to the RNA samples. The differences in abundance of reads that are generated from these RNAs then allows inferring the different distortions that apply to the samples.

Besides the aforementioned design layout of the study, also the choices of sequencing parameters in the library preparation and sequencing protocol are important and should be considered when designing the experiment. This aspect involves identifying the fraction of the transcriptome that is intended to be sequenced and determining how to sequence it in order to maximise the information yield for the question of interest. The former part typically invokes establishing filtering steps and choosing the appropriate sequencing protocol. If mRNAs should be sequenced, it is advised to select RNA molecules with a poly(A)-tail and having an insert size between 300 and 500 bp (pers. comm. Lisa Smith). For sequencing miRNA for example it is suggested to filter for fragments that have a length similar to the length of miRNAs [122]. Besides these examples there are numerous other selections and filters that were proposed (e.g. selection for 5' or 3' ends) which will not be discussed here.

The role of sequencing parameters in the experimental design has been studied relatively little; to our knowledge, a systematic conclusive analysis of the effect of parameters such as the insert site, read type and library size does not exist. Anecdotal reports on these effects have been made in [13, 83]. In the former study ([83]) it was shown that paired-end read information and a small insert size variability facilitates the assignments of reads to transcripts. This was done for selected representative gene models. However, a genome-wide analysis was not performed yet. In the latter study ([13]) the effect of the insert size on the detection of structural variation was investigated. The authors also performed a calculation of the minimal number of reads to be sequenced (*sequencing depth*) to identify a certain fraction of expressed transcripts (transcript identification). However, the authors assumed unrealistically that there is only one transcript per gene and did not account for the effect of the read type. Therefore, their results on transcript identification are not conclusive.

In this chapter we will derive a probabilistic framework to model the utility of RNA-Seq experiments for the task of the identification of expressed transcripts (transcript identification)when the gene annotation is known, which is an important task in order to determine the cellular state and is fundamental for understanding the regulation of RNA processing We will, therefore, analyse the combined effect of various parameters of RNA sequencing experiments on the information gain of this analysis.

We will also discuss how this framework can be adapted in order to optimise parameter choices for other tasks than transcript identification. In particular, we will adapt the framework to model differential splicing and present results on how detection of differential splicing is influenced by different parameter choices.

Finally, we will present experimental results that highlight the benefit of parameter optimisation, we will show limitations of the commonly used sequencing approach for transcript identification and we will discuss the detection limits of differential splicing.

3.2. Methods

3.2.1. Modelling of Transcript Identification

A key advantage of high-throughput sequencing is that it provides an unbiased view on the transcriptome. In particular, it allows identifying and quantifying the expressed transcripts, thus revealing the biological state of the sequenced cells. This can be achieved by first identifying the transcripts to which the sequenced reads map and then counting these reads.

Uniquely mapping a sequenced read to the corresponding transcript has varying complexity

depending on the number of isoforms of the gene from which the read originates: If a gene that does not overlap other genes has only one isoform, it is obvious that each read that maps to the respective transcript shows that the isoform was expressed. However, in the case of genes with multiple isoforms, several transcripts can stem from the same genomic locus and thus share stretches (*subsequences*) of the same sequence. As a consequence, even if a read maps uniquely to a gene, it can still map to several transcripts (i.e. the mapping can be *ambiguous*). In this case we can categorise the reads into three classes: (1) A read that maps to all the other transcripts in the gene indicates the expression of the gene it originates from, but cannot be used to draw conclusions about the transcript it stems from (*uninformative reads*). (2) One that maps only to a single transcript identifies the respective transcript (*unique reads*). (3) Finally, the reads that are not in any one of the previous categories are called *informative reads*. An informative read on its own cannot be used to identify a transcript but there are approaches that use all reads in order to infer the expressed transcripts (e.g. [14, 24, 173]). For simplicity we refer to the set of positions that these reads map to in the following, as *unique*, *informative* respectively *uninformative* positions. Depending on the context, we also refer to these positions as *regions*.

The ability to identify transcripts depends on the number of reads that can be mapped non-ambiguously. This number is influenced by several parameters of the read generation process, such as the type of reads (single-end or paired-end), the insert size and the length of the reads, as well as the on the sequencing depth.

When transcript identification is intended in an experiment then the sequencing parameters should be chosen that yield the highest number of identified transcripts. In order to determine the optimal parameters, we therefore propose to systematically investigate the effect of the sequencing parameters on transcript identification. For this, we formulate a probabilistic model of the effect of sequencing parameters on transcript identification. As only the unique reads provide direct evidence for the expression of a specific isoform, we focus on those in our model.

An assumption that facilitates the probabilistic modelling significantly is that the reads are sampled independently of each other. This assumption can be justified as in a typical RNA-Seq experiment the amount of RNA transcripts in the sample is orders of magnitudes larger than the number of reads that are sequenced. Therefore, it is unlikely that two reads stem from the same RNA molecule, which could introduce dependences between reads. The advantage of this assumption is that the probability of identification of a transcript is determined by the probability of identification by a single randomly chosen read. Therefore a strategy to model the probability that a transcript can be identified for given parameters is:

1. Establishing the necessary and sufficient property of a read to identify a transcript
2. Computing the probability that a randomly drawn read exhibits these properties
3. Deriving the probability that a transcript can be identified using all reads of a library

We formalise the problem as follows: Let g be a gene of length L at the genomic positions I that has the sequence $S = (s_i)_{i \in I}$. Assume furthermore that g has T isoforms $t_1, \ldots, t_T$ having length $L_1, \ldots, L_T$. By definition of a gene the sequence of each transcript $S_j = (s_i)_{i \in J}$ is a subsequence of S, that is, $J \subseteq I$. Then, if N_t is the number of copies of a transcript t in the sample under investigation, t is expressed if $N_t > 0$. For the remainder of this chapter we also assume that only expressed transcripts can generate reads and that all reads are correctly mapped. Finally, to simplify the notation, we omit the index for the gene when possible. We begin by introducing a compact representation of the set of isoforms of a gene (*gene structure*) to facilitate the notation of the model.

Gene Structure Representations

The most commonly used representation of the gene isoforms is the splice graph [68]. This is a directed graph $G = (V, E)$ that represents the exons of a gene and their junctions (see Fig. 3.1 for an example). Formally, this graph is defined as follows. If $(e^i_1, \ldots, e^i_{E_i})$ is the ordered list of all E_i exons e^t_j of transcript t, then the set of splice graph vertices V is given by:

$$V := \bigcup_{t=1}^{T} \bigcup_{j=1}^{E_t} e^t_j$$

and the set of its edges E by:

$$E := \bigcup_{t=1}^{T} \bigcup_{j=1}^{E_t-1} (e^t_j, e^t_{j+1}).$$

In this graph a transcript t corresponds to a path from its first exon e^t_1 to its last exon $e^t_{E_t}$. The splice graph has the advantage that it provides a compact representation of the exons and their junctions. It is therefore commonly used to represent the gene structure. However, a limitation of this representation is that the transcripts cannot be reconstructed from it as the long distance dependencies between the exons are not contained in this representation.

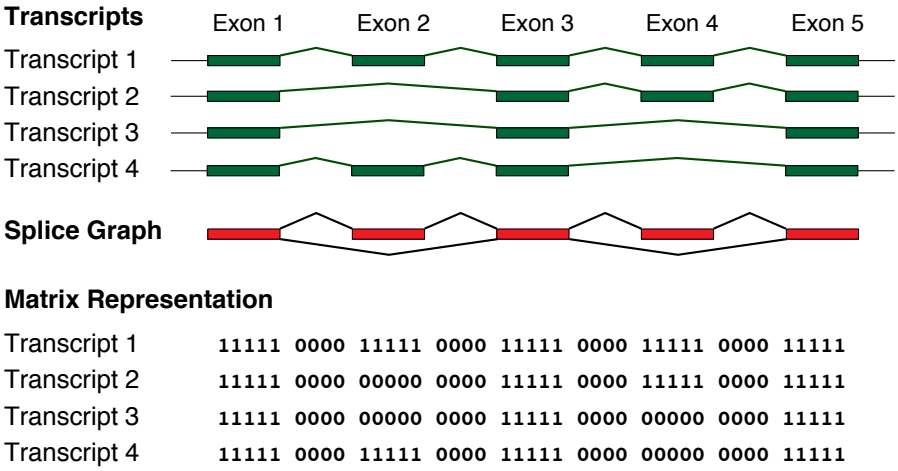

Figure 3.1.: Illustrated are two representations of a gene structure. Shown on top is a gene with four transcripts and five exons (green) that are spliced together in different combinations. Exons 1, 3 and 5 are constitutive exons whereas exons 2 and 4 are alternatively spliced. Shown in the middle is the splice graph representation (red), where exons constitute nodes and paths between them represent their junctions. Shown below is the matrix representation. In this representations positions that are included in a transcript are represented by a one and the others by zero in the row of the matrix that corresponds to the respective transcript.

For transcript identification the splice graph is therefore not rich enough. This drawback can be resolved by the use of a *matrix representation* of the gene structure (see for example [24]). For a gene g this representation is given by a $\{0,1\}$-valued matrix $A_g \in \{0,1\}^{L \times T}$, where the entry $a^g_{i,t}$ of A_g is 1 if and only if transcript t contains position i and 0 otherwise (see Fig. 3.1 for an example). This representation contains all the positional information of the transcripts and reconstruction of transcripts from it is trivial. We furthermore suggest using

the same matrix representation for the reads. A read r is represented in the same manner as the transcript by a matrix $A_r \in \{0,1\}^{L\times 1}$ with non-zero entries a_i^r at the positions where it maps to the genome.

Single-end Reads

We begin by establishing the necessary and sufficient property for a single-end read to identify a transcript. We then derive the probability that a randomly drawn read has these properties. For single-end reads we have two experimental parameters. The first one is the read length and the second one is the number of reads to be sequenced. In the following we assume that all reads have the same length l.

In order to identify a transcript t, a read r firstly has to stem from the transcript and secondly must not map to any other transcript than t. This can be checked, using the matrix representations, in the following way. Let i^r_{start} and i^r_{stop} be the first and the last genomic position, respectively, where r maps to. Then r can originate from t, if and only if the representations of the read and the transcript agree in all position, between the start and the stop of the read:

$$a^g_{i,t} = a^r_i, \quad \forall i \in \mathbb{N}, i^r_{start} \leq i \leq i^r_{stop}$$

or equivalently if:

$$\min_{i^r_{start}\leq i\leq i^r_{stop}} \delta_{a^g_{i,t}=a^r_i} = 1, \tag{3.1}$$

where δ_x is the identity function. This function is one if x is true and zero otherwise. We say that a single-end read agrees with a transcript if criterion 3.1 is full filed. Furthermore, we introduce the notation $a^s(t,r)$ to denote this property for single-end reads and define it by:

$$a^s(t,r) := \min_{i^r_{start}\leq i\leq i^r_{stop}} \delta_{a^g_{i,t}=a^r_i}.$$

Therefore, $a^s(t,r)$ equals one if the read r agrees with t and otherwise zero. Using this, we can determine whether a read can identify a transcript. As this is the case if and only if t is the only transcript that r maps to we have that

$$a^s(t,r)\sum_{t'=1}^{T} a^s(t',r) \quad = \quad 1 \tag{3.2}$$

if and only if r identifies t. This can be seen, as the left hand side of Eq.3.2 is zero if r does not stem from t and bigger than one if it maps to another transcript beside t. To furthermore simplify the notation, we denote the property of a single-end read r to identify t by $I^s(t,r)$. We define $I^s(t,r)$ to be 1 if the read r identifies t and 0 otherwise.

We then use this notation to derive the probability that a randomly drawn read r can identify the transcript t. The probability of this depends on two factors. First, the probability $p(r)$ that r is drawn and secondly, whether r can identify t. Therefore, the probability that t can be identified is given by:

$$P(t \text{ is identified}) = \sum_{r\in\mathcal{R}} I^s(t,r)p(r), \tag{3.3}$$

where $\mathcal{R}$ is the set of all possible reads that can be generated from t. If we assume that we have an uninformative distribution where all reads are equally likely, the probability 3.3

can explicitly be computed. In this case the probability of read starts is uniform for all positions but the last $l-1$ where it is 0 (reads of length l that start at these last positions would exceed the transcript and thus cannot exist). The set of reads $\mathcal{R}$ is therefore given by $\mathcal{R} = \{r_1, \ldots, r_{L_t-l+1}\}$, where r_i denotes the read that starts at position i of t. The probability is hence given by:

$$P(t \text{ is identified}) = \frac{\sum_{i=1}^{L_t-l+1} I^s(t, r_i)}{L_t - l + 1} \tag{3.4}$$

This formula shows that the probability of identification depends on the fraction of unique reads that identify t. The unique regions to which these reads map are typically alternatively spliced parts or exon junctions. However, these regions may not always exist and their existence depends on the read length. For example in Fig. 3.1, the transcripts only can be identified when the read length is greater than the exon lengths. Only then reads can exist that can identify the two alternative splice events that are separated by the middle exon.

The computation of the probability of identification 3.4 can be further simplified. For transcript identification the exact position of the reads is often not necessary and it suffices to consider with which combination of exons a read overlaps, i.e. it does not matter where in an exon a read maps in order to identify it. Therefore, all reads that map to the same set of exons provide the same information for transcript identification. We use this observation to motivate a *reduced matrix representation*. In this representation the columns do not represent positions but combinations of exons. This representation can be derived in the following way:

In a first step, we group together all positions between two splice sites of any transcript, thereby, defining K regions $R = \{R_1, \ldots, R_K\}$, where $K+1$ is the number of splice sites in the gene. Here, we use a broad definition of splice site, meaning any genomic locus where an exon starts or ends. These newly defined regions have the property that any two positions in a region have the same matrix representation and that there is no splice site in a region. We then determine all combinations c^r of regions from R that can be covered by a single read r. It should be noted that this definition also applies for paired-end reads. We furthermore define for each of the combinations c_r the length l_{c^r} of c_r as the number of distinct reads that map to this combination. Finally, we define $C^l = \{c_1, \ldots, c_{n^l}\}$ to be the set of all these combinations of regions for reads of length l, where n^l is typically much smaller than the length of the gene L. A region of C^l is characterised by its property that all reads that map to it have the same capability of identifying transcripts, i.e. either all reads that map to it identify a transcript or all reads cannot identify it.

Using the regions from C^l we can define the reduced matrix representation. This representation is given by a matrix $H \in \{0,1\}^{n^l \times T}$, where its entries $h_{i,j}$ are one if reads from transcript j map to the region combination c_i and zero otherwise. We use the same notation $a^s(t, r)$ as for the matrix representation to denote that r can stem from t. However, instead of agreeing in all positions, then a read has to agree in all regions $c^r \in C^l$ between its start and stop. By using this definition, the other notations can be defined analogously as before.

We can then use the reduced matrix representation to express the probability 3.4 in a simplified way:

$$P(t \text{ is identified}) = \frac{\sum_{i=1}^{n^l} I^s(t, c_i) l_{c_i}}{L_t - l + 1}$$

Here, only summation over n^l terms instead of $|\mathcal{R}|$ is necessary and thus this provides a more efficient way of computing the probability. Moreover, this expression also provides a more

comprehensive characterisation of the probability and the effect of parameters. It can be seen that the size of the exon combinations that are unique for a transcript are the determinant of the probability of transcript identification and also how this depends on the read length l.

Paired-end Reads

Establishing the necessary and sufficient properties for paired-end reads to identify a transcript is, due to their complex structure, slightly different than for single-end reads. We begin by first defining the structure of paired-end reads first and then discuss the properties of them that allow identifying transcripts.

In the paired-end sequencing protocols the RNA fragments are sequenced from both ends, resulting in two short single-end reads of length l from the same transcript. Furthermore, since the fragments have been filtered for their length, also their expected lengths (*insert size*) L^i, are known. By this we can determine the distance between the two read-ends $L^i - 2l$. In the case where $L^i = 2l$ read-ends are adjacent and the paired-end read becomes a single-end read of length $2l$. The sequencing protocol also allows for reads to overlap: This is the case, when the insert size is less than $2l$ and the reads become a single end read of length smaller than $2l$. In the following, we denote the left and the right end of a paired-end read r with r^l respectively r^r.

Both ends of a paired-end read have on their own the same properties as single-end reads and thus can identify transcripts individually. However, beyond this, the information that they both stem from the same transcript allows also a joint identification by both ends. For example when one of the ends shows that the read cannot stem from a certain transcript, we can conclude that the other cannot as well, even though it alone does not provide the information for this conclusion. Hence, a paired-end read r agrees with a transcript t only when both ends agree, that is $a^s(t, r^l)a^s(t, r^r) = 1$. We denote by $I^p(t, r)$ the property of a paired-end read r to identify a transcript t. We, therefore, have that $I^p(t, r)$ is only zero if both $I^s(t, r^l)$ and $I^s(t, r^r)$ are zero as well.

Beyond the information contained in read-ends, paired-end reads also contain information on the distance between the ends. This information can also be exploited to identify transcripts. This can be illustrated in the example shown in Fig. 3.1. If the two read-ends map completely to exon 1 respectively in exon 3 and it is known that the insert size is smaller than one exon, then it can be concluded that the read stems from a transcript that does not contain exon 2. Otherwise, the read would have to be longer than one exon. This shows that information on the insert size can be used to exclude further potential transcripts of origin, thus increasing the power of identification. Formally this can be achieved in the following way: Let i_{min} and i_{max} denote the minimal and maximal insert size of the insert size distribution. Then a read r does not agree with a transcript t based on the insert size if the distance between the read start and read end on the transcript is not compatible with i_{min} and i_{max}. In the following, we denote this property with $a^i(t, r)$. This information is especially useful if the insert size variability is small because then, the two extreme values are close and allow excluding more potential transcripts. When the insert size distribution is large also more strict cutoffs can be chosen for i_{min} respectively i_{max}. In this case, one should be aware, however, that the false positive rate is not zero any more as true potential transcripts of origin can be rejected. We denote the property that a read r does not agree with t, based on one of the above mentioned properties, by $a^p(t, r)$ and defined it as:

$$a^p(t, r) := a^i(t, r)a^s(t, r^l)a^s(t, r^r)$$

In the same manner as for single-end reads we also define the property that a paired-end read r identifies a transcript t by:

$$I^p(t,r) := a^p(t,r) \sum_{t'=1}^{T} a^p(t',r),$$

where $I^p(t,r)$ is one if r identifies t and zero otherwise. Given the reduced matrix representation the probability of identification can be computed in the same manner as shown for-single end reads.

Parameter analysis

Given the above characterisations of properties and probabilities for reads, we can then compute the probability that we can identify a transcript with a library of reads. Furthermore, we show how this probability depends on the sequencing depth N. As established before, the reads can be considered independent, thus the probability of identifying transcript t with a randomly drawn read r depends on two quantities: Firstly, the probability $p(I(t,r)=1)$ that a random read r from t can be used to identify t and secondly, the probability $p(r|t)$ of a read r to come from isoform t in the sequenced library The latter probability, $p(r|t)$, is a function of the overall abundance of isoform t in the library compared to all other isoforms. Consequently, higher expressed transcripts are generally easier to identify than lowly expressed transcripts. We can, therefore, derive the probability of t not being identified by a random read:

$$P(t \text{ is not identified}) = (1 - p(r|t)p(I(t,r)=1))$$

and thus the probability that we can identify t with N reads as:

$$P(t \text{ is identified}|N) = 1 - (1 - p(r|t)p(I(t,r)=1))^N$$

Therefore, if we assume that all isoforms are equally likely to generate a read (i.e. $p(r|\cdot)$ is a uniform distribution), then we can compute the expected number of identified transcripts E_t for library size N:

$$E_t = \sum_{g \in G} \sum_{t_i^g \in T_g} 1 - (1 - \frac{1}{T} p(I(t,r)=1))^N, \tag{3.5}$$

where T_g are the transcripts of gene g and T is the total number of isoforms in the genome. It should be noted that we in this derivation assumed that genes are non-overlapping. When this is not the case overlapping genes can be merged to estimate the number of expected number of identified transcripts.

For the identification as outlined above, it was required that there was at least one read that confirmed that a transcript was identified. In some cases, however, it can be of advantage to have a stricter criterion. This could be an absolute criterion for the identification, such as a certain number of reads that must confirm the transcript or a relative criterion such as a certain percentage. In both cases, the expected number of identified transcripts with the respective identification criterion can be computed similarly as before. We can generalise formula 3.5 by replacing $P(t \text{ is identified}|N)$ with $C(x)$, where C is the CDF of the binomial distribution $\mathcal{B}_{N,\frac{1}{T}p(I(t,r)=1)}$. Then $C(N-y)$ and $C(N(100-z))$ yield the expected number of transcripts having at least y and $z\%$ of the reads that map to a transcript, respectively.

3.2.2. Models for Identification of Differentially Spliced Genes

Another task that is frequently performed when analysing the transcriptome is the detection of genes for which splicing changes between two libraries (*differential splicing*). Typically, changes in the regulation of alternative splicing lead to production of different transcripts, but the total number of transcribed transcripts from a gene remains constant (see Sec. 4.1 for a detailed explanation). Therefore, only the abundances of transcripts from the gene relative to another (*relative abundance*) changes. Genes that are differentially spliced between two samples can, therefore, be determined by detecting changes in the relative abundance of transcripts between the samples. For this task various approaches have been proposed (see Sec. 4.1). However, as for the task of transcript identification, a systematic analysis of the effect of various parameter choices on the detection of differential splicing has not been performed yet.

To address this issue, we adapt the framework that we have derived in Sec. 3.2.1 to the detection of differentially spliced genes. As is outlined in detail in Sec. 4.1, this is based on one key observation: Besides the unique reads also the informative reads can indicate a changes in splicing of a gene. Briefly, this is because every change in the splicing leads to a change in the relative and absolute abundance between transcripts. Thus, this can lead to a change in the number of reads of the regions, where the transcripts of a gene are different, i.e. the informative and unique regions. In splicing events that have a simple architecture, such as a gene with two isoforms that only differ by one skipped exon, the informative and the unique regions coincide. This can be different in more complex splicing patterns, such as the one illustrated in Fig. 3.1. In this example, some changes can still be detected, even though there does exist a unique region (e.g. whether exon 2 is skipped more often). This highlights the value of the informative region for detection of differential splicing. In the following, to simplify the notation, we will consider informative regions to include the unique regions. We, therefore, propose to adapt our framework by not only considering unique regions but also informative regions.

For the detection of changes in splicing two aspects are important: The first is the direction of the change, that is, how are the relative abundances changing. The second aspect is the absolute expression of the transcripts.

We begin by first investigating the direction of the change. For this, denote by $x_C(t)$ the relative abundance of transcript t in condition C and by $\mathbf{x}_C$ the vectors of these relative abundances. Let furthermore H be again the reduced matrix representation with entries $h_{i,j}$ for region c_i and transcript j. We then call the fraction of all reads from the gene under investigation that map to region c_j, the relative abundance of the region c_j. Then, a difference in the relative abundance of transcript t between two conditions A and B, leads to a difference in the expected relative abundance of region c_j of transcript t that is proportional to $h_{j,t}(x_A(t) - x_B(t))$. Consequently, a difference in the relative transcript abundances between $\mathbf{x}_A$ and $\mathbf{x}_B$, leads to a difference in the expected relative abundance of region j that is proportional to $< H_j, \mathbf{x}_A - \mathbf{x}_B >$. It should be noted that the entries of the vector of relative abundances $\mathbf{x}_C$ sum to one. Therefore, $< H_j, \mathbf{x}_A - \mathbf{x}_B >$ can only be non-zero if the region c_j is an informative region. Finally, we can derive a criterion for when a direction of change $\mathbf{x}_A - \mathbf{x}_B$ in the relative transcript abundance can be detected. As a change in any of the expected relative abundance of the regions indicates differential splicing, a necessary and sufficient criterion for its detection is:

$$\|H^T(\mathbf{x}_A - \mathbf{x}_B)\|_2 > 0 \tag{3.6}$$

This shows, that changes of direction $\mathbf{x}_A - \mathbf{x}_B$ that are in the kernel of the matrix H cannot be detected. However, because this result holds only for the expected number of reads, 3.6 is only a sufficient criterion for infinite sequencing depth.

For finite numbers of reads, the relative abundances are random variables and thus also the relative abundances of the informative regions are random variables. In order to determine whether the difference in the relative abundances of regions is significant, it is therefore best to use statistical tests. These account for the uncertainty in the relative abundance estimates when determining the significance of a difference.

In order to model the detection of differential splicing, we use a Poisson homogeneity test on the informative regions of a gene. Briefly, this method assumes that the number of reads that map to a region follows a Poisson distribution. It then tests for differences in relative abundance for each informative region and combines the resulting p-values by Bonferroni's correction (see Sec. 4.2.1 for details).

Under the assumption that the reads are distributed uniformly along the transcripts the distribution of reads along the transcript that maps to a gene g in a condition C is given by two parameters. The distribution n_g^C from which the total number of reads is sampled and the relative transcript abundances $\mathbf{x}_C$. The probability of detecting a change in gene g between two sample A and B can therefore be computed by:

$$P(g \text{ is detected} | n_g^A, n_g^B, \mathbf{x}_A, \mathbf{x}_B, \alpha) \tag{3.7}$$

Here, α is the confidence level of the statistical test. Depending on the assumption on the distributions of n_g^A and n_g^B, analytical solutions for the probability may exist. When this is not the case, a Monte Carlo estimation strategy can be used to estimate the probability 3.7. In this strategy, reads are repeatedly sampled according to the gene expression distributions, the relative transcript abundances and the uniform distribution of the reads along the transcripts. Subsequently, the probability 3.7 can be estimated by the fraction of times the test is significant.

Finally, using the probability 3.7 the expected number of detected changes in splicing can be computed as:

$$E_t = \sum_{g \in G} P(g \text{ is detected} | \mathbf{x}_A, \mathbf{x}_B, \alpha),$$

The expected number of detected changes in splicing can then be used in order to determine optimal parameters for sequencing. It should be noted that when the choice of the statistical test should be optimised it is important to compare the tests at the same Type I error rate. Otherwise, a test that calls all genes significant would be the best choice.

3.3. Results and Discussion

3.3.1. Transcript Identification

We applied our probabilistic model to assess the influence of the sequencing parameters on the identification of transcripts. We did this for two organisms with differing splicing complexity according to their gene annotations: *H. sapiens*, where almost 95% of the genes have multiple isoforms and the nematode *Caenorhabditis elegans*, where only about 25% have more than one annotated isoform. To fit the models, we used the WS199/200 genome annotation for *C.*

elegans and the hg19 annotation for *H. sapiens*. We assumed that the probability of sequencing a read from a transcript was uniform over all transcripts and also that the probability of reads starting at a certain position was uniform in the sense that was introduced in the model derivation. These uninformative distributions allow for an unbiased analysis of the effect of parameter choices. For our analysis, we also assumed that for paired-end reads the sequenced ends were 76 bp long and that for single-end reads the reads were 152 bp long (being the same as a paired-end reads where the read-ends were next to each other). Furthermore, we assumed that the insert size for paired-end reads had a deviation of at most ±12.5% of the chosen library insert size.

We first analysed how many of the 5,718 *C. elegans* and 66,654 *H. sapiens* transcripts of genes with multiple isoforms can be identified with single-end reads when the sequencing depth can be arbitrarily high. For this analysis, we assumed infinite coverage in order to obtain an upper bound on the expected number of transcripts. Therefore, we counted the number of transcripts, where the probability of identification, $p(I(t, r) = 1)$ was bigger than zero for any read r. Thereby, we found that 4,411 (77.1%) transcripts could be identified for *C. elegans* and 61,716 (92.6%) for *H. sapiens*.

We then analysed the utility of different insert sizes for transcript identification using paired-end reads. We computed the expected number of transcripts for all insert sizes from 152 to 1,152 bp (see Fig. 3.2). We found that for *C. elegans* the optimal insert size was 315 bp, which is similar to the commonly used 300 bp. With this optimal insert size, 4,468 (78.1%) transcripts could be identified. For *H. sapiens*, we found that the optimal insert size was 241 bp, which is slightly shorter than the commonly used insert size. For this optimal insert size 62,549 (93.8%) transcripts could be identified.

Next, we analysed the minimal distance between the read ends that is necessary for transcript identification. The cumulative distribution of the expected number of identified transcripts for these distances is shown in Fig. 3.3. We found that among the 260 isoforms of *C. elegans* that are generated from genes with more than two annotated isoforms, 260 of them can be identified with paired-end reads that have a total insert size between 77 bp and 1,152 bp. Out of those, 40 (15.4%) isoforms could be detected with overlapping read ends, another 73 (28.1%) were identifiable with libraries of insert size up to 300 bp and 111 (42.7%) isoforms required an insert size between 300 bp and 800 bp. For *H. sapiens* 1,917 isoforms of 5,688 (33.7%) were identifiable with overlapping paired-end reads, while 2,037 (35.8%) were identifiable by reads that were at most 148 bp apart. For an insert size of between 300 bp and 800 bp 1,460 (25.7%) were identifiable.

In the previous analyses, we assumed that we had unlimited sequencing depth and that one read suffices for transcript identification. In practice, however, it is also important how efficient detection is. We, therefore, analysed how many of the transcripts can be detected if we required that at least 10% of the isoform reads mapped to it. When requiring this, we found that only 2,704 (47.3%) and 47,608 (71.4%) transcripts are expected to be found for *C. elegans* respectively *H. sapiens* using single-end reads. Similarly, for paired-end reads a much lower number of transcripts can be identified efficiently. For *C. elegans*, only 3,053 (53.3%) transcripts can be identified for the optimal insert size (294 bp) and 3,782 (66.1%) if all insert sizes would be taken. For *H. sapiens*, only 52,571 (78.9%) transcripts can be identified for the optimal insert size (288 bp) and 60,017 (90.0%) if all insert sizes would be taken. We observed that for both organisms the decrease in efficiency, if a sub-optimal insert size was chosen, was more pronounced (see Fig. 3.2) compared to the previous criterion.

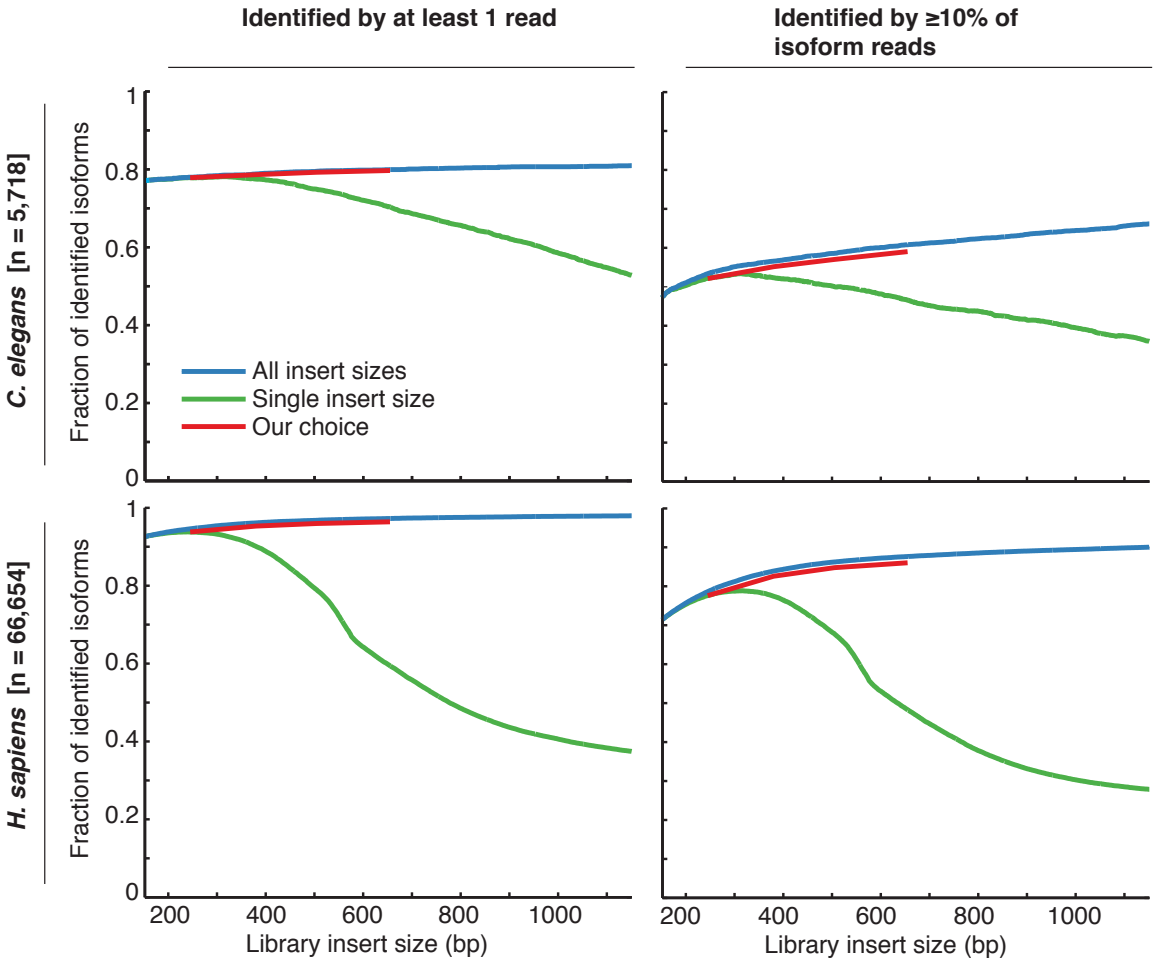

Figure 3.2.: Shown is the insert size utility for *C. elegans* and *H. sapiens* using two identification criteria. Shown in blue is the expected number of identified transcripts with all libraries of all insert sizes up to a given insert size and in green the expected number for a given insert size. The insert size utility for our choice of four libraries is shown in red. This figure has been adapted from our publication [158].

In this analysis, information on the approximate distance between reads was used. To elucidate how much information is contained in the specific distance between the read-ends we further computed the expected number of identified transcripts in absence of the distance information, when using all libraries. In this scenario, the same number of transcripts as with distance information is expected to be identified for *C. elegans*. For *H. sapiens* 57,839 (11.4% less) were expected to be identified. Interestingly, for *C. elegans* 3,635 (3.9% less) were expected to be identified efficiently and 49,889 (5.1% less) for *H. sapiens*. This suggests that particularly for organisms with high splicing complexities, the paired-end distance is a valuable source of information for transcript identification.

The before mentioned criterion for efficient detection required that at least 10% of the reads identified the transcripts. However, other criteria are available, such as requiring at least a certain number of reads to identify transcripts. To investigate how these different criteria affect transcript identification, we computed the expected number of identified transcripts when requiring at least 2, 5 and 10 reads as a function of the number of reads that were

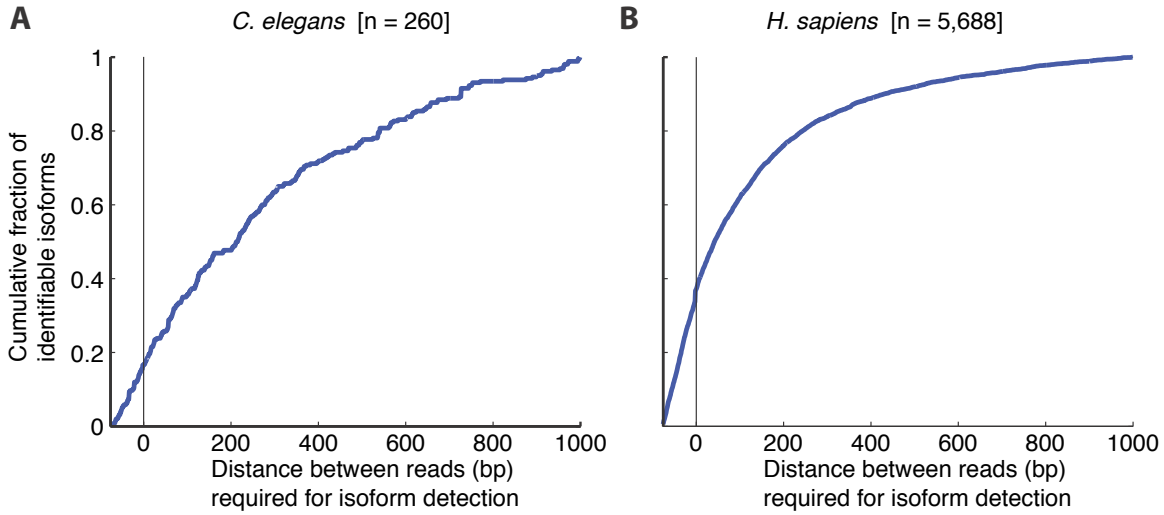

Figure 3.3.: Shown is the cumulative fraction of identifiable isoforms from genes with three or more annotated isoforms for *C. elegans* (**A**) and *H. sapiens* (**B**). Insert sizes on the left size of the vertical bar at 0 indicate overlapping read-ends. This figure has been adapted from our publication [158].

mapped per transcripts (see Fig. 3.4). We found that for both, *C. elegans* and *H. sapiens*, the expected number of identified transcripts for all required read criteria converged to the one, where only one read was required as the number of reads increased. For the original length criterion the information gain saturated after the first 100 reads per transcripts, but for higher read threshold saturation did occur only at higher coverage.

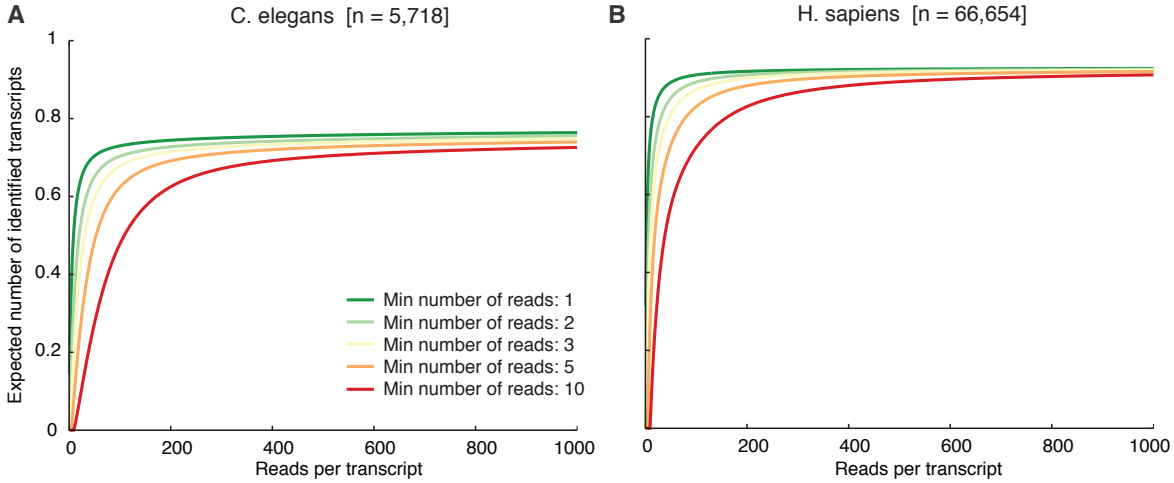

Figure 3.4.: Shown is the expected number of identified transcripts for different numbers of required reads for transcript identification for *C. elegans* (**A**) and *H. sapiens* (**B**).

Our model shows that some transcripts can only be identified, when two distant splice events are jointly observed by the two read ends of a read. Therefore, only reads with a specific insert side can identify these transcripts. This implies that even with the choice of the optimal insert size, some transcripts cannot be identified that are identifiable with a another insert size. We,

therefore, suggest sequencing multiple insert sizes for transcript identification. To investigate the gain of having multiple insert sizes, we computed the expected number of transcripts if all insert sizes up to a certain insert size were to be sequenced. This allows obtaining an upper bound on the information gain that can be achieved with multiple insert sizes (see Fig. 3.2). We, therefore, computed the fraction of transcripts that can be identified if all insert sizes were to be used. We found that if all insert size were taken $4,629$ (81.0%) and $65,313$ (98.0%) could be identified for *C. elegans* and *H. sapiens* respectively, showing that additional information can be gained when all insert sizes are used. In practice, sequencing all insert sizes separately is prohibitive because of the resources required for sequencing. Also mixing of libraries with different insert sizes to reduce the number of sequenced libraries is also not an option, as thereby the distance information is lost and transcripts would not be identifiable any more based on the distance of their events. Therefore, if the aim of the study is to get a full picture of the transcriptome, we suggest using a small subset of different insert sizes that are uniformly distributed to approximate the selection of all inserts sizes. The effect of such a strategy is exemplified in Fig. 3.2. For this, we have shown the effect of using four insert sizes (215 bp, 350 bp, 475 bp and 625 bp). It can be seen that this strategy allows identifying almost as many transcripts, as if all insert sizes were taken. This can be achieved with only a limited increase in the sequencing effort.

To further show that these theoretical considerations are indeed of practical importance, we generated four libraries for *C. elegans* with the afore mentioned insert sizes and aligned the reads using PALMapper [75] (for details on the experimental protocol we refer to [158]). We predicted which exon skips and intron retentions could be detected using our libraries and we also predicted novel isoforms (for details see [158]). We found surprisingly little overlap between the annotated exon skips and intron retentions. Only 201 of the $1,021$ (16.4%) novel intron retentions and 343 of 973 (35.3%) of the detected exon skips were annotated. When combining all libraries we found 993 novel transcripts. However, only 441 (44%) of these unannotated isoforms were supported by all libraries. We also found between 49 and 58 of the transcripts being private to only one of those libraries. This shows that multiple insert sizes can help to obtain a more complete catalogue of expressed transcripts.

A representative example for *C. elegans* that further illustrates the potential of multiple insert sizes to get a broad view on splicing is shown in Fig. 3.5. This example shows the gene mdt-28 that has five annotated isoforms. As can be seen from the coverage in Fig. 3.5(B) roughly 6.5% of the transcripts contain the longer exon 3. According to the annotation all transcripts that contain this exon contain the short exon 7. This is not contradicted by the two libraries with the short insert sizes. However, examining the coverage of all reads from the longer two insert sizes that contain exon 3 and exon 7 clearly shows that the longer exons 8 and 9 were used (see Fig. 3.5(C)) in these transcripts and thus showing that the gene annotation is incomplete.

In summary, our aim was to study the combined effects of parameter choices on the information gain of the experiment. To achieve this, we have derived a probabilistic model of the utility of several parameters: read type, insert size, read length and sequencing depth. We have applied it to two organisms with distinct splicing complexity in order to show that it is generally applicable to guide experimental design.

Our framework provides several insights when a single library is used. Firstly, we have shown that paired-end information helps in identifying transcripts. Furthermore, we have shown that the fraction of transcripts that can be identified depends strongly on the chosen insert size. This is especially the case if a strict criterion (e.g. requiring at least ten reads) is used. Furthermore, using our model we were able to determine the optimal insert sizes for both

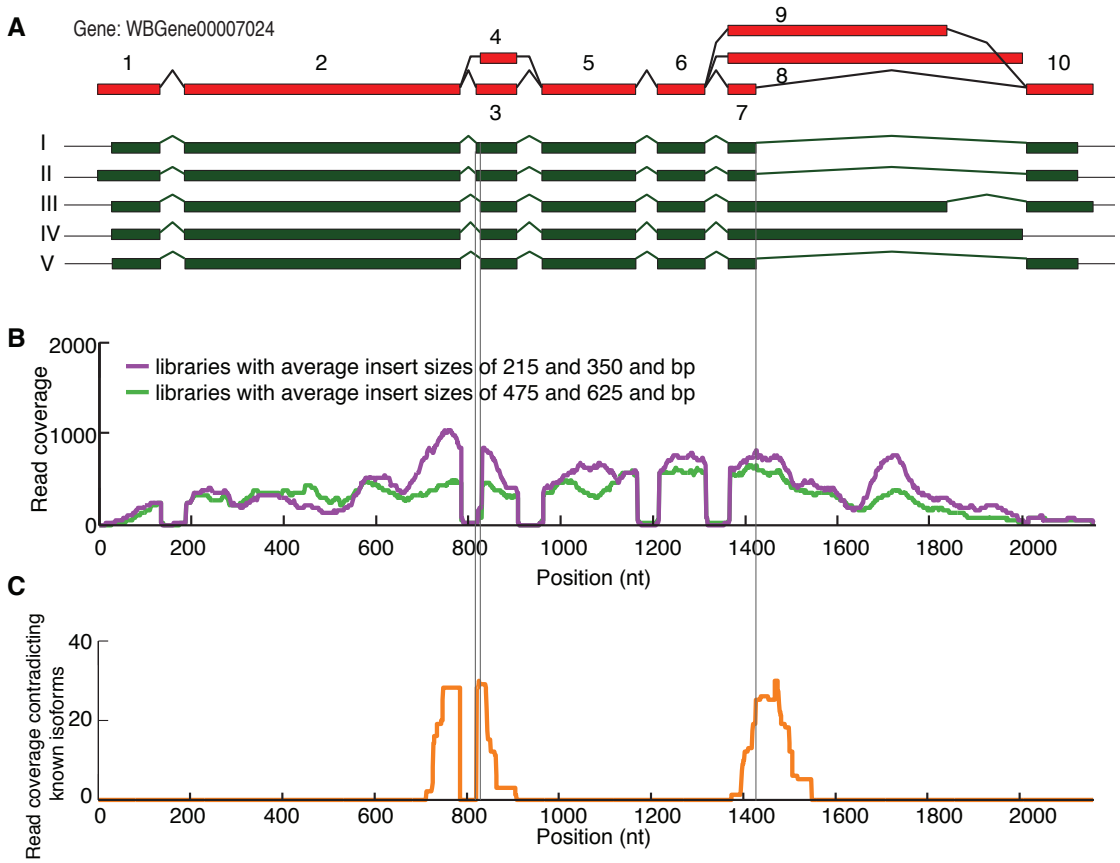

Figure 3.5.: Example of utility of multiple insert sizes. (**A**) The Wormbase gene model WBGene00007024 (mdt-28) splice graph (red) and its individual transcripts (dark green). (**B**) Read coverage of the gene mdt-28 from libraries with average insert sizes of 215 and 350 bp (purple) and 475 and 625 bp (light green). (**C**) Reads from the 475 and 625 bp insert libraries where one pair of the read covers exon 3 and the second read extends beyond exon 7, indicating that they must originate from exons 8 or 9. This figure has been adapted from our publication [158].

organisms and the number of reads per transcript for saturation of the identification. We, therefore, believe that our model provides valuable information for improving the design of RNA-Seq experiments.

Besides this, we could show that only a subset of the identifiable isoforms can be identified with a given selection of parameters. This was shown both in a theoretical analysis and was also experimentally confirmed. Based on this insight, we showed that if a combination of four libraries with different insert sizes is sequenced, then a bigger fraction of the identifiable isoforms can be identified, thus allowing a more comprehensive view on the transcriptome.

In our model we assumed uninformative distributions for the transcript abundance and also for the read distribution along a transcript. This has the advantage of providing an unbiased analysis. If, however, for specific analyses, prior information on the distribution of transcript abundances and the positional distribution of reads is available, we suggest incorporating this in to the model in order to adapt the analysis. For estimation of the expected number of transcripts, we suggest to use a Monte Carlo estimation using the two distributions.

Based on our results, we suggest different read-types depending on the organism and the identification criterion. For organisms with a low splicing complexity, we observed that the difference in information gain between single-end and paired-end reads of the same length is minor. For organism with a high splicing complexity such as *H. sapiens*, there is a larger difference when requiring efficient identification. This suggests that for these organism paired-end reads can provide a significant in information gain compared to single end-reads.

3.3.2. Detection of Differential Exon Usage

We applied the probabilistic framework to reveal the dependence between the power of statistical tests to detect differential splicing and various parameters, namely the sequencing depth, the fold-change and the length of splicing events. Here, we only considered single-end reads of length 80 and considered exon skips of exons with different lengths (25 bp, 50 bp, 100 bp, 134 bp, 200 bp, 300 bp, 500 bp and 750 bp). We determined for these skipped exons the coverages and the minimal fold changes that are necessary for detection using a Poisson homogeneity test (see Sec. 4.2.1). For this, we tested on the number of reads in the exon and assumed that an event can be detected if the p-value was smaller than 0.05.

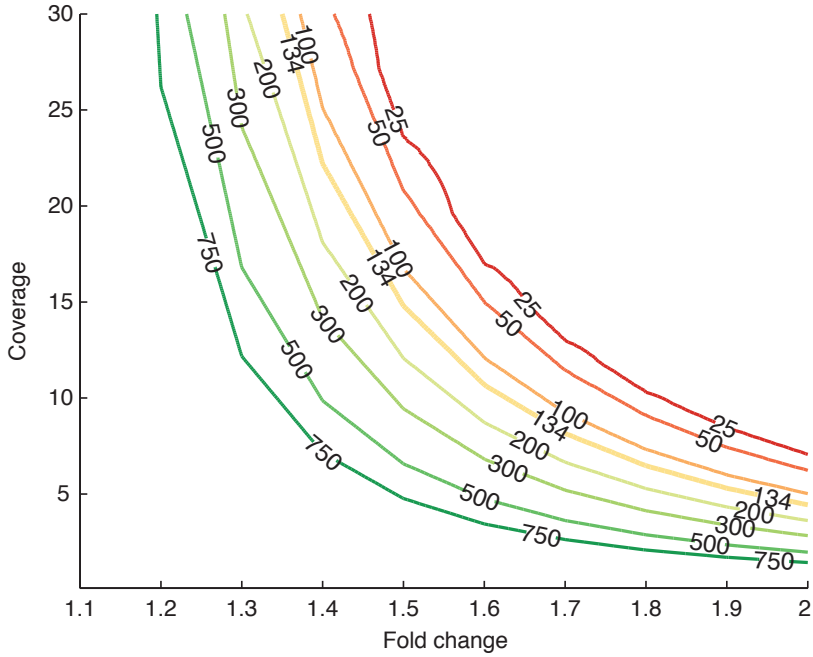

Figure 3.6.: Shown are minimal lengths of detectable splicing events of various lengths given the average coverage and the fold change of the event. Fold changes that are on the right side of the respective curves are detectable.

We observed that as the coverage increased the minimal fold change for detection decreased (see Fig. 3.6). For example for an exon of 134 bp length (the average length of a *C. elegans* exon) at least a coverage of 4.6 was required to detect a 2-fold change, whereas with a coverage of 15 an 1.5-fold changes could be detected.

Furthermore, we found that changes in longer exons are easier to detect than in smaller regions. This can be explained longer regions have, for the same coverage, more reads that map into it than shorter regions. Therefore, as there are more observations, the statistical power to detect a change is higher. Interestingly, for very short exons the detectability was

still reasonable. This is because for a short exon also all the reads that only partially overlap are informative. Therefore, the effective length of the region where informative reads can start is almost one read length longer than the exon itself.

As discussed in detail in Sec. 4.2.3, the used Poisson-based test does not account for biological variance. Therefore, the detection boundaries can be optimistic in cases, where the biological variance is not negligible.

3.4. Summary

The design of an experiment has a great influence on the success of the experiment. Therefore, careful planning of the experiment is important. In this chapter, we have presented the first genome-wide model to systematically assess the influence of parameters such as insert size, the read type or library size on transcript identification. For this, we have developed a probabilistic model to compute the effect of different parameter choices. Furthermore, we have shown that this probabilistic model is very general and can easily be adapted to model other tasks, such as the detection of differentially spliced genes. Using our model, we were able to determine the optimal parameter choices for transcript identification and derive guidelines for transcript identification. We have also shown that it is necessary to sequence multiple insert sizes in order to identify all transcripts. Finally, we were able to gain insights on the factors that influence the detection of differential splicing. This showed the value of this approach in order to understand the effect of various parameters and to improve the design of RNA-Seq experiments for the questions at hand.

4. Detection of Differential RNA Processing

4.1. Motivation

The introduction of high-throughput sequencing technologies such as RNA-Seq allowed examining the transcriptome to an unprecedented extent. The richness of RNA-Seq data has inspired the development of many new methods in order to fully exploit the potential of this new data. These methods have in turn led to many fundamental insights on the mechanisms that shape the transcriptome [130].

In this chapter we focus on the approaches to analyse post-transcriptional RNA processing (e.g., alternative splicing or 3' cleavage) using RNA-Seq data. However, although the analysis of RNA processing is the focus of this chapter, we find it instructive to first discuss a closely related problem, namely the analysis of gene expression regulation. The questions and challenges that arise in each of these analyses are fundamentally similar. Nevertheless, typically the analysis of gene regulation is less complex than the analysis of RNA processing.

Among the processes shaping the transcriptome, one that was investigated first was the regulation of gene expression (e.g. [126]). In the beginning, the main challenge was to quantify gene expression. This required the development of strategies for read mapping. The main challenge here was efficiency, as typically there are millions of reads to be mapped. For this task, several approaches have been proposed, e.g. PALMapper [75] or TopHat [172]. Using the mapped reads, the expression of a gene was then usually defined as the number of reads that map to this gene, corrected for library size.

However, quantifying the gene expression alone only provides a steady-state view on the transcriptome, while to understand the regulation of gene expression, one needs to study the dynamics of the transcriptome. A natural way to study these dynamics would be to detect genes that change upon perturbation of the environment. These genes would then allow to identify the involved pathways and to reveal the regulatory architecture of these genes. This motivated the development of methods to detect changes in gene expression.

For this purpose, several statistical tests have been proposed. These tests rely on the assumption that the number of reads that are expected to map to a gene is a monotone function of the expression of this gene. Therefore, changes in gene expression are indicated by read count changes. The statistical models that have initially been proposed to detect differential gene expression model the number of mapped reads as either a binomial distribution or one of its limits, the Poisson distribution [110, 132].

More recently, however, it has become clear that these methods underestimate the gene expression variability. This is because these models only account for the variability in the measurements due to sequencing (*shot noise*) and do not account for additional biological variability of gene expression. Therefore, they tend to have higher false positive rates than estimated, especially among highly expressed genes. Consequently, novel methods have been proposed that account for the additional variance to remedy this shortcoming (e.g. [5, 66, 139, 140]). Most of these methods have in common that they assume that the counts are distributed according to a negative binomial distribution.

Besides providing a means to understand gene expression regulation, RNA-Seq furthermore allows to examine other processes that shape the transcriptome, such as RNA processing. In fact, the reactions that are part of RNA processing define which isoform is produced and thus also determine cellular transcript abundance. These abundances in turn determine the read distribution that results from sequencing. Therefore, the *read distribution* of a gene provides information on the transcripts and thereby also on the reactions involved in the RNA processing (see Fig. 4.1 for an illustration). However, compared to the relatively straightforward quantification of gene expression, the quantification of transcript expression is considerably more challenging. This difficulty stems from two issues: First, the transcripts need to be reconstructed from the reads. Second, the transcript from which a read originates can often not be uniquely identified (see Sec. 3.2.1) and thus the original abundances need to be inferred. To address these problems, there have been numerous methods proposed to either quantify annotated transcripts based on existing annotations or additionally also to reconstruct transcripts that were not yet annotated (e.g. [14, 24, 55, 63, 76, 83, 136, 173]). These methods typically achieve the deconvolution of transcript abundances by either explicitly or implicitly assigning the reads to the transcripts. Some of these methods, such as MISO [83] or BitSeq [55] perform a full Bayesian inference to model the uncertainties in the read assignments.

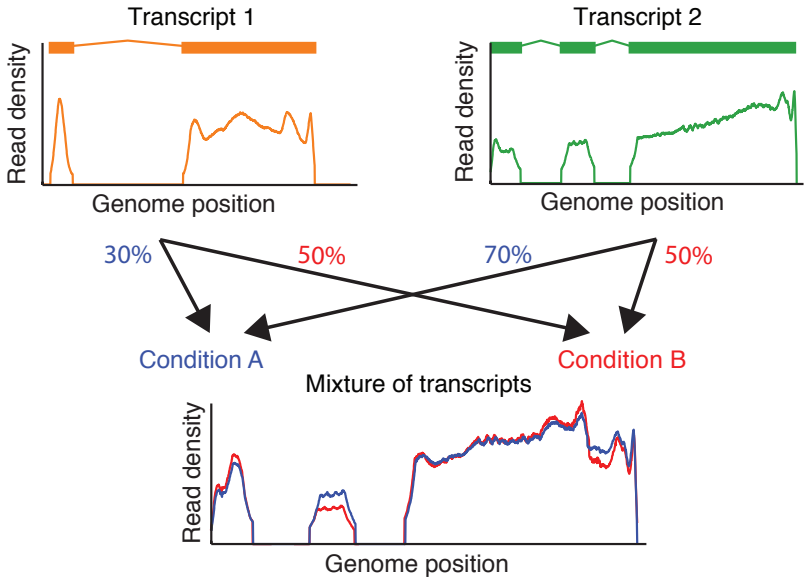

Figure 4.1.: Mixture of two transcripts that results from different RNA processing reactions. On the top, the two resulting transcripts of two RNA processing reactions are shown together with the read density one would observe if they were sequenced separately from each other (orange and green). On the bottom, the read densities for two mixtures of the transcripts are shown. The intensity of the reactions for the conditions A (blue) and B (red) is different, which is reflected by the difference of the read densities.

Similar to the development of methods to analyse gene expression, the availability of quantification approaches for transcript abundance motivated methods to detect differential transcript abundance, in order to shed light on the underlying regulatory processes. All of these proposed methods exploit the fact that changes in RNA processing lead to the synthesis of different transcripts and these changes are reflected by changes in read distributions of the

corresponding genes (see Fig. 4.1). Therefore, changes in RNA processing can be detected by changes in read distribution. The observed changes in transcript abundance can be factored into two components: Into (1) a change in relative abundance and (2) a change in gene expression. In order to analyse RNA processing, it is advantageous to investigate changes in relative abundance. This is, because changes in RNA processing tend only to influence the relative abundance of transcripts. In contrast changes in gene expression regulation tend to change the overall transcript expression but not the relative transcript expression. Therefore, focusing on changes in relative abundance allows analysing RNA processing without confounding from gene expression. In the following, we will thus concentrate on the changes in relative transcript abundance.

In order to detect changes in relative abundance and, thereby, genes with differential RNA processing, several approaches have been proposed. These approaches can be divided into two groups. Firstly, the group of methods to detect changes that are restricted to a certain locus in the transcript, e.g. the skipping of an exon or the retention of an intron. For this problem, standard statistical tests have been proposed. For example in [181], the ratio of reads that confirmed the two different splice forms was compared, using a 2x2 contingency table and Fisher's exact test. More elaborate approaches that have been proposed recently additionally also account for biological variance and other confounding factors [6]. One shortcoming of this group of methods is, however, that only one splicing event at a time can be examined and that they thus are restricted to the study of single events. The second group of methods tries to detect changes in relative abundance of entire transcripts. For this approach, typically a gene annotation is used. This has the advantage that the knowledge on the dependence of distant splice events that is encoded in the annotation can be exploited. Typically, these methods quantify in a first step the abundances of transcripts and in a second step use these estimates in order to detect differential relative abundance. This approach provides a high interpretability as the changing isoforms can be easily identified. On the other hand, there exist several shortcomings of this approach. One is the problem that there may exist several optimal read assignments to different transcripts and thus optimal quantifications [70, 94]. Therefore, establishing meaningful distances between quantifications is challenging. Furthermore, variability of read densities can be amplified and propagated to the quantifications, which leads to unstable estimates (see Fig. A.3 for an instructive example). This problem can be partially overcome by estimating confidence intervals for the quantifications, such as by conducting a Bayesian inference using Markov Chain Monte Carlo approaches [55, 83] or evaluating the Fisher information matrix [76]. A drawback of these methods is , however, that they are computationally expensive and they require many assumptions in their models, which may not always be satisfied in practical applications. Therefore, there is still a need for robust approaches to detect differential relative transcript abundances that do not require the challenging transcript quantification.

In principle, it is beneficial to use the gene annotation for detecting differential transcript abundance. However, it should be kept in mind that these annotations are derived from existing experimental results. Hence, novel previously unseen transcripts, such as expected in splicing factor knockdown mutants or cancer, are likely not to be included in the annotation. In these situations, a comprehensive analysis of RNA processing is therefore not possible when relying only on the gene annotations. The critical dependence on existing annotations of many methods is especially cumbersome with newly sequenced organisms, where an annotation is usually of poor quality. To detect differential RNA processing in cases where the genome annotation is missing, a solution is to first infer the transcripts (e.g. using [14, 59, 154, 163, 173]) followed by quantification and testing for differential RNA processing. However, a complex pipeline with many assumptions is statistically hard to track and the reliability of

the obtained predictions is unclear.

But, there are also cases in which it is not established how a meaningful annotation could be defined (e.g. for footprinting or antisense transcription data). In this situation the approach, to first reconstruct the annotation, then quantify and finally test, cannot be applied. An alternative to this complex approach is to use established nonparametric statistical tests such as the Kolmogorov-Smirnov-test (K-S-test). However, this approach typically fails to account for the high dimensional nature of the reads and for the biological variation. Therefore, practical methods that can detect changes in RNA processing without needing a gene annotation, are still needed.

In summary, the approaches presented above are all either not capturing complex splicing events or solving a much more complicated task to detect differential RNA processing than necessary. Also there is still a lack of methods that can be applied, when the gene annotation is incomplete or missing.

In this chapter, we will propose a novel approach to directly test for differences in relative isoform abundance change without the need to quantify the transcripts. We will present a series of tests for these two settings. For situations when a complete gene annotation is available, we will present a novel test (*rDiff.poisson*). Furthermore, we will show how the nonparametric MMD-test can be applied for testing without gene structure. A similar approach has recently also been taken by [157] in order to detect differential splicing and by [153] for detecting shape changes in ChIP-Seq data sets. We will then show how our two tests can be extended to also account for biological variance (*rDiff.parametric* and *rDiff.mmd*) and present an extension of rDiff.mmd to increase the power for detecting differential splicing (*rDiff.nonparametric*). Besides this, we will also show that rDiff.mmd can be adapted to detect differential RNA secondary structure (*sDiff*). Lastly, we will present a generalisation of rDiff.mmd for an association testing setting (*rDiff.gmmd*). In this chapter we will furthermore, present an evaluation of the proposed methods on simulated and experimental data. Finally, we will show applications of our methods to reveal regulatory mechanisms of RNA processing.

4.2. Methods

4.2.1. Detection of Differential RNA Processing with Gene Annotation

As established before, changes in the relative abundance of transcripts translate to changes in the read distribution of a gene. Therefore, these abundance changes can be detected by studying changes in the read distribution. For this purpose, however, not all regions of a gene are equally informative. For example, when there are two isoforms that only differ by an exon skip, then a change in the relative abundance will lead to a change in the expected number of reads in that skipped exon and the neighbouring exon junction. However, in the other parts of the gene the expected number of reads will remain unchanged. These parts of the gene are therefore not informative for detecting a change of relative transcript abundance. In general, the regions that reflect the changes are the informative ones, i.e. those contained in at least one but not all transcripts (see Sec. 3.2.1). It is therefore reasonable to focus on the change in relative abundance of these regions, for detection of differential relative transcript abundance. Based on the contained transcripts, the informative regions can be further grouped into larger non-contiguous regions (alternative regions), as shown in Fig. 4.2. Thereby, all positions are grouped together that have the same directionality of change upon a change in transcript abundance. As will be shown later, this also has the advantage that the number of regions is

reduced, which increases the power of tests based on these regions.

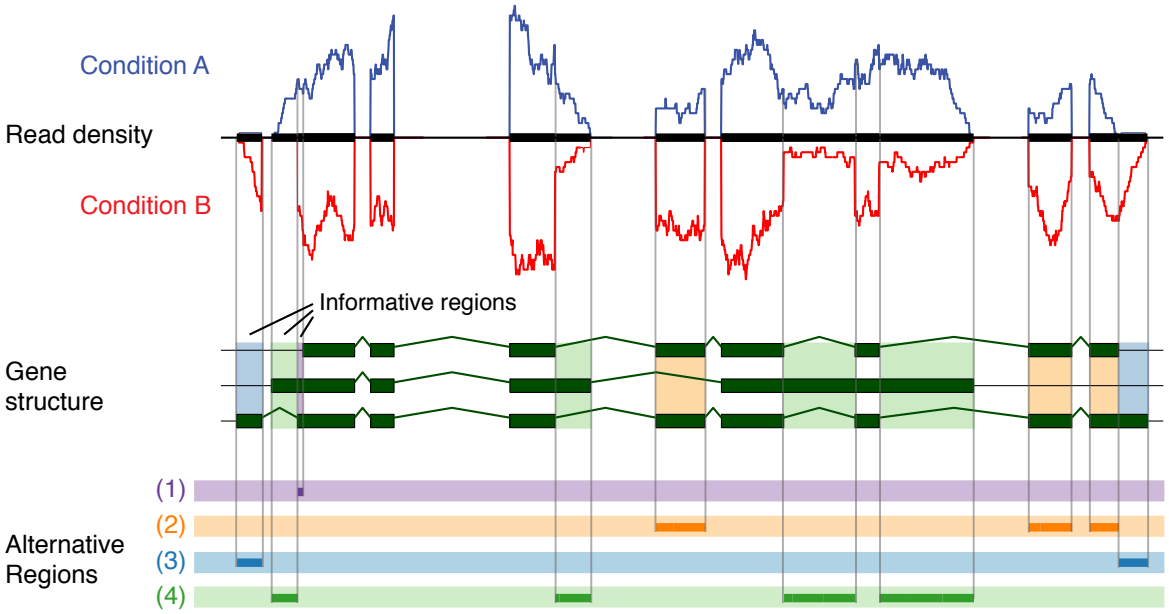

Figure 4.2.: Illustration of alternative regions used by our methods, rDiff.poisson and rDiff.parametric. Shown on top is the coverage in two conditions, condition A (blue) and condition B (red) and the exonic regions (thick black lines). Shown below is the gene structure (green). The informative regions of this gene are shown as coloured regions in the gene structure (light violet, light orange, light blue and light green). Informative regions that emerge from the same transcript combinations are shown in the same colour. The derivation of the alternative regions is shown below. They are obtained by grouping together the informative regions with the same colouring. Differences in the read distribution of non-alternative regions are only due to shot noise.

This distinction between different regions of genes can be used to derive statistical tests for the detection of differential relative transcript abundance, as we will show below. This can be done for each gene independently as follows: As the testing is carried out for each gene g individually, we omit its index for simplicity of the notation whenever possible. Assume that we would like to determine whether a gene g has differential relative transcript abundance between two conditions A and B. Assume furthermore that $R_g = \{r_1, \ldots, r_{n_g}\}$ are its alternative regions and that N_r^A and N_r^B are the number of reads that map to these regions in condition A and B respectively. Assume finally that E_g^A and E_g^B is the expression of the gene g in condition A respectively B (computed as the number of reads mapping to all non-alternative regions).

When the biological variance is small it is plausible to assume that the number of reads at given loci follows a Poisson distribution. We therefore assume that $N_r^A \sim \mathcal{P}(\lambda_r^A)$ and $N_r^B \sim \mathcal{P}(\lambda_r^B)$ are distributed according to a Poisson distribution with (unknown) intensities λ_r^A and λ_r^B respectively. Under these assumptions the Null hypothesis $\mathcal{H}_0$ for testing is then that the ratio of the intensities equals the ratio of the gene expressions, i.e. the change in

intensities can be explained by a changed gene expression:

$$\mathcal{H}_0 : \quad \frac{\lambda_r^A}{\lambda_r^B} = \frac{E^A}{E^B}$$

Accounting for differences in library size is hereby not necessary as both ratios are equally affected by it. It has been established [135] that for this test conditioning on the total number of reads observed $N_r^A + N_r^B$ does not change the ratio and thus the probabilities of the observed counts under the Null hypothesis is given by:

$$P(\mathcal{H}_0 | N_r^A, N_r^B, E^A, E^B) = \mathcal{B}_{N_r^A + N_r^B, \frac{E^A}{E^B}}(N_r^A)$$

Using this distribution the p-value p_r for the alternative region r can be computed as:

$$p_r = 2 \sum_{i=0}^{\min(N_r^A, N_r^B)} \mathcal{B}_{N_r^A + N_r^B, \frac{E^A}{E^B}}(i)$$

This testing approach is optimal in the sense that it is a uniformly most powerful test for an alternative region [98]. For large number of reads the computation can be sped up using the *de Moivre-Laplace approximation* [37] of a binomial distribution with a normal distribution:

$$\mathcal{B}_{n,p} \approx \mathcal{N}_{np, np(1-p)}$$

For this approximation, an analytic expression for the p-value can be obtained by integrating the tails of the Gaussian distribution.

Finally, in order to obtain a p-value p_g for the gene, the p-values from the alternative regions can be combined using the Bonferroni correction [25]:

$$p = |R_g| \min_{r_i \in R_g} p_{r_i}$$

This correction provides a conservative estimate of the p-value for the gene g. It is also possible to combine the evidence from the alternative regions using other methods such as Holm's step-down method [71]. In this work we will refer to this testing approach as rDiff.poisson.

4.2.2. Gene Annotation Free Detection of Differential RNA Processing

When the gene annotation is not available, parametric methods such as rDiff.poisson or quantification-based approaches cannot be directly applied. This is because rDiff.poisson depends on the annotation in order to determine the alternative regions and the quantification-based approaches need the annotated isoform in order to assign reads to them. Another more elegant solution in this situation is to reformulate the problem as a testing for the identity of the compared read distributions, i.e. as homogeneity testing. As the parametrisation of the read distributions is typically unknown, nonparametric tests such as the K-S test can be applied. However, this test has the limitation that it is only defined for one-dimensional spaces and thus the higher dimensional structure of the reads cannot be captured. To overcome this limitation we suggest to use the Maximum Mean Discrepancy (MMD) test (see Sec. 2.5.4) to detect differential RNA processing. This can be done as follows.

Let again g be a gene with length l_g, $\mathcal{X}$ the space of all reads that map to it and $X^A = \{x_1^A, \ldots, x_{N^A}^A\} \subseteq \mathcal{X}$ and $X^B = \{x_1^B, \ldots, x_{N^B}^B\} \subseteq \mathcal{X}$ be the reads mapping to g in condition A respectively B. As discussed in Sec. 2.5.4, we need to establish a reproducing kernel Hilbert

space (RKHS) $\mathcal{H}$ and a mapping $\phi : \mathcal{X} \to \mathcal{H}$ of the reads into it. For this we choose $\mathcal{H}$ to be $\mathbb{R}^{l_g}$ with the Euclidean scalar product. The image $\phi(X_i)$ for a read x_i is defined to be vector the $\phi(x_i) \in \mathbb{R}^{l_g}$, where $\phi(x_i)[j]$ is the j-th entry of $\phi(x_i)$ such that $\phi(x_i)[j] = 1$ if x_i maps to the j-th position and 0 otherwise. Therefore, the resulting kernel $k(\cdot, \cdot)$ is the linear kernel on the embedding defined by ϕ. We can then compute for each sample $C \in \{A, B\}$ the mean embedding:

$$\mu^C = \frac{1}{N^C} \sum_{i=1}^{N^C} \phi(x_i)$$

The test statistic D is then the distance between these means of A and B (discrepancy) in $\mathcal{H}$:

$$D := \|\mu^A - \mu^B\|_2$$

When using the linear kernel, the mean μ^C is the mean coverage at each position. Therefore, the discrepancy is the L^2-norm of the difference of the two mean coverages. One advantage of this embedding is that the discrepancy can be computed in linear time compared to the quadratic time that is in general necessary when using other kernels. However, as the linear kernel is not universal (in the sense defined in [164]), the mapping into the RKHS is not injective and therefore some differences in read distributions that do not lead to changes of the coverage cannot be detected. However, since these types of changes are rare (data not shown) the benefit of faster computation outweighs this limitation.

The observed discrepancy between the two means alone does not allow concluding how unlikely it is to observe it under the Null hypothesis. In order to obtain a p-value for the observed discrepancy D we therefore perform bootstrapping. This is done by comparing D to discrepancies from two means that are sampled from the Null hypothesis. As under the Null hypothesis the two distributions from which the reads are drawn are the same, the mixture of them is also the same. We therefore can sample from the union $X^A \cup X^B$ of the reads two new samples of size N^A respectively N^B from the Null distribution and compute the discrepancy D_t between the means of those two random samples. This Null discrepancy can then be compared to D to determine whether the discrepancy between the observed samples is bigger than the discrepancy observed by chance. This can be done T times to obtain a stable empirical discrepancy distribution under the Null hypothesis, with which a p-value p can be computed:

$$p = \frac{1}{T} \sum_{i=1}^{T} \delta(D \leq D_i),$$

where D_i is the discrepancy i-th random permutation of the reads. For the pseudocode of the algorithm see Alg. 4.1.

Algorithm 4.1 MMD-test for read distributions

$S \leftarrow 0$
$D \leftarrow \|\text{coverage}(X^A) - \text{coverage}(X^B)\|_2$
$X \leftarrow X^A \cup X^B$
for $i \leftarrow 1, T$ **do**
  $X_p \leftarrow \text{permute}(X)$
  $X_p^A, X_p^B \leftarrow \text{split}(X_p, \text{size}(X^A), \text{size}(X^B))$
  $D' \leftarrow \|\text{coverage}(X_p^A) - \text{coverage}(X_p^B)\|_2$
  if $D \leq D'$ **then**
    $S \leftarrow S + 1$
  end if
end for
$p \leftarrow \frac{S}{T}$

Alternative Mean Embeddings

One of the key advantages of using MMD to test for differential RNA processing is that it provides flexibility in how to represent reads and how to define the similarity between them. Therefore, more expressive representations can be easily incorporated such as embeddings that contain information on where mismatches and introns are located or the insert size for paired-end reads. We propose the following kernels to exemplify how this can be done.

A kernel that leverages the splice information can be constructed in the following way. Assume that the observable junctions $(J_i)_{i\in\mathcal{L}}$ are enumerated by an index set $\mathcal{L}$. Then we can define an embedding $\phi_S : \mathcal{X} \to \mathbb{R}^{|\mathcal{L}|}$ by $\phi_S(r)_i = 1$ if read r supports junction J_i and 0 otherwise. This induces a positive-semi-definite kernel $k_S(r, r') :=< \phi_S(r), \phi_S(r') >$, where $< \cdot, \cdot >$ is the Euclidean scalar product on $\mathbb{R}^{|\mathcal{L}|}$. This kernel has the advantage over the kernel k presented before, that differential splicing events that cause only a small or no difference in coverage (e.g. slightly shifted acceptor sites) can be better detected.

If paired end reads are used, it is can be of advantage to consider the insert size information during testing (see Sec. 3.2.1) and therefore a kernel that accounts for this information is desirable. Assume that P is the distribution of the insert sizes as obtained for example from reads that map to genes with only one isoform or measured during library preparation. Then a kernel that compares the distance between reads can be obtained as follows. Denote the genomic distance between the read-ends of a read r by $d(r)$. Then the embedding $\phi_I : \mathcal{X} \to \mathbb{R}^{\mathbb{N}}$ of a read r can be defined as the convolution of the empirical insert size distribution P and the Dirac delta function $\delta(d(r))$: $\phi_I(r)_i = \delta(d(r)) * P$. The induced kernel $k(\cdot, \cdot)_I$ for this function is then given by $k_I(x, y) :=< \phi_I(r), \phi_I(r') >$, where $< \cdot, \cdot >$ is the Euclidean scalar product on $\mathbb{R}^{\mathbb{N}}$. This kernel then allows comparing the empirical insert size distribution from the samples, while accounting for the overall variability in the insert size.

Different kernels can also be easily combined as positive linear combinations of kernels form again a kernel. For this, we suggest to scale the kernels for each gene such that the empirical Null distributions of the kernels have the same variance and mean in order to make them comparable. The weighting of the individual contributions can be made prior to testing. For this we suggest a constant weighting of the kernels as results from our preliminary studies show that this provides better results than any using each kernel individually (see Sec. A.3.1).

4.2.3. Extensions

For the variance of gene expression σ^2_{total} it has been established that it is composed of two components [5, 66, 140, 141]:

$$\sigma^2_{total} = \sigma^2_{shotnoise} + \sigma^2_{biological},$$

the so called shot noise $\sigma^2_{shotnoise}$ that arises from the random sampling of the sequencing procedure. This noise typically has a Poisson linear mean-variance relationship. The second type of noise is the biological variation $\sigma^2_{biological}$. This over-dispersion, compared to a Poisson variance, arises from variation in transcript abundance before sequencing and therefore has a quadratic mean-variance relationship. This latter variation can be caused by cell-to-cell variation of transcript abundances but despite the name can be also due to changes in the experimental condition or the use of different barcodes. These two distinct types of noises are dominant in different regimes of gene expression. The technical variation is typically dominant for low counts, whereas the biological variation is the major source of variation for high read counts. Since the biological variance is a phenomenon that is not restricted to gene expression but also occurs for of transcript abundances, the read counts in alternative regions and the coverage can also be over-dispersed. Therefore, in order to account for overdispersion when testing for differential RNA processing it is important to estimate the variance.

The strategy to estimate the distinct variance components depends on the number of replicates at hand. In cases where a large number of replicates is available, such as the case for association studies, the variance can be estimated by the empirical variance within the sample. However, in typical controlled experiments the number of replicates is much smaller. In this case the empirical variance is not stable enough to estimate the variation. An alternative in this case is to pool the individual estimates from multiple genes in order to obtain stable variance estimates [5]. For this it is assumed that the variance σ^2_{total} of read counts is a function of the mean μ number of read counts. This allows to estimate the function $\sigma^2(\cdot)$ by fitting a local polynomial throughout the empirical mean expressions and variances $(\hat{\mu}_g, \hat{\sigma}_g)$ of all genes. This *variance function* $\sigma^2(\cdot)$ can then be used to predict a stable estimate of the gene expression variation for each gene based on the observed mean. In cases where no replicates are at hand a workaround is to consider the second sample as a replicate when estimating the variance function [5]. The assumption hereby is that the number of true differentially expressed genes is small and that the rest behaves as replicate data. Then the influence on the former genes on the variance estimation is small and the estimate variance is only a slight overestimation of the actual biological variance. This approach therefore leads to a conservative call off differential expression.

One of the crucial assumptions of rDiff.poisson and the MMD-test is that the biological variation (see Sec. 2.5.2) is small. If this assumption is not fulfilled, then these tests are oversensitive for genes that are high expressed as they underestimate their variance. Consequently, we propose two extensions of these tests that account for biological variance. These extensions work by first estimating the extra variation and then accounting for this during testing, thus providing better-calibrated test statistics.

We first show how the biological variance can be estimated and then present the two extensions. For this we follow the approach established by [5] for detection of differential gene expression.

Biological Variance Modelling

We propose to estimate the biological variance for each sample separately. In the following we will assume that G denotes the set of all genes and assume that the biological sample R for which the variance is to be estimated consists of a set of replicates $r \in R$. Furthermore, we will assume that for each of the alternative regions $j \in J_g$ of gene g, $c^r_{g,j}$ are the read counts and that N^r_g are estimations of the gene expressions. We then estimate a variance function $f(\mu) = \sigma^2_{total}(\mu)$, which describes the mean-variance relationship, using the replicate information in the following way: We first compute a normalising constant s^r_g to capture the variation caused by changes in gene expression and library size:

$$s^r_g := \frac{|R| N^r_g}{\sum_{r' \in R} N^{r'}_g}$$

As we would like to detect changes in the relative abundance of transcripts and not those that are due to a changed gene expression, we then use the derived normalising constants in order to compute normalised counts

$$\hat{c}^r_{g,j} := \frac{c^r_{g,j}}{s^r_g}$$

This provides counts that are comparable across replicates. Differing from the approach presented in [5], we do not need to correct for changes of the library size separately as these are modelled as changes in gene expression. With these normalised counts we then compute for each region $j \in J_g$ the empirical mean

$$\mu^R_{g,j} = \frac{1}{|R|} \sum_{r \in R} \hat{c}^r_{g,j}$$

and the empirical variance

$$\sigma^{2\,R}_{g,j} = \frac{1}{|R| - 1} \sum_{r \in R} (\hat{c}^r_{g,j} - \mu^R_{g,j})^2.$$

In the last step we perform a local regression on the tuples $((\mu^R_{g,j}, \sigma^{2\,R}_{g,j})_{j \in J_g})_{g \in G}$ in order to estimate the variance function (see Fig. 4.3 for example). For this we use the Locfit [104] package that is part of Chronux 2.00 (obtained from `http://chronux.org`), using local polynomials of degree two, Mallows's CP criterion for bandwidth selection and the and gamma distribution as local likelihood function.

rDiff.parametric

We will now present how rDiff.poisson can be extended to also account for additional variance. One limitation of the Poisson assumption on the read counts is that extra variance cannot be modelled. We therefore propose to model the read counts using a negative binomial distribution. This class of distributions can be seen as a generalisation of the Poisson distribution (see Sec. 2.5.2).

In the following we assume again that we have two samples $A = \{A_1, \ldots, A_u\}$ and $B = \{B_1, \ldots, B_v\}$, where u and v are the number of replicates in condition A and B, respectively. To simplify the notation we omit again the index for the gene g whenever possible.

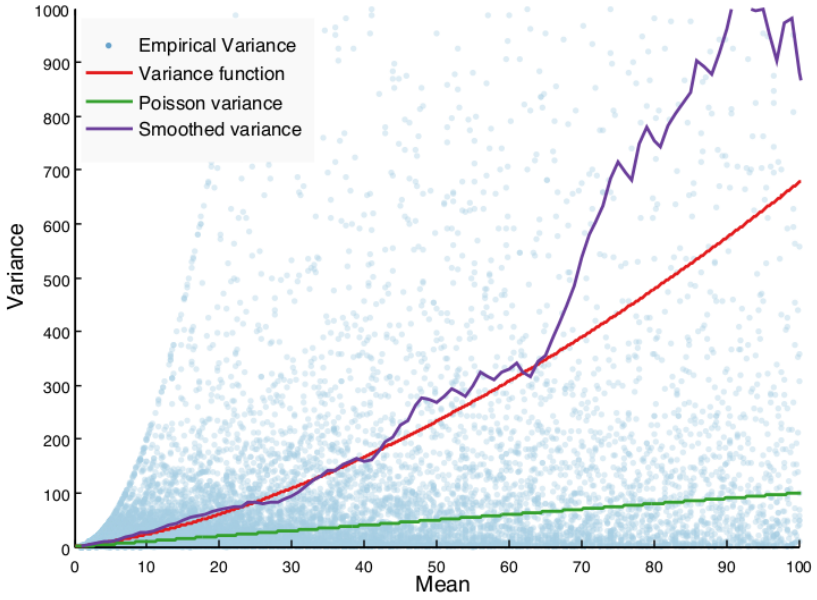

Figure 4.3.: Example of variance function fit. Shown are the observed mean variance tuples (light blue), the variance of the Poisson distribution (green), the sliding-average smoothed empirical variance (purple) based on a 5 bp sliding window and the fitted variance function (red).

Similar to the Poisson case we propose as Null hypothesis $\mathcal{H}_0$ that the two means of the distributions of counts μ_j^A and μ_j^B are the same, when accounting for differences in gene expression. We then can calculate the expected expression under the Null hypothesis by first averaging the normalised read counts in both samples:

$$q_j := \frac{1}{|A| + |B|} \sum_{r \in A \cup B} \frac{c_j^r}{N^r},$$

where N^r is the gene expression in replicate r and c_j^r is the number of reads mapping to region j in replicate r. Using this average we then calculated the expected number of counts μ_j^A we expect to see under $\mathcal{H}_0$:

$$\mu_j^A = \frac{q_j}{|A|} \sum_{r \in A} N^r$$

and μ_j^B analogously. These expected values then can be used to identify the probability distributions $\mathcal{NB}_{\mu_j^A, f_A(\mu_j^A)}$ and $\mathcal{NB}_{\mu_j^B, f_B(\mu_j^B)}$ from which the counts are drawn. For any pair of counts (k, l) we can therefore compute the probability of observing them:

$$p(k, l) = \mathcal{NB}_{\mu_j^A, f_A(\mu_j^A)}(k) \cdot \mathcal{NB}_{\mu_j^B, f_B(\mu_j^B)}(l)$$

This probability in turn, can be used to compute the p-value for the observed counts in region j. For this let $C_j^A = \lceil \frac{1}{|A|} \sum_{r \in A} c_j^r \rceil$ and $C_j^B = \lceil \frac{1}{|B|} \sum_{r \in B} c_j^r \rceil$ be the rounded up average number of observed reads in a region j. Denote furthermore the total read counts in region j as $C_j = C_j^A + C_j^B$. Then the p-value p_j of the observed counts C_j^A and C_j^B under the Null

hypothesis H_0 is given by:

$$p_j\left(C_j^A, C_j^B | H_0\right) = \frac{\sum_{k+l=C_j} \delta_{p(k,l) \leq p(C_j^A, C_j^B)} p(k,l)}{\sum_{k+l=C_j} p(k,l)}$$

where δ_T is an indicator function that is 1 if T is true and 0 otherwise. Here we condition again, as for the Poisson case, on the total number of counts observed C_j in order compute the p-value. Finally, we combined the p-values across regions into a p-value for the gene g of relative transcript abundance variability using the Bonferroni correction [25]:

$$p_g = |J_g| \min_{j \in J_g} p_j\left(C_j^A, C_j^B | H_0\right).$$

In this work we will refer to this testing approach as rDiff.parametric.

Alternatively to combining the test for all alternative regions, the information as to which specific testing region is differentially expressed can be used directly, which is similar as the approach taken in [6].

rDiff.mmd

The MMD test is well suited to detect differences in read distributions when the biological variation is minor. However, despite being nonparametric it still suffers from the same oversensitivity for highly expressed genes as rDiff.poisson. This is again because even between two identical cells there will be a variation and therefore the expectation discrepancy of the sample means of two finite samples is strictly larger than zero. However, as a consequence of the strong law of large number (e.g. [53]) the variation of the means that are drawn during bootstrapping will tend towards zeros as the number of reads increases. Hence, also the discrepancy between two random samples from the Null distribution will tend towards zeros as the number of reads increases. This means that the observed discrepancy will be almost certainly bigger than most of the discrepancies expected under the Null hypothesis if the number of reads is high. Thus, highly expressed genes will be prone to be detected as strongly significant independent of their true difference (for an illustration see Fig. 4.4). Therefore, accounting for extra variability during testing is crucial in order to obtain a well-calibrated statistical test.

To achieve this we propose to correct for the excess variation during the computation of the empirical Null distribution by sampling random samples with a realistic variance. As the variance of the mean is a function of the subsample size, the subsample size can be chosen such that the variance of the subsample equals the expected biological variance.

For this we first compute the variance of the subsample $\sigma^{2^r}_{\text{subsample}}$ as a function of the sample size n. As the drawing of a subsample is a drawing without replacement, the counts can be well described by a hypergeometric distribution (see Sec. 2.5.2). Therefore, when drawing a subsample of n reads from the total of N^r reads the distribution of the coverage at a position p follows a hypergeometric distribution $\mathcal{H}_N(N^r, n^r, C_p^r)$, where C_p^r is the fraction of reads covering the position p, N^r is the number of reads in the sample r and n^r is the size of a subsample. Consequently, the variance $\sigma^{2^r}_{\text{subsample}}$ of the coverage of a subsample of size n^r is given by:

$$\sigma^{2^r}_{\text{subsample}} = n^r \frac{C_p^r}{N^r} \frac{N^r - C_p^r}{N^r} \frac{N^r - n^r}{N^r - 1}$$

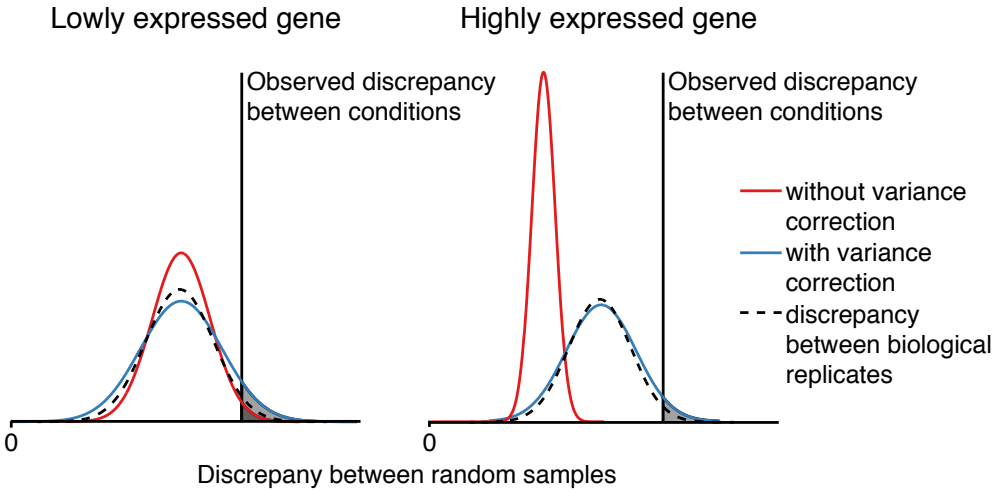

Figure 4.4.: Illustration of variance of the discrepancy between random samples from the Null distribution for lowly and highly expressed genes. The density of the discrepancy between two biological samples is shown as a dashed black curve, the one between two random samples when not correcting for biological variance in red and when correcting for biological variance in blue. The resulting p-value for rDiff.mmd corresponds to the area of the grey surface, which is the expected number of random samples that have a larger discrepancy than the observed discrepancy between the two conditions. The density difference between random samples converges to zero for highly expressed genes, thus leading to an unrealistically small P-value. Drawing samples while correcting for the biological variance leads to a better approximation and therefore to better p-value estimates.

Therefore, the variance of the read density is given by:

$$\begin{aligned}\sigma^{2^r}_{\text{subsample-density}} &= n^r \frac{\frac{C^r_p}{N^r}\frac{N^r - C^r_p}{N^r}\frac{N^r - n^r}{N^r - 1}}{(n^r)^2} \\ &= \frac{f_r(1-f_r)}{N^r - 1}\frac{N^r - n^r}{n^r},\end{aligned}$$

where $f_r := \frac{C^r_p}{N^r}$ is the fraction of reads covering position p.

After computing the dependence of the sub-sampling variance on the sample size as described above, we next determine the sample size for which the biological variance $\sigma^2_{\text{biological variance}}$, as estimated using a fitted variance function, equals the sub-sampling variance $\sigma^{2^r}_{\text{subsample}}$at position p. We therefore solve the following equation for n^r:

$$\begin{aligned}\sigma^{2^r}_{\text{biological variance}} &= \sigma^{2^r}_{\text{subsample-density}} \\ \frac{f(C^r_p)}{(N^r)^2} &= \frac{f_r(1-f_r)}{N^r - 1}\frac{N^r - n^r}{n^r}\end{aligned}$$

For simplification of the notation we further define $c^r := \frac{f_r(1-f_r)}{N^r - 1}$, which leads to the desired sample size n^r:

$$n^r = \frac{c^r f_r}{c^r + \frac{f(\text{median}_p(C^r_p))}{(N^r)^2}}$$

In order to match the variances at multiple levels of the total coverage $C = C_A + C_B$ we perform this matching in 10 equally sized bins defined by the position $b_j, j \in \{1, \ldots, 10\}$ where the coverage is in the same 10% quantile of positive coverage. This matching is performed for each of those bins and all samples r in order to obtain subsample rates n_j^r The new random sample from the Null distribution can then be computed as:

$$\mu_r = \sum_{j=1}^{10} \frac{\sum_p C_p|b_j}{\sum_p C_p} \frac{1}{n_j} \sum_{r=1}^{n_j^r} \phi(\mathbf{x}^r_{\sigma(r)}|b_j),$$

where σ is a permutation of $1, \ldots, N_A + N_B$. We refer to the resulting statistical test as rDiff.mmd. It should be noted that rDiff.mmd is a general test for differences in read distributions. It can therefore also be applied to other data than RNA-Seq and we show an example of the application of this test to ribosome footprinting data in Sec. 4.3.4.

rDiff.nonparametric

The power of Diff.mmd to detect differential RNA processing can further be increased by applying *contrasting*, an extension that is described in the following. This extension is motivated by the observation that the power of rDiff.mmd increases when only considering regions where the total coverage in all samples is less than half of the maximal total coverage in the gene. This counter-intuitive heuristic can be explained by the fact that regions that are maximally covered tend to be contained in all transcript, as otherwise regions that are contained in more transcripts would have a higher coverage, thus contradicting the assumption that the former regions where maximally covered. These regions tend to be uninformative, as established in Sec. 4.2.1, and therefore not considering them during testing decreases the noise. The risk in this strategy is however, that changes in highly covered regions cannot be detected as efficiently.

We therefore propose a heuristic to which we refer to as contrasting, where we exploit the aforementioned heuristic but also consider changes in highly covered regions, by performing a series of tests using rDiff.mmd on regions below an increasing threshold. Formally, this is done by first applying rDiff.mmd test on the 10% of the positions that have the lowest positive coverage, leading to a p-value p_{10}. Next, we repeat the same procedure on the lowest 20% of the positions that have the lowest positive coverage and so forth until we have 10 p-values, $p_{10}, \ldots, p_{100}$. These are then combined using the Bonferroni correction resulting in the final p-value. In the following we will refer to this extension of rDiff.mmd as rDiff.nonparametric.

When the ranking of genes is important we propose to take the lexicographic order of the sorted p-values $p_{10}, \ldots, p_{100}$ in order to resolve ties. This captures also the information contained in the other p-values besides the strongest one for tie breaking. Alternatively, also a small quantity $\frac{\max_{j=1,\ldots,10} p_{j*10}}{\text{number of permutations}+1}$ can be added to each p-value which has a similar effect in practice.

4.2.4. Biases in the Detection of Differential RNA Processing

A challenge in the analysis of RNA-Seq data is that there can be systematic differences in the probability of seeing a read from a fragment (biases) between libraries (e.g. [65, 100, 137]). This means that fragments generate reads at a different rate in different libraries. Therefore, transcripts that have the same abundance can have different numbers of reads, even after correcting for library size. There are two types of biases: The ones that cause differences in

comparisons between libraries (*inter-library biases*) and the ones that cause differences when comparing read counts in a library (*intra-library biases*). The latter biases are typically not of importance for detection of differential RNA processing, as they do not affect the Null hypothesis, i.e. a gene that is not differential will also not have different read distributions. However, inter-library biases can affect the Null hypothesis, as these biases can lead to changes in the read distributions between libraries even though the gene is not differential. These biases are typically caused by different barcodes or different experimental conditions but also by differing bioinformatics pre-processing steps (e.g. different alignment approaches [46]). To our knowledge there exist no approaches that can reliably model these biases and thus can be used to remove the contribution of biases to quantifications. Therefore, experiments should be designed and performed such that samples all libraries are treated in the same way in order to minimise inter-library biases (see Sec. 3.1 for a discussion of this subject).

4.2.5. Detection of Changes in RNA Secondary Structure

Aside from RNA-Seq there exist further sequencing based assays to study other aspects of the transcriptome. RNA secondary structure for example can be examined using sequencing protocols such as the *parallel analysis of RNA structure* (PARS) [84], *FragSeq* [175] or *ShapeSeq* [107]. For PARS this is achieved by splitting the RNA sample under investigation up into two samples. Next, the two resulting samples are treated with different structure specific enzymes; one is treated with a nuclease that preferentially cuts single stranded RNA (*S1 nuclease*) while the other one is treated with a nuclease (*RNase V1*) that preferentially cuts double stranded RNA. The resulting libraries then have fragments that start where the prevalent structure was single-stranded and double-stranded, respectively. These libraries can then be sequenced and aligned. Finally, the prevalent local secondary structure in the transcripts can be inferred by determining which of the two coverages is higher. As a quantitative measure of this prevalence the log ratio of the coverages has been proposed (*PARS score*) [84].

To detect changes in secondary structure prevalence between samples an extension of the PARS score, *StrucDiff*, has been suggested [180]. If V_i^A and V_i^B are the two double-strand library size normalised coverages at a position i and S_i^A and S_i^B are the single stranded ones of two conditions A and B then this score in its general form is given by:

$$\text{StrucDiff} := \frac{1}{n}\sum_{i=1}^{n}\left|\log_2\frac{V_i^A+5}{S_i^A+5} - \log_2\frac{V_i^B+5}{S_i^B+5}\right|,$$

where n is the length of the transcript. This score was used by [180] in order to detect the most differential region by choosing $n = 5$. One of its advantages is that it is very efficient in computation. However, major disadvantages of it are that it neither does provide a probabilistic interpretation nor does it account for the discrete nature of the reads.

We, therefore, propose an adaptation of the MMD-test to robustly detect changes in secondary structure that does not suffer from disadvantages of StrucDiff. Assume for this that V^A and V^B are the read densities for the V1 nuclease and S^A and S^B the ones for the S1 nuclease for condition A and B respectively. Furthermore, let $\mu_{V^A}, \mu_{V^B}, \mu_{S^A}$ and μ_{S^B} be their respective RKHS embeddings. We then propose to measure the difference in structure using the following measure (*sDiff*):

$$\text{sDiff}(\mu_{V^A}, \mu_{V^B}, \mu_{S^A}, \mu_{S^B}) := \|(\mu_{V^A} - \mu_{S^A}) - (\mu_{V^B} - \mu_{S^B})\|_{\mathcal{H}} \tag{4.1}$$

This measure has, similar to StrucDiff, the property that changes in the read distributions are not considered as long as it is the same in both libraries, thus making it more robust than

simply testing for changes in any of the four distributions. The difference to StrucDiff is that changes that have the same difference aren't considered whereas for StrucDiff changes that have the same ratio aren't considered.

Similar to MMD, sDiff can be computed using a kernel expansion (see Lemma A1 for a proof):

$$\begin{aligned}\mathrm{sDiff}(\mu_{V^A}, \mu_{V^B}, \mu_{S^A}, \mu_{S^B})^2 &= \|(\mu_{V^A} - \mu_{S^A}) - (\mu_{V^B} - \mu_{S^B})\|^2_{\mathcal{H}} \\ &= \mathbb{E}_{v^A, v'^A \sim V^A,} k(v^A, v'^A) - 2\,\mathbb{E}_{v^A \sim V^A,\, s^A \sim S^A} k(v^A, s^A) \\ &\quad + \mathbb{E}_{s^A, s'^A \sim S^A} k(s^A, s'^A) + \mathbb{E}_{v^B, v'^B \sim V^B} k(v^B, v'^B) \\ &\quad - 2\,\mathbb{E}_{v^B \sim V^B,\, s^B \sim S^B} k(v^B, s^B) + \mathbb{E}_{s^B, s'^B \sim S^B} k(s^B, s'^B) \\ &\quad + 2[\mathbb{E}_{v^A \sim V^A,\, s^B \sim S^B} k(v^A, s^B) - \mathbb{E}_{v^A \sim V^A,\, v^B \sim V^B} k(v^A, v^B) \\ &\quad\quad - \mathbb{E}_{s^A \sim S^A,\, s^B \sim S^B} k(s^A, s^B) + \mathbb{E}_{v^B \sim V^B,\, s^A \sim S^A} k(v^B, s^A)],\end{aligned}$$

where we denote by $x, x' \sim X$ that x and x' are two independent identical distributed random variables that are distributed as X. For this test statistic an estimator $\widehat{\mathrm{sDiff}}$ can be obtained analogously as for the MMD estimator (see Lemma A2). To compute the significance for a test statistic we propose to use bootstrapping. However, differing from the computation of the MMD p-value, we suggest permuting reads only between the samples of the same nucleases as the Null hypothesis is that the distributions for each of the nucleases are the same and not that all distributions have to be identical. This is thereby a statistically robust and well calibrated estimation. In order to account for biological variance we suggest following the approach outlined for rDiff.mmd and to estimate a variance function for each sample and each nucleases. Realistic samples for the Null distribution can then be computed in the same manner as described for rDiff.mmd. In the following we will refer to this approach as sDiff.

4.2.6. Association of Changes in RNA Processing

As already mentioned before, estimating the biological variance by sharing information across genes allows stable estimation of the biological variance when the number of replicate is small. In cases where many samples that are at hand, such as it is often the case in association studies, alternative approaches are possible. In this case the variance estimate of each gene is sufficiently stable and thus sharing information across genes is not necessary. This allows to more accurately account for the individual variances of genes during testing. We here show, how rDiff.mmd can be generalised naturally to incorporate information from a large number of replicates to estimate biological variance for each gene individually. For this, we propose to first restate the problem setting and then derive a model of the data generation for which we finally outline an RKHS embedding. To simplify the notation we omit as before the index for the gene.

In the following, we assume that we have two populations of cells or individuals X and Y. From these populations m respectively n samples are drawn that are subsequently sequenced. Our aim is then to determine based on the reads in all libraries whether the RNA processing in the two populations was different.

If we assume that RNA processing of a cell can be parametrised by a parameter $\theta \in \Theta$ (representing for example transcript abundances and other cellular characteristics) and that P and Q are the two probability distributions on Θ that describe the two populations X and Y, then data generation procedure can be formalised by the following *generative process* (see Fig. 4.5):

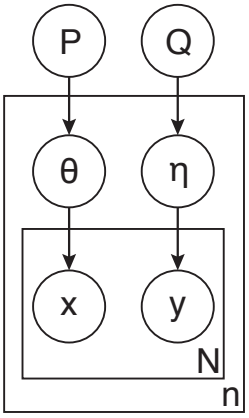

Figure 4.5.: Plate diagram of generative process.

1. From the population P and Q the parameters $\theta_1, \ldots, \theta_n \sim P$ and $\eta_1, \ldots, \eta_n \sim Q$ are drawn. For simplicity, we assume that the number of parameters that are drawn are the same for both populations. These parameters correspond to the parameters of the cells in the samples and are not observable.
2. For each of the sampled parameters, N observations (reads) are drawn (sequenced) to create the sets of observations $x_1^i, \ldots, x_N^i$ and $y_1^j, \ldots, y_N^j$, for $i, j \in \{1, \ldots, n\}$. We again assume for simplicity that a constant number of N reads is drawn for all parameters and that the sequencing is described by the sampling from a distribution $S(\cdot|\theta)$ parametrised by $\theta \in \Theta$.

We can then define an embedding in an RKHS for the model that we have defined above in the following manner: Assume that $\mathcal{H}$ is a universal kernel reproducing Hilbert space that is induced by the kernel k. Assume furthermore that μ_p and μ_q are the kernel mean embeddings of the marginal distributions of the two populations $p(x) = \int S(x|\theta)P(\theta)\,d\theta$ and $q(x) = \int S(x|\theta)Q(\theta)\,d\theta$. The problem is then to test whether $P = Q$, based on the read samples $x_1^i, \ldots, x_m^i$ and $y^j, \ldots, y_n^j$, $i, j \in \{1, \ldots, n\}$. Unfortunately, this is not possible in general as it could be that two identical mean embeddings stem from different parameter distributions. In this case we can only test for identity of the marginal distributions. This can be the case if the sequencing distribution does not capture all the properties that are determined by the parameter space, e.g. some transcripts are filtered out during the fragment length filtering or the solution to the transcript abundance estimation has multiple solutions. Therefore, we can only test for observable changes in the reads distribution. If, however, the mapping from the marginal distribution to the population densities is injective, then testing for identity of the mean embeddings is a test for the identity of the populations. We therefore have that $p = q$ if their maximum mean discrepancy is 0.

If we write the mean embeddings as expected values of the conditional mean embeddings μ_θ for the parameter θ, then we have:

$$\begin{aligned}
\mathrm{MMD}[\mathcal{H}, p, q]^2 &:= \|\mu_p - \mu_q\|_{\mathcal{H}}^2 \\
&= \|\mathbf{E}_\theta \mu_\theta - \mathbf{E}_\eta \mu_\eta\|_{\mathcal{H}}^2 \\
&= \mathbf{E}_{\theta,\eta}\left[\langle \mu_\theta, \mu_\theta\rangle_{\mathcal{H}} + \langle \mu_\eta, \mu_\eta\rangle_{\mathcal{H}} - 2\langle \mu_\theta, \mu_\eta\rangle_{\mathcal{H}}\right] \\
&= \mathbf{E}_{\theta,\eta}\left[\mathop{\mathbf{E}}_{x'|\theta} k(x, x') + \mathop{\mathbf{E}}_{y'|\eta} k(y, y') - 2\mathop{\mathbf{E}}_{y|\eta} k(x, y)\right]
\end{aligned} \tag{4.2}$$

The inner part of the bracket is also a valid kernel $h(\theta, \eta)$ as shown in [26]. For the inner kernel the estimator presented in Sec. 4.2.2 can be used but also a linear time estimator $\hat{h}(\theta, \eta)$ can be derived [62]:

$$\hat{h}(\theta, \eta) := \frac{1}{n_2}\left[\sum_{i=1}^{n_2} k(x_{2i-1}, x_{2i}) + \sum_{i=1}^{n_2} k(y_{2i-1}, y_{2i}) - \sum_{i=1}^{n_2} k(x_{2i-1}, y_{2i}) - \sum_{i=1}^{n_2} k(x_{2i}, y_{2i-1})\right],$$

where $n_2 := \lfloor \frac{N}{2} \rfloor$ and x_i and y_j are the samples from θ respectively η. This estimator can be computed significantly faster than the originally proposed one and has a comparable performance if the number of samples is large [62]. In the setting where many replicates are available this can be of advantage as the runtime is only linear in the number of reads. We therefore suggest using this estimator, when kernels other than the linear kernel are used. The outer expectation of Eq. 4.2 can then be estimated as:

$$\widehat{\mathrm{MMD}}[\mathcal{H}, p, q]^2 := \frac{1}{n^2} \sum_{i,j=1}^{n} \hat{h}(\theta_i, \eta_j) \tag{4.3}$$

in order to derive a test statistic. The significance for the observed test statistic can be computed using a two step bootstrapping approach, wherein the first bootstrapping step is performed for the reads of pairs of replicates and then second is performed for the outer expectation. In cases where time efficiency is of importance and the number of samples is large, the Null distribution can be approximated by performing only the outer bootstrapping. For the application in association studies a variation of the bootstrapping approach can be employed to account for confounding effects such as population structure or gene expression. This consists of not randomly assigning the replicates to the samples but to assign them according to randomly selected variants, i.e. to assign all replicates to the sample that have the same allele at a given randomly selected variant.

Preliminary experiments have shown that differential gene expression is a major confounding factor when testing for differential RNA processing in an association study setting (data not shown). We therefore suggest to account for confounding gene expression by matching the gene expression between the replicates comparable. This can be done by subsampling reads from replicates where the gene expression is high in order to obtain samples with a similar number of reads per gene in all libraries. For this number, we suggest the minimal median gene expression in both samples. In the following, we will refer to the generalisation of rDiff.mmd that is outlined above as rDiff.gmmd.

4.2.7. Data Simulation

A systematic evaluation of methods to detect differential RNA processing requires a large number of genes that have been experimentally validated. The state-of-the-art validation technique for this is RT-qPCR. This technique, however, only allows validating a few dozens of genes with reasonable resources, which falls short of the requirements for robust evaluation. In order to assess our methods on a dataset for which the ground truth is known, we therefore simulated data. We designed the simulation strategy such that many key properties of realistic datasets were reflected. The insights we obtained are therefore transferable to experimental datasets for which no ground truth is known.

Differential Relative Transcript Abundance

To assess methods to detect differential alternative splicing we simulated reads. We did this for all 5,875 mRNA coding genes with multiple annotated isoforms of *A. thaliana* using the TAIR10 genome annotation. In order to evaluate the performance for different strengths of biological variation we simulated two datasets: One for low and the other one for a large biological variance. For each of these datasets, we simulated two samples (referred to as A and B in the following) consisting of two replicates. For a realistic simulation of the datasets we proceeded in the following way:

First, we measured gene expression and transcript variability on a real dataset that we have generated (see Sec. 4.2.8). For gene expression, this was done by counting the number of reads in the non-informative regions of all genes. We obtained two estimates, one from seedlings grown with 0 h light exposition and the other from seedlings grown with 1 h light exposition. The transcript variability was estimated from the read counts in alternative regions. In order to study the effect of different strengths of biological variance on the performance of methods to detect differential RNA processing we obtained two variability estimates: One for when biological variability is small such as for individuals grown in a laboratory where the environment and the sample preparation time point is controlled. For this we used the two seedlings grown with 0 h light exposure. The second estimate was obtained from a sample with 0 h light exposure and one from a sample that was exposed to light for 1 h. These were considered as replicates in order to mimic uncontrolled environments and collection, such as it is the case for studies involving *H. sapiens*.

Second, we determined the transcript abundances for both samples of the two settings as follows. We began by choosing the pairs of gene expression for both samples from the empirical distributions. We then sampled for each gene g and each transcript $t_j \in \{t_1, \ldots, t_{T_g}\}$, where T_g is the number of transcripts of gene g, relative transcript abundances $e \in [0, 1]$ from a uniform distribution and normalised them to sum to one. These relative transcript abundances provided the basis for the generation of the data for two samples. For half of the genes, we then perturbed the relative transcript abundance. This was done by first choosing for both samples a vector $v_j^{A,B} \in [-0.5, 0.5]^k$ that determined the directions of change and the strength of the change $c_j \in [0, 1]$. Both the strength and the change vector were drawn from uniform distributions in the respective spaces. For the sample A we perturbed e_j by adding cv_j^A to it and for the gene in sample B by adding cv_j^B to it. If any e_i^j was negative we set it to zero and if all e_i^j were negative we repeated the procedure above until we obtained valid relative transcript abundances. The final relative transcript abundances for the samples were obtained by again normalising the transcript abundances to sum to one. From those relative abundances we calculated the mean transcript abundances by multiplying the relative transcript abundances with the gene expressions.

Third, we sampled the transcript abundances for each replicate. This was done such that the resulting read counts followed a negative binomial distribution that was in accordance with the estimated biological variances. This was achieved by first sampling the transcript abundances from the gamma distribution:

$$\Gamma_{\frac{e_j^2}{f(e_j)-e_j}, \frac{f(e_j)-e_j}{e_j}}$$

These abundances were then used to simulate the Poisson noise generating read simulation thus leading to negative binomial distributed reads [23]. The read generation was performed with FluxSimulator [63] (build 20100611), a tool that simulates all the sample preparation and sequencing steps, thus providing realistic simulated dataset. For the read simulation, we used the default parameters to simulate 26 million reads of length 70 bp per replicate.

Differential Relative Transcript Abundance for Association

To simulate RNA-Seq data for an association testing setting, we followed an approach similar as in Sec. 4.2.7. First, we randomly selected 500 genes together with their transcript abundance from the genes used for the previous simulation (see Sec. 4.2.7), such that 250 genes

were differential. These genes were subsequently used to simulate the replicates for each sample. For this we sampled for each of the two conditions 100 replicates. Differing from the strategy described above we used for each gene g a different variance function for the variance of the transcript abundance:

$$\sigma = \mu + a_g \mu^2,$$

where a_g was sampled for each gene g from the uniform distribution on $[0, 0.1]$. This generalisation poses fewer assumptions on the biological variance of genes than the variance model that was previously assumed (see Sec. 4.2.3). It allows genes to have distinct levels of stability in expression that are independent of the strength of expression. We finally simulated reads using FluxSimulator [63] (build 20100611) to generate 1 million reads of length 70 bp per replicate.

Differential Secondary Structure Prevalence

As for the assessment of differential alternative splicing events to our knowledge there does not exist a gold-standard dataset to evaluate detection of differential secondary structures from high-throughput sequencing data. We therefore simulated a dataset that allows this analysis. For this dataset we assumed that the biological variance is negligible and that the nucleases have no sequence specificity. We generated data for 1000 randomly selected genes from the TAIR10 genome annotation of *A. thaliana*. For each of these genes, we predicted for one of its transcripts the ten secondary structures with the lowest free energy. This was done using the RNAfold package v.1.6 [106]. The base abundance of the ten distinct secondary structures was then sampled from the uniform distribution between 0 and $1,000$. To simulate the cleavage by the single-strand-cleaving nuclease we sampled for each position in the transcript the read-starts for each secondary structure from a Poisson distribution. For this, we used a Poisson distribution that had intensity 0 if the respective position of the transcript was double-stranded and otherwise had an intensity of the base abundance of the structure divided by the transcript length. We simulated the read-starts for the double-strand-cleaving nuclease analogously. Finally, we joined the read-starts for all secondary structures of a gene and computed reads of length 30 from the read-starts.

4.2.8. Preparation of Sequencing Data

To demonstrate that our methods are also generally applicable in practical situations and provide novel insights, we applied them to real datasets from different species. These were obtained and prepared as described below.

A. thaliana

We used libraries that were generated from *A. thaliana* seedlings that were grown in darkness and then exposed to white light for 0, 1 and 6 hours. We furthermore used an additional library from a *cry1cry2* light receptor knockdown seedling grown under the same conditions as the 0 h wild type (wt) seedling. These libraries were sequenced with Illumina GAIIx platform that provided per lane on average $\sim 3.9 \times 10^7$ reads of length 80 bp. The libraries were then aligned to the *A. thaliana* genome using the TAIR10 genome annotation. For details we refer to our publication [42].

D. melanogaster

We furthermore used an existing dataset [28] from *D. melanogaster* consisting of two samples, one wild type and the other from Pasilla knockdown mutants, each containing two paired-end libraries and a single-end library. We downloaded the paired-end read libraries (GSM461177, GSM461178, GSM461180, GSM461181) from the NCBI Gene Expression Omnibus. In order to get a small variance in the read counts, we refrained from using the single end-library. Before aligning the reads we trimmed them down from the end to have a common length of 36. We then aligned the reads using TopHat v.1.3.1 [172] and the following parameters:

```
-segment-length 18
-max-insertion-length 0
-max-deletion-length 0
-g 10
```

For this we used Flybase, r5.22 genome annotation. After the alignment we treated the both ends of the read-pairs as independent single-end reads for a simplification of the analysis.

H. sapiens

We also used a third dataset consisting of two samples. A sample from *H. sapiens* mesenchymal stem cells and a sample from patient derived Ewing's sarcoma cells (pers. comm. Ahmet Zehir). Each sample consisted of three replicates. For these replicates $\sim 4.0 \times 10^7$ paired-end reads of length 50 bp per read-end were generated. We performed a variant aware alignment for the reads using PALMapper [75] against the *human* hg19 genome and allowed for at most 1 mismatch. Finally, we filtered out reads that mapped optimally to more than one locus. During the alignment we allowed for at most one mismatch and one insertion or deletion (*indel*). Doing this, we obtained between 3.0×10^7 reads and 3.8×10^7 read-pairs.

4.2.9. Application of Methods

Unless not mentioned otherwise, we applied all methods using their default parameters.

Application to simulated data

We applied MISO [83], CuffDiff [173] and our methods to the simulated and experimental *A. thaliana* datasets described above as follows.

rDiff.parametric We used all reads that were in concordance with the gene annotation for differential testing. We estimated the variance function on the counts in the alternative regions. For the analysis of the *A. thaliana* dataset we estimated one variance function using the two samples from 0 h and used this variance function for all samples.

rDiff.poisson We applied rDiff.poisson in the same way as rDiff.parametric except that the replicates were merged instead of used separately.

rDiff.nonparametric We estimated the variance function for rDiff.nonparametric by considering each nucleotide as an alternative region and estimating the gene expression using all reads that mapped to the gene. To speed up the computation on the real data we used at

most 10,000 reads per gene. If more were present we sub-sampled down to that number. For the computation of the p-values we performed 1,000 permutations. Furthermore, we added to each p-value a small quantity $\frac{\max_{j=1,\ldots,10} p_j}{\text{number of permutations}+1}$ in order to resolve ties for genes that have the same p-value. This value is always smaller then the absolute difference between two of the untied p-values and therefore does not affect the ranking of genes that have different p-values

rDiff.mmd We applied rDiff.mmd in the same way as rDiff.nonparametric.

CuffDiff For all experiments we used CuffDiff from cufflinks-1.3.0 for differential testing. Contrary to the observation in [55] that version 0.9.3 performed better than version 1.3.0 in identifying differential transcript expression, we found that version 0.9.3 performed worse than version 1.3.0 for identifying differential relative transcript expression. The experiments were carried out using the default parameters except for the following ones:

```
-num-bootstrap-samples 200
-num-importance-samples 10000
-max-mle-iterations 50000
```

For these, we increased the default values to get better estimated for the p-values. The resulting p-values of the computation for each transcript were then combined using Bonferroni's correction to obtain a p-value per gene.

MISO We used the MISO [83] package that we downloaded from the MISO website on 8/6/2011. For all our experiments we used the default parameters. The ranking of the genes was computed with the Bayes factor as ranking criterion. As MISO cannot account for replicates we merged all the replicates into one sample.

rDiff.gmmd For computation of the significance we performed bootstrapping only for the outer expectation (see Sec. 4.2.6). For this bootstrapping we performed 10,000 permutations. In order to reduce the confounding effect of gene expression, we sampled in each replicates the number of reads down to the minimal median gene expression in both samples.

sDiff For the computation of the p-values we performed 1,000 permutations.

Detecting Differential Translation

rDiff.mmd Instead of the default 1,000 permutations during bootstrapping we performed 10,000 permutations to increase the detection power.

4.3. Results and Discussion

4.3.1. Detection of Differential Alternative Splicing

Evaluation on Synthetic Data

We assessed the performance of our methods on the two simulated datasets with distinct strengths of biological variance (see Sec. 4.2.7). Furthermore, we also compared our methods

to *MISO* [83] and *CuffDiff* [173], two state-of-the-art quantification-based methods. For a detailed description how the methods were applied see Sec. 4.2.9. We did not include *FDM* [157] in this comparison as preliminary results (data not shown) showed that it performed considerably worse than all other methods.

In a first experiment we investigated the performance of the different methods to detect differential RNA processing. We computed for each method the *receiver operator characteristic* (ROC) using the p-values respectively the Bayes factor for MISO, as ranking criterion. This curve shows the *true positive rate* (TPR) for given *false positive rates* (FPR). As typically the predictions for a low false positive rate (the genes where predictions are reliable), are of interest in experiments, we compared the TPRs for genes with an FPR smaller than 0.2 (see Fig. 4.6). We quantified the performance using the *area under the ROC* curve (auROC) between 0 and 0.2, which will be denoted by auROC20 in the remainder of this work.

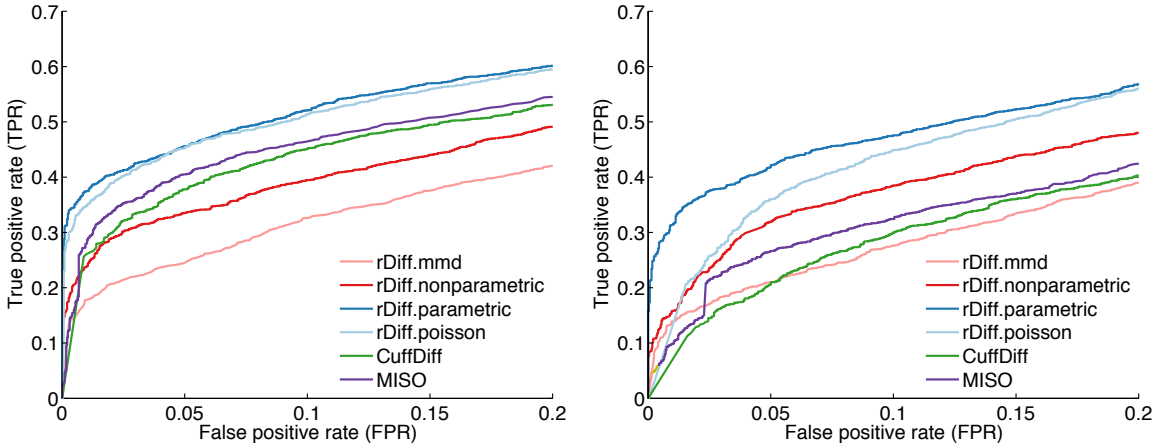

Figure 4.6.: ROC curves for rDiff, CuffDiff and Miso. Shown on the left are the ROC curves in the interval $[0, 0.2]$ for the simulated dataset with small biological variance and on the right for the simulated dataset with the larger biological variance.

These resulting auROC20 values are shown in Tab. 4.1. Overall, we found that the performances of all methods decreased dramatically for the dataset with the bigger biological variance compared to those with the small biological variance. This was expected as the biological variance induces noise and therefore changes are harder to detect when the biological variance increases. In the dataset with the smaller variance we found that the best two methods were rDiff.parametric and, with a comparable performance, rDiff.poisson, followed by MISO, CuffDiff and rDiff.nonparametric. On this dataset rDiff.mmd performed considerably worse than the other methods, showing the effectiveness of contrasting for detection of differentially processed RNA. In the second dataset rDiff.parametric showed the best performance. However, in this dataset the performance of rDiff.poisson was, although being still second best, inferior to the one of rDiff.parametric. Differing, from the previous dataset the third best method was rDiff.nonparametric, which performed better than the quantification-based methods when the biological variance was large. The next best method was MISO followed by rDiff.mmd and CuffDiff. In summary, we saw a decrease in performance of the methods that cannot account for biological variance, especially for the most significant predictions, reflected by the lower TPRs for small FPRs. This highlights the importance to account for biological variance during testing. To our surprise, also CuffDiff appeared to suffer more strongly from the increased biological variance than MISO even though accounting for it. We believe that this is due to unsatisfied modelling assumptions in the estimation of the variance.

The comparison on the two datasets showed that testing without prior quantification is indeed a promising approach as on both datasets rDiff.parametric outperformed all other methods. Furthermore, the performance of rDiff.nonparametric was comparable to MISO and CuffDiff even though not needing the gene annotation. This observation also holds when considering all FPRs (see Tab. A.1 and Fig. A.1), showing that testing without gene annotation is feasible. We also excluded the possibility that the difference in performance is due to differential gene expression. For this we sampled for each gene the same number of reads in each replicate, thus removing differential gene expression. This, however, did not change the qualitative performance (data not shown).

Table 4.1.: Area under the ROC-curve in the interval $[0, 0.2]$ (auROC20) for rDiff, CuffDiff and MISO. The comparison is shown on the two simulated datasets with small and large biological variance (see Sec. 4.2.7).

Method	auROC20	
	small biological variance	large biological variance
rDiff.mmd	0.062	0.054
rDiff.nonparametric	0.077	0.073
rDiff.parametric	0.101	0.093
rDiff.poisson	0.099	0.082
CuffDiff	0.085	0.055
MISO	0.089	0.061

A second, often neglected aspect of the performance of statistical tests is their *calibration*. This describes whether the predicted significance of a test is reflecting the true (usually unknown) significance. To study the calibration we, therefore, computed the false discovery rate (FDR) and compared it to the empirical FDR (see Fig. 4.7). We computed the FDR as described in [165] and the empirical FDR as the fraction of genes below a certain threshold that are false positives. As MISO only provides the Bayes factor and no p-value, we could not compute the FDR for MISO. When comparing the FDR and the empirical FDR of the methods, we observed that the distance to the diagonal was much lower for high biological variance than for a low biological variance. For the low variance dataset we found that rDiff.mmd was closest to the diagonal, indicating a very good calibration. The next two methods closest to the diagonal were rDiff.parametric, being too conservative, and rDiff.nonparametric that was too optimistic. The two remaining methods were both overly optimistic with CuffDiff having an empirical FDR of 0.08 while in the limit of the true FDR going to 0. This again indicates a tendency for false positives. For the high variance dataset rDiff.parametric exhibited the best calibration, even though being to conservative. Except for rDiff.nonparametric, the empirical FDR did not converge to 0 as the FDR did. On this dataset, the empirical FDR of CuffDiff and rDiff.poisson both converged to about 0.2 as the FDR decreased. This shows that even under the most significant genes predicted by these methods a large fraction are false positives, when the biological variance is large. Overall, these results shows that our methods were better calibrated than existing methods. The calibration of all methods, however, has still room for improvement and further research on the calibration of tests is needed.

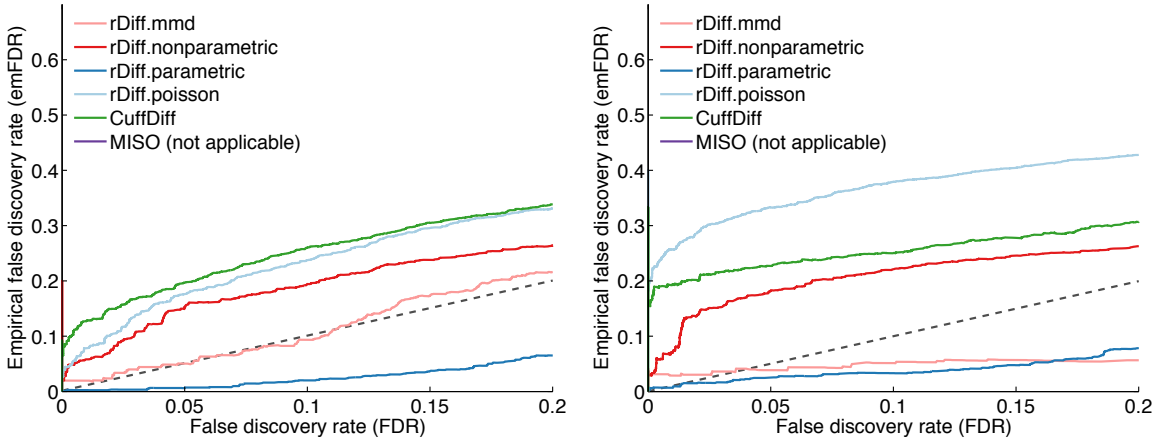

Figure 4.7.: Calibration of rDiff, CuffDiff and MISO. Shown on the left is a comparison of the empirical FDR with the predicted FDR in the interval $[0, 0.2]$ for the dataset with small biological variance. On the right this is shown for the dataset with the larger biological variance. The grey dashed line indicates the diagonal.

Evaluation on Experimental Data

To show that our proposed methods also perform well on experimental data, we evaluated their performance to detect differential splicing on a dataset from *A. thaliana* and one from *D. melanogaster*.

Evaluation on *A. thaliana* We first applied rDiff.parametric, rDiff.nonparametric as well as CuffDiff to a dataset from *A. thaliana* (see Sec. 4.2.8). Briefly, this dataset consisted of samples derived from *A. thaliana* seedlings that were grown in darkness and were exposed to 0 h, 1 h resp. 6 h of light. For the sample with 0 h light exposure a replicate was available. Preliminary analyses indicated that rDiff.poisson showed a strong oversensitivity for highly expressed genes as shown in Fig. 4.8. We therefore excluded it from the remaining analyses. We also did not consider rDiff.mmd since rDiff.nonparametric showed in the analysis of the artificial data to have a much better performance for detection of differential splicing (see Sec. 4.3.1).

In a first experiment, we assessed how the predicted significance of the events relates to the actual relative fold change (see Fig. 4.9). For this, we measured the relative fold change between conditions using RT-qPCR. We did this for 5 randomly chosen genes that were predicted to be significant from rDiff.nonparametric between the samples for all three time points (for details on the procedure we refer to our publication [42]). We then computed the correlation ρ for the log p-values and the fold change. In order to obtain a comparison that is robust to extreme values and monotone transformations we used Spearman's rank correlation. We found that rDiff.parametric had the highest correlation with the determined strength of change ($\rho = 0.84$) and that both rDiff.nonparametric and CuffDiff had a similar correlation of $\rho = 0.66$ resp. $\rho = 0.68$. This was well in line with the observation on the synthetic dataset that rDiff.parametric is the most accurate method in detecting differential alternative splicing and that rDiff.nonparametric and CuffDiff have a comparable similar performance. This results also show that the genes detected by our methods are indeed differentially spliced.

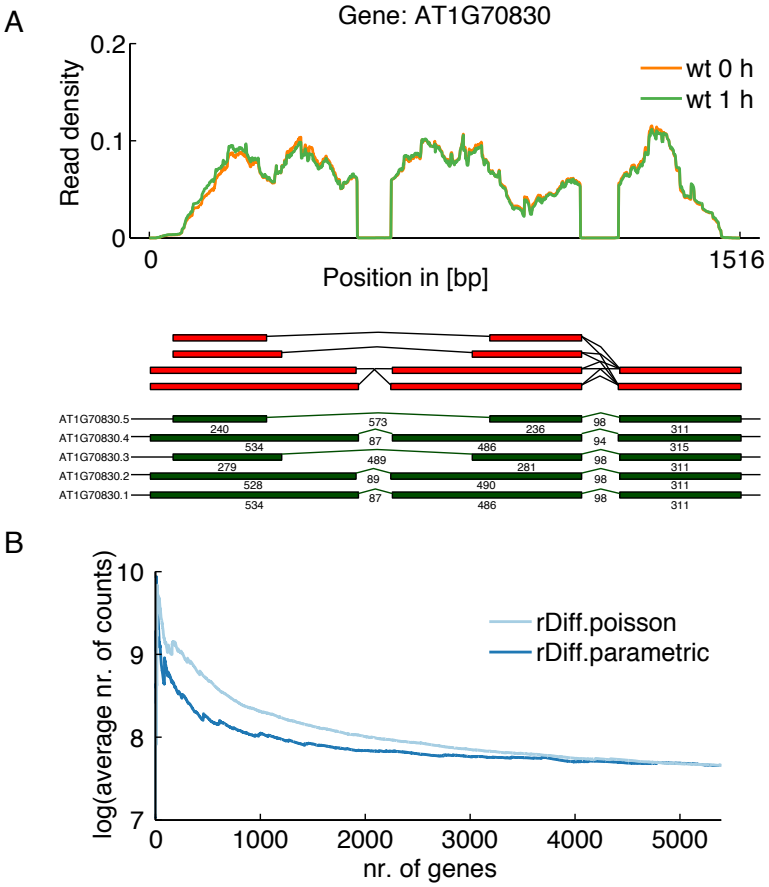

Figure 4.8.: Shown on top (**A**) is the read density for the gene AT1G70830 in 0 h (orange) and 1 h (light green) together with its splice graph (red) and transcripts (dark green). This gene is predicted to be highly significant between 0 h and 1 h by rDiff.poisson ($p \leq 2.67 \times 10^{-7}$) but not by rDiff.parametric ($p \leq 0.897$). Shown below (**B**) is the log-mean expression of the top genes with the lowest p-value for rDiff.poisson (light blue) and rDiff.parametric (dark blue) between wt 0 h and wt 6 h. The x-axis shows the number of genes that were used to compute the log-mean.

In a second experiment, we investigated how many genes were detected as differentially processed between the three conditions using rDiff.parametric and rDiff.nonparametric. We furthermore examined the overlap between the significant genes found by both methods (see Tab. 4.2). For this, we computed from the p-values of both methods the FDR as described in [165] and called genes significant differential if their FDR was smaller than 0.1. We observed that both methods found most changes between 0 h and 6 h of light exposure, which is expected, since the difference in experimental conditions were largest for this comparison. Furthermore, we observed that rDiff.nonparametric detected substantially more genes than rDiff.parametric but that the overlap between their predictions was modest. For the genes detected by rDiff.parametric, this can be explained as rDiff.parametric is more powerful than the nonparametric test and therefore, more of the events that are tested can be detected. This was confirmed by the lower p-values of rDiff.parametric compared to rDiff.nonparametric for the validated genes (see Fig. 4.9). Further, in order to understand, why there were many genes detected only using rDiff.nonparametric, we examined these. We found that $\sim 60\%$ of

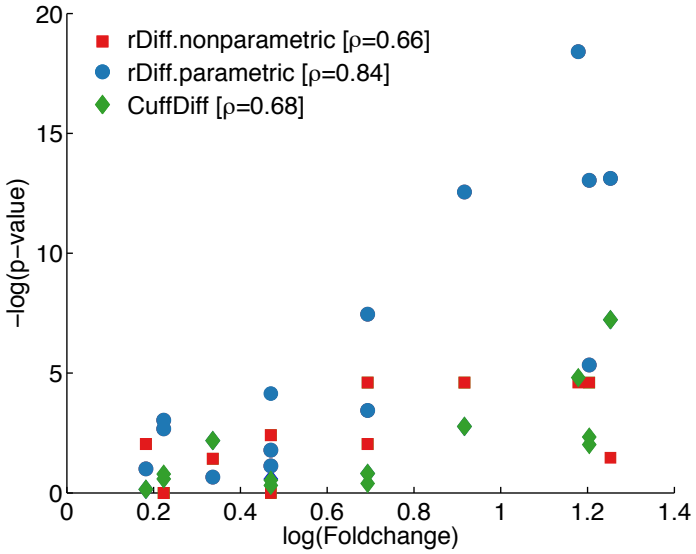

Figure 4.9.: Plot of the estimated -log p-values against the measured log fold change for rDiff.parametric (blue), rDiff.nonparametric (red) and CuffDiff (green). The Spearman's correlation coefficient ρ for the methods is given in the legend.

the genes had only one annotated isoform, which explains a large fraction of the difference between the numbers of detected genes. Visual inspection of these events suggested that they are truly unannotated events (see Fig. 4.11 for examples) that are not tested for by methods that rely on the gene annotation. This underlines the value of rDiff.nonparametric to get an unbiased view on changes in alternative splicing and in situations were the gene annotation is incomplete.

Table 4.2.: Overlap in detected genes (FDR ≤ 0.1) between methods for 0 h vs. 1 h / 0 h vs. 6 h/ 1 h vs. 6 h. The events written in bold are the number of events predicted by the respective methods.

Method	rDiff.parametric	rDiff.nonparametric
rDiff.parametric	**39 / 80 / 54**	
rDiff.nonparametric	18 / 29 / 16	**213 / 219 / 138**

To further analyse the nature of the changes, we classified where in the gene the unannotated events occurred. For this, we computed for all of the significant genes the window of 100 bp that contributed the most to the test statistic of rDiff.nonparametric and then determined into what type of region this window bp fell (see Tab. 4.3). We found that most of the changes resided in intronic regions. However, we did not find any evidence for a time point specific enrichment of certain class of events between different comparisons ($p = 1$ using Fisher's exact test and Bonferroni correction). When normalising the counts by the length of the sequence, we found that the highest fraction of reads was still in the intronic regions but almost the same fraction could be observed in 5'UTR and in the 3'UTR (see Fig. 4.10).

Table 4.3.: Categorization of the regions that contained the most differential 100 bp detected by rDiff.nonparametric (FDR≤ 0.1). This is shown for wild type (wt) for all three comparisons between the three samples.

Event	0 h vs 1 h	0 h vs 6 h	1 h vs 6 h
Intronic regions	118	126	77
5' UTR	30	36	23
3' UTR	46	47	23
First exon	29	22	13
Last exon	30	32	13
Other exons	18	10	14

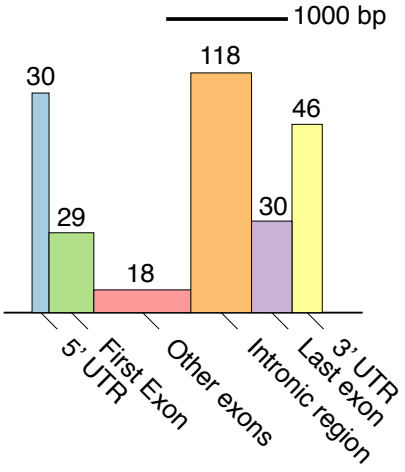

Figure 4.10.: Shown is the number of the most differential 100 bp detected by rDiff.nonparametric (FDR≤ 0.1) between 0 h and 1 h. The width of the bars indicates the average length of these regions in all genes. The area of the bars is proportional to the number of hits and the height is proportional to detection intensity.

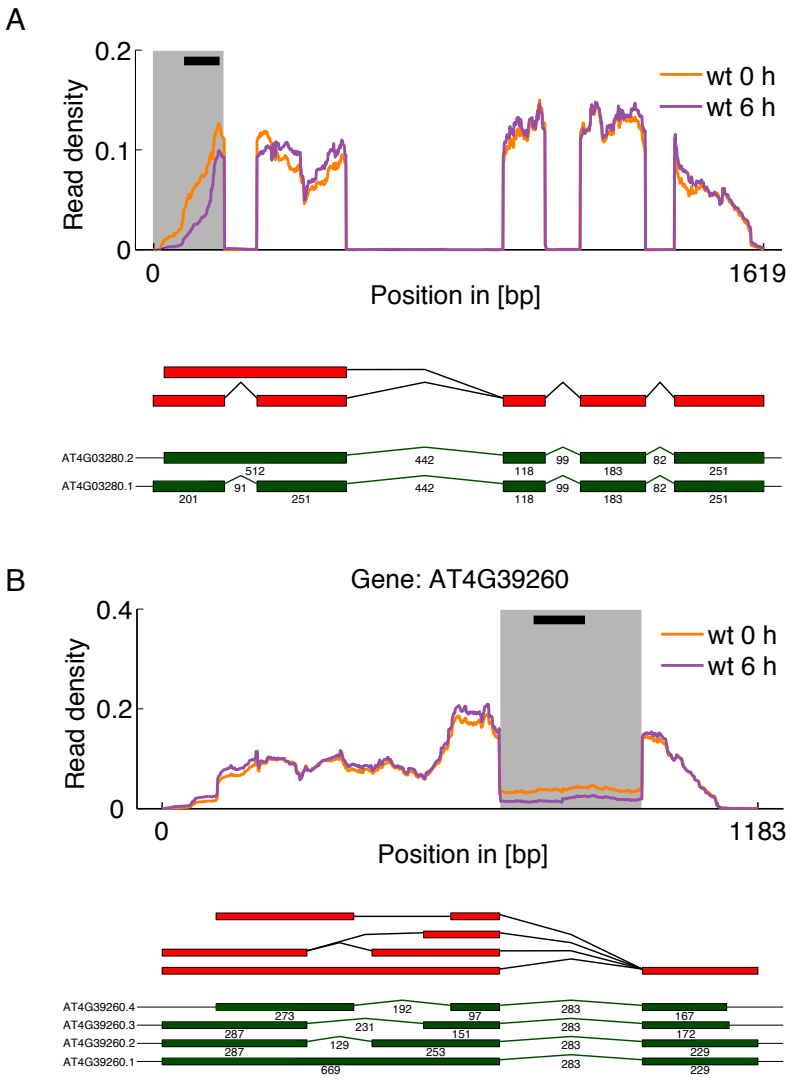

Figure 4.11.: Examples of gene AT4G03280 (**A**) and gene AT4G39260 (**B**) detected by rDiff.nonparametric between 0 h and 6 h. Shown are the read densities for the two respective conditions in orange resp. purple and the gene structures in green. The grey area indicates the regions where the change was detected and the black bar shows the window of 100 bp length that contributed the most to the test statistic of rDiff.nonparametric.

Evaluation on *D. melanogaster* We also evaluated the performance of rDiff.parametric and rDiff.nonparametric on a publicly available dataset from *D. melanogaster* [28] (see Sec. 4.2.8). In a previous study for this dataset, 323 genes were found to be differentially alternatively spliced and 16 of these events were further validated [28]. In this study, these genes were detected using a Fisher's exact test based testing strategy, similar to the one used in [182]. When applying rDiff.parametric and rDiff.nonparametric, we detected 71 respectively 278 genes with differential relative isoform expression (FDR$\leq$ 0.1) on this dataset. Next, we determined the fraction of the validated events from [28] that we could detect. To account for the different calibrations of the methods, we also considered the top 323 genes of our predictions. The 323 most significant genes from rDiff.parametric contained 12 of the validated genes and 11 for rDiff.nonparametric. For three of the remaining validated genes the read coverage was too low to make statistical statements, due to our string alignment strategy. This large fraction of detected validated genes that could be detected, again shows that our methods allow robust detection of differential RNA processing.

4.3.2. Detecting Changes in Secondary Structure

We evaluated the performance of sDiff on a simulated dataset (see Sec. 4.2.7). In order to put the performance of sDiff into perspective, we compared it to two alternative approaches. The first is the straightforward approach of applying rDiff.mmd to the sets of reads of the two nucleases and then adding the two rDiff.mmd test statistics. We refer to this approach here as rDiff.mmd. The second approach that we compared against was the general form of StrucDiff (see Sec. 4.2.5). As we did not simulate biological variation in the dataset, we did not perform subsampling correction of rDiff.mmd and sDiff to account for biological variance. To compare the performance of these three approaches, we computed their auROC (see Fig. 4.12 (A)). We found that sDiff had the highest auROC (0.74) followed by rDiff.mmd (0.70). The original StrucDiff had the lowest auROC (0.58). We believe that this low score is due to the un-tuned pseudo count parameter in the original formulation and that optimal choice can provide better performances.

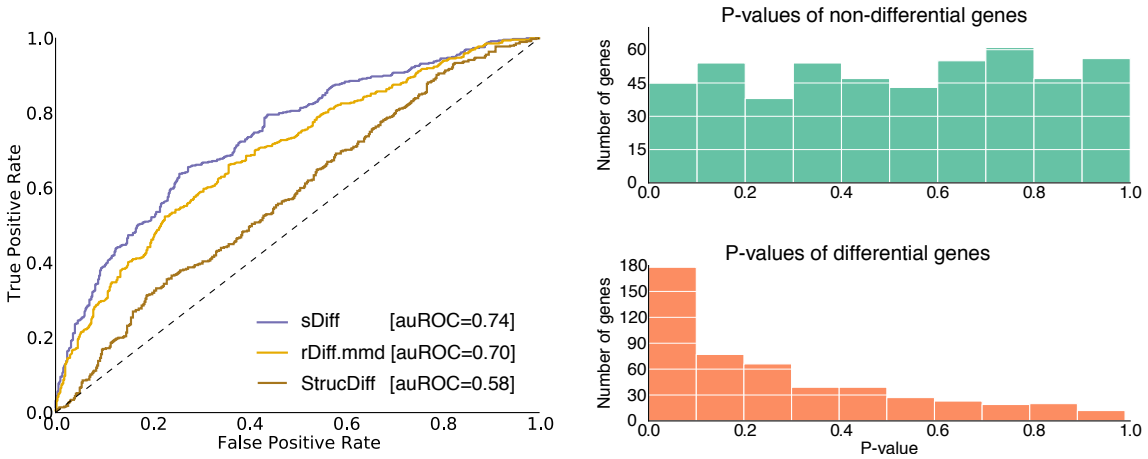

Figure 4.12.: Shown on the left is the ROC curve for sDiff (purple), rDiff.mmd (yellow) and StructDiff (brown). The auROC for the ROC curves is given in the legend. Shown on the right are the p-value distributions of sDiff for the non-differential genes (green) and the differential genes (orange).

Furthermore, we examined the sDiff p-value distribution of the differential and non-differential genes (see Fig. 4.12 (B)). We found that for the non-differential genes the p-values appeared uniformly distribute, showing that the test is well calibrated. The p-values of the differential genes were clearly enriched for small values, showing further that the test is powerful in detecting differences in secondary structure between samples. Overall, these results suggest that sDiff is a powerful tool for detecting differential secondary structure.

4.3.3. Association of Changes in RNA Processing

To evaluate the potential of rDiff.gmmd to associate changes in RNA processing to genetic variants, we evaluated it on simulated data (see Sec. 4.2.7). Briefly, this dataset consisted of two samples, each having 100 replicates for 500 genes, half of which were differential. On this dataset, we evaluated the performance of rDiff.gmmd and compared it to the performance of rDiff.mmd in order to determine how well associations with an allele frequency of 0.5 can be detected. Specifically, we measured the auROC of the methods, when using an increasing number $(5, 10, 20, 30, \ldots, 100)$ of replicates per sample (*sample size*) (See Fig. 4.13 (A)). Although the auROC estimates were slightly unstable we found that for less than 70 replicates per sample, rDiff.mmd had consistently a higher auROC, whereas for bigger sample sizes the opposite was true.

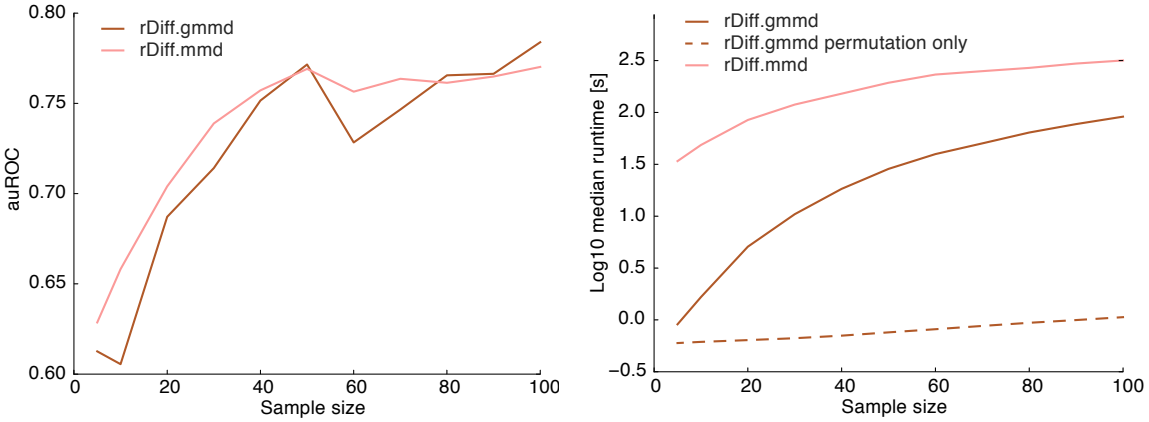

Figure 4.13.: Shown on the left is the auROC of rDiff.mmd (pink) and rDiff.gmmd (brown) for various sample sizes. Shown on the right is the median computation time per gene for rDiff.mmd (pink) and rDiff.gmmd (brown) for different number of replicates per samples. Shown in dashed brown is the median computation time for the bootstrapping for rDiff.gmmd.

Besides analysing the power to detect differential genes we also examined the calibration of the methods. For this, we compared the empirical FDR with the FDR of both methods for a representative subset of the sample sizes (see Fig. 4.14). We observed that rDiff.mmd consistently underestimated the FDR, whereas for rDiff.gmmd the empirical FDR was well in accordance with the FDR, except for the smallest sample sizes. We believe that this reflects the different variance estimation approaches of the two methods: rDiff.mmd estimates the variance function globally, which is more stable and particularly well suited for small sample sizes. In contrast, rDiff.gmmd estimates the variance for each gene independently, which is not as stable for small sample sizes. This disadvantage, however, is out-weighted for larger numbers of replicates by the ability to get more accurate estimates of the biological variance and thus obtain a better-calibrated test statistic.

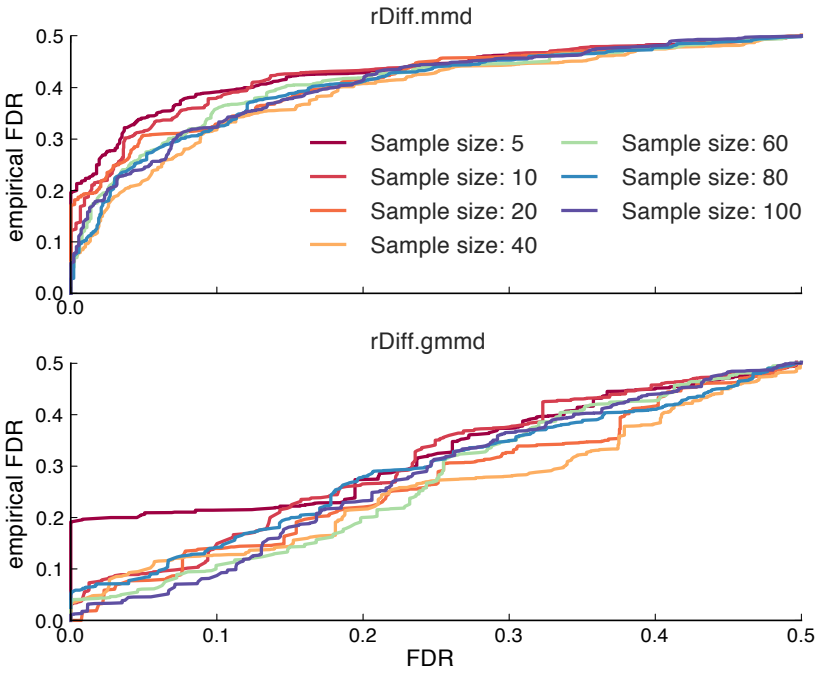

Figure 4.14.: Shown is the comparison of the empFDR and FDR of rDiff.mmd on top and rDiff.gmmd below. The comparison is shown for a subset $(5, 10, 20, 40, 60, 80$ and $100)$ of the sample sizes.

A critical aspect of association tests is the computation time, as typically for each gene thousands of variants have to be tested. We, therefore, compared the median computation time of all genes for both methods (see Fig. 4.13 (B)). For the times for rDiff.mmd we did do not include the time needed to estimate the variance functions, which are therefore optimistic estimates. We found that rDiff.gmmd is for 50 replicates per sample 7 times faster and for a sample size of 100 still 3 times faster than rDiff.mmd. However, for testing the most time consuming part is computation of the kernel matrix, which only has to be done once per gene for rDiff.gmmd. Thus, for other tested associations only the permutation step needs to be repeated. As can be seen in Fig. 4.13 (B), this is much more efficient in computation. For subsequent tested associations, rDiff.gmmd is for 50 replicates per sample 255 times faster and for a sample size of 100 even 298 times faster than rDiff.mmd.

In summary, our results showed that rDiff.gmmd has a similar power to detect differential RNA processing as rDiff.mmd. Furthermore, it has a better calibration and also considerably lower computational complexity. It is therefore, well suited for association studies to investigate the genetics of RNA processing

4.3.4. Applications of rDiff

Besides the evaluations presented before, we used the methods that we have developed in order to investigate other aspects of RNA processing. These range from the investigating mechanisms that alter translation, nonsense mediated decay, but also identification of novel oncogenes in cancer. We will briefly describe, how we have applied our methods in these studies and our findings.

Detecting Differential Translation

The nonparametric tests that we have developed provide the means to detect changes in read distribution while accounting for biological variation. As discussed before, they are also well suited to study changes in RNA processing for which there is no obvious parametrisation. To show this, we applied rDiff.mmd to investigate ribosome binding to mRNA. This was done using a ribosome foot printing dataset that we generated [187], which consisted of the sequenced *H. sapiens* mRNA fragments that were bound by ribosomes. The dataset comprised two samples, one collected after 45 minutes of treatment with the drug *silvestrol* (a drug specifically reduces translation of many oncogenes) and an untreated control. The short time span between administration of the drug and the collection of the RNA allowed investigating changes in translation without the confounding by changed isoform abundances. Both samples had three biological replicates. For details on the procedure we refer to [187]. To this dataset we applied rDiff.mmd in order to detect differential RNA processing by the ribosomes. We did this using the *H. sapiens hg*19 genome annotation to determine the mRNA locations and $10,000$ permutations for the bootstrapping procedure. We found that 847 ($p \leq 0.0001$) genes had a significant change in their ribosome distribution. Further inspection of these genes that showed a significant change showed an accumulation of reads in their 5'UTR. To understand the cause of this accumulation of reads in the 5'UTR, we searched for motifs in these regions. This was done using the discriminative Motif finder DREME [12]. By this, we found a 12-mer motif CGGCGGCGGCGG (see Fig. 4.15) that was significantly overrepresented in the detected genes ($p = 2.2 \times 10^{-16}$). From the 641 of 847 genes that had an annotated 5'UTR, 232 had at least one 12-mer motif. The motif, which we found, resembles closely a secondary structure inducing *G-quadruplex* (GQS) motif that consists of at least four repeats of the nucleotide triplet GGC.

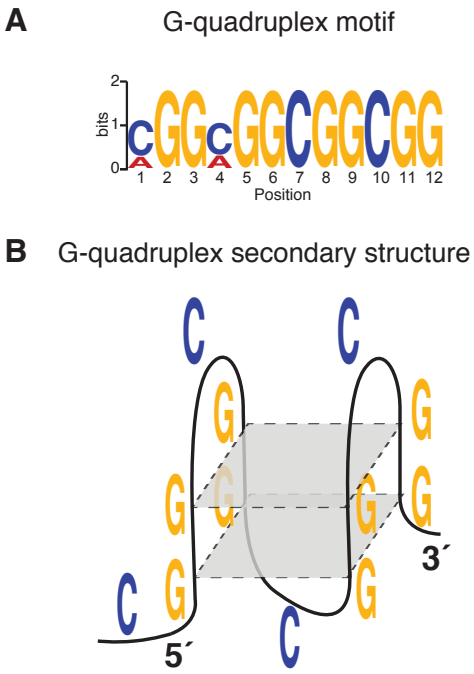

Figure 4.15.: Shown in (**A**) is logo of the detected 12-mer motif. An illustration of the G-quadruplex structure formed by the motif is shown in (**B**). The Hoogsteen-hydrogen bonding is depicted by the grey squares.

To validate that the motif is indeed linked to the GQS we examined the colocalisation of the motif and predicted GQS. For this we predicted GQS using the tool RNAfold (v. 2.1.0) [106]. We confirmed that the motif we found overlaps a GQS 38.4% of the times. Besides the 12-mer motif, we also found three 9-mer motifs that also were predicted to colocalise with GQS (data not shown). We, therefore, hypothesise that the motif induces a GQS in the 5'UTR and that administration of the drug reduces the efficiency of the ribosome to resolve this GQS-structure, which in turn reduces the initiation rate of translation. Therefore production of proteins of genes with a GQS in their 5'UTR is reduced upon administration of silvestrol. This hypothesis is further supported by validation experiments that showed that translation efficiency is decreased, when this motif is cloned into a reporter construct (for details see [187]). This finding shows again that our tests perform well in practical application and provide the community with novel strategies to investigate the mechanisms RNA processing.

Nonsense Mediated Decay Another application of rDiff.parametric includes a study, in which the alternative splicing coupled nonsense mediated decay in *A. thaliana* was investigated [41]. As a first step of this study a detection of unannotated splice events was carried out. Following this each splicing event was tested independently using rDiff.parametric. For the testing, also the direction of change was determined. For this study a *A. thaliana* wild type samples treated with the translation inhibitor cycloheximide and two samples from NMD factor homologs UP FRAMESHIFT1 (UPF1) and UPF3 (see our publication [41] for details) knockdowns were used. In this study, it was shown that 92.3% of the NMD-responsive mRNAs exhibit known classical NMD-inducing features. Furthermore, it could be shown that also many noncoding RNAs and transcripts derived from intergenic regions are subject to NMD. Overall, it was shown that nonsense mediated decay is a central pathway in quality control of the transcriptome and that it has a fundamental role in gene expression regulation.

Testing for Differential Splicing

Application to Cancer Specific Splicing We applied rDiff.parametric to understand the role of splicing in Ewing's Sarcoma, a rare small-round-blue cell tumour in *human*. For this, we used a dataset (see Sec. 4.2.8) consisting of a control and a cancer sample, each comprising three replicates. For each replicate between 3.0×10^7 reads and 3.8×10^7 read-pairs were available. Using rDiff.parametric, we found in total $3,675$ genes (FDR$\leq$ 0.1) that were differentially spliced. This large number of genes that were detected suggests that RNA processing is disturbed in Ewing's Sarcoma. When analysing the detected genes, we found many known oncogenes. The analysis of the top 10 genes showed that 6 out of 10 genes are known cancer associated genes (TPM4 [97], PRSS23 [85], PDE3A [49], RTN4 [144], KIAA1199 [117], PCSK7 [18]). This indicates that many of the remaining genes that we found are indeed oncogenes that could be potential targets for a treatment.

4.4. Software and Webservice

To provide researchers access to our methods and facilitate their use, we have packaged the methods together with an extensive documentation. The methods are available under then GNU General Public License (`https://www.gnu.org/copyleft/gpl.html`) and can be downloaded from: `http://www.bioweb.me/rdiff`.

Furthermore, we provided a wrapper for the Galaxy web platform [22, 54, 56], which can be

downloaded from: `https://github.com/ratschlab/rDiff/tree/master/galaxy`. We also integrated rDiff into Oqtans [161], the first online transcriptome analysis platform, which we developed and that can be obtained from `http://oqtans.org`.

4.5. Summary

The mechanisms that underlie RNA processing are still enigmatic. High-throughput sequencing of transcriptomes provides the data for a better understanding of these mechanisms. However, there is still a lack of robust methods to mine this data. In this chapter, we have proposed several statistical tests for detection of differential RNA processing to address this need.

The tests that we propose can be distinguished by their use of the gene annotation. The two tests, rDiff.poisson and rDiff.parametric, use the gene annotation in order to define the regions where to test. The counts in these regions are modelled using a Poisson respectively Negative binomial distribution. The latter of these two tests, additionally allows accounting for biological variance during testing.

The other tests that we have proposed, namely rDiff.mmd, rDiff.nonparametric, rDiff.gmmd and sDiff, do not depend on the gene annotation for testing. These tests are nonparametric tests that are based on an RKHS embedding of the read distribution. They provide great flexibility in their application, in the integration of different information sources and they also can account for biological variation during testing. Furthermore, their independence from an annotation for testing allows their application, when an annotation is not available. This allows reliable detection of differential RNA processing on non-model organisms, where in general the annotation quality is poor. Besides, it also allows their application in situations, when it is not clear how to define an annotation such as for secondary structure or for ribosome footprinting. Finally, we have proposed the first statistical test that can be applied to associate genome-wide RNA processing variation (rDiff.gmmd).

We have thoroughly evaluated our methods on realistically designed and real datasets. On these datasets, we have shown that the parametric tests outperform state-of-the-art quantification-based approaches and we have shown that the nonparametric approaches are on a par with these quantification-based methods. Furthermore, we have shown that our methods are better calibrated than existing approaches, thus providing more reliable predictions.

Lastly, we have applied our developed methods in several studies. Using our methods we were able to describe changes in the transcriptome and shed light on the underlying mechanisms. This shows that our new methods are a valuable contribution and provide means to advance the study of RNA processing.

5. Genetic Determinants of Gene Expression Changes in *A. thaliana*

5.1. Motivation

Natural populations are shaped by a process of variation and selection. As a consequence of this process, a natural population typically exhibits a broad spectrum of phenotypic variation. This phenotypic variation in the population is largely driven by underlying variations in the genomes of the population; these genetic variants impact parts of the cellular regulation and thus can induce variation in phenotypes.

A perturbation of the genome by a genetic variant can act in numerous ways on molecular and non-molecular phenotypes. For molecular phenotypes of genes, such as its expression, genetic variants can be distinguished by whether they act directly (*cis*) on the gene (e.g. by changing its expression) or indirectly (*trans*) if they affect regulators of the gene. For non-molecular phenotypes that lack genetic localisation this distinction cannot be made. But not all genetic variants necessarily induce a change of phenotypes. Typically, genetic variants in non-functional regions do not cause any changes in phenotype.

A genetic variant can be classified by whether it increases (*beneficial*), decreases (*deleterious*) or does not affect (neutral) the fitness of an organism. In natural populations, the frequency of a variant in a population (*allele frequency*) is linked to its effect: On a long-term run, beneficial variants have a selection advantage and thus typically have a higher frequency. In contras deleterious variants have a selection disadvantage and thus have lower allele frequencies. However, the frequencies of variants are not independent. This is because variants that are close to each other on the genome are unlikely to be separated during meiosis by chromosomal crossovers. Hence, these variants tend to be inherited together [119]. This co-inheritance of variants that are close is commonly referred to as *linkage*. Therefore, slightly deleterious variants in linkage with beneficial variants can be selected for on a short term.

The study of genetic variation in natural populations provides several opportunities: Firstly, it allows to shed light on the evolutionary history of populations by analysing the variant distribution in a population. Individuals or groups that are closely related have a greater overlap in their genetic variants than distant ones. Therefore, the extent of variant sharing can be used to infer the degree of relatedness and thus to retrace the phylogeny of the population. Besides this, allele frequencies that are higher than expected by chance also allow to identify phenotypes (*traits*) that have been selected for in a population and thus to understand the process of adaption.

Secondly, the study of natural populations allows to reveal regulatory mechanisms that are responsible for phenotype variation in the populations and to understand their genetic architectures. This can be done by identifying the candidate variants whose presence correlates with a change in a phenotype (*association mapping*). Thereby, potential stretches of the genomes can be identified that affect the trait under investigation (*quantitative trait loci* (QTL)). For a molecular phenotype such as gene expression, this allows for example to identify potential regulatory elements of transcription.

To map the QTLs, several approaches have been proposed (e.g. [10, 58]). These approaches typically regress a linear model on the phenotype and the genetic variants to identify associations. Many of these approaches also account for confounding factors (e.g. batch effects or population structure).

A major challenge in the analysis of the QTLs is, however, that often not a single variant is associated with a trait, but rather a block of variants. This is because variants that are close to each other on the genome are in linkage. Therefore, these variants will be strongly correlated with each other, meaning that if a variant is correlated with a trait also the other variants in linkage with this variant tend to be correlated with the trait. Consequently, the resolution of QTL mapping is limited to the scale of linkage, which thereby restricts the precise identification of causal variants. Therefore, revealing the true regulators of the genes is still one of the major challenges in QTL analyses.

In this chapter, we will first perform a comprehensive analysis of the gene expression variability in a population of *A. thaliana* of 19 accessions (*strains*) that are part of the 1001 Genome Project [129]. For this, we develop a strategy to robustly estimate gene expression, when the genomes are not identical. We will then characterise the observed expression and differential gene expression patterns in this population. After this, we will investigate how the combination of transcription factor binding information and associations helps to determine the regulatory factors that cause the variability of the gene expression. For this, we use an extensive collection of experimentally derived *A. thaliana* transcription factor binding profiles. Using these profiles we show that a large fraction of the highly significant variants act by changing the transcriptional regulation of genes. Finally, we will quantify the extent of gene expression change that is caused by changing transcriptional regulation in a set of *Multiparent Advanced Generation Inter-Cross* (MAGIC) lines [92] that are derived from the 19 *A. thaliana* strains.

5.2. Methods

5.2.1. Data Preparation

Multiparent Advanced Generation Inter-Cross

Natural populations are a useful resource for studying the regulation of various phenotypes. In contrast to F2-intercrosses, which are created by repeatedly crossing the offspring of two parents, natural populations have the advantage that they have a higher number of variants [92]. Therefore, on one hand, they provide a broader spectrum of genetic perturbations of the regulatory elements and thus offer a better resolution to detect regulatory elements. On the other hand, they have the disadvantage that usually the minor allele frequencies are low, which reduces the statistical power to detect an association [33, 92]. Recently, a novel approach has been proposed that combines the advantage of both of the previously mentioned populations for genetic mapping in *A. thaliana*. This approach consists of using a population of so called Multiparent Advanced Generation Inter-Crosses [92] for genetic analyses. These crosses are obtained by first crossing a natural population of 19 A. thaliana inbred lines (*founder lines*) with one another. Next, the offspring are repeatedly crossed with each other for several generations, to create mosaic genomes. The last generation of the offspring are finally inbred in order to obtain the homozygous lines, the *MAGIC lines*. Through the recombination of the founder population, the resulting strains have parts of the genome from many of the MAGIC founders (see Fig. 5.1). These MAGIC lines have the advantage that the

expected minor allele frequency is at least $\frac{1}{19}$ while they still have a high density of genetic variants.

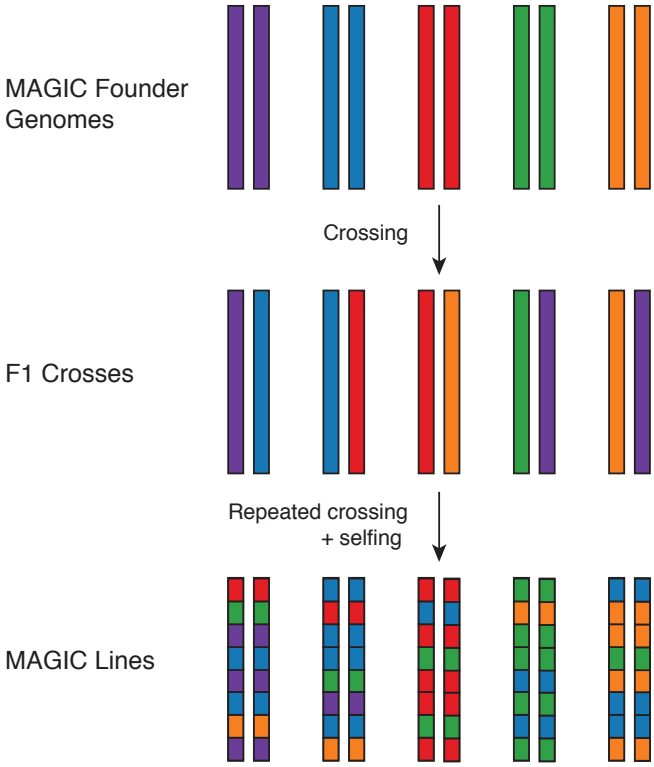

Figure 5.1.: Shown is the MAGIC design. From a population of inbred founder strains (shown on top) first, F1-crosses are derived (shown in the middle). The offspring of these F1-crosses are subsequently crossed for several generations such that genomes of the last generation of crosses are a mosaic of the founder's genomes. Lastly, the strains of the last generation are inbred to obtain the homozygous MAGIC lines (shown on the bottom). This figure has been adapted from an illustration in [91].

MAGIC Founders We sequenced, assembled and annotated the genomes of all the MAGIC founder strains except *Col-0* (for details see [52]). In the following, we refer to the novel genes that were detected during annotation of the genomes as *new genes*. Furthermore, we performed transcriptome sequencing of all 19 founder strains. For these strains, we obtained RNA-samples from root-tissue of 10 days old seedlings (for details see [52]). For each of the lines two biological replicates were sequenced. By this, we generated on average 5.0×10^6 single-end reads of length 78 bp per library. From these reads we could align on average 4.7×10^6 (95.0%) per library against their respective genome using PALMapper [75] (for details see [52]).

MAGIC Lines We also analysed the transcriptome of 208 MAGIC lines. For this, the transcriptome of samples generated from seedlings were sequenced. This yielded on average 13.6×10^6 paired-end reads of length 100 bp per library. For these libraries a variant aware alignment was performed using PALMapper [75] (pers. comm. Andre Kahles). For the alignment of the reads, at most three mismatches and no gaps were allowed. Furthermore,

all TAIR10 junctions as well as all junctions and variants from the MAGIC founder genomes were used for the alignment. Additionally, the genomes of the MAGIC lines was imputed using the RNA-Seq libraries (pers. comm. Robert Greenhalg).

Transcription Factor Binding Data

For the analyses that involved the transcription factor binding profiles, we used a collection of 254 positional weight matrices from *A. thaliana* transcription factors that were experimentally determined using protein binding arrays [16] (pers. comm. Matthew Weirauch and Timothy Hughes). These transcription factors were selected in order to sample a broad range of binding profile in *A. thaliana* and thus are representative for the entirety of transcription factors in *A. thaliana*.

5.2.2. Gene Expression Quantification Strategies for Populations

One of the most basic tasks for the analysis of RNA-Seq is the estimation of gene expression. These estimates provide the foundation for many subsequent analyses, such as the calling of differential gene expression, network analyses or association studies to name only a few. The estimation of gene expression is usually accomplished by counting the number of reads that map to a gene. When comparing the gene expression estimates of two or more libraries, these counts are typically normalised by the library size in order to account for changes that are caused by differences in library sizes. Some authors propose, to further normalise the expression for differences in gene length in order to derive estimates that correlates better with the abundance of RNA molecules, leading to the so called measure reads per *kilo base per million mapped reads* (RPKM) [121]. Even more elaborated approaches first estimate the transcript expressions and then use these expressions, to estimate the gene expression (e.g. [55, 173]). This provides a certain degree of robustness against changes in the number of reads due to inclusion of introns or skipping exons.

These distinct approaches to quantify gene expression all have in common that they critically rely on the assumption that the gene expression is a monotone increasing function of the number of reads mapping to the respective gene. In particular, they all assume that if the library size is constant, the expected number of reads only changes if the gene expression does. However, in RNA-Seq experiments, this is not always the case: Reads can map to multiple locations or stem from unannotated loci, thereby mixing the gene expression estimates of different genes. Furthermore, biases can influence the estimation of gene expression [137]. Another factor that can affect the estimation when comparing gene expression between samples, is genomic variability in some samples (*structural variation*). This is illustrated by the following example: Assume that the expression of a gene is to be compared between two individuals and that for one individual the gene is partially deleted. Then, when abundance of transcripts in both individuals is the same, we still expect to observe fewer reads for the individual with the deletion in the gene under investigation. In this case, the fallacy is that not the expression of the same gene is compared, but rather the expression of two slightly different genes.

Therefore, it is advantageous to use a gene expression estimation strategy that is robust against biases and structural variation, in order to avoid drawing wrong conclusions on the cause of the change. This is especially important when the transcriptomes of different individuals or species are compared. However, to our knowledge, such estimation strategies are still not available.

To remedy this lack of strategies, we therefore propose a read filtering procedure that can be applied prior to the estimation of gene expression. In this procedure, reads are filtered in several steps to remove reads that can potentially confound the estimation of gene expression. For this, we propose to apply the following filtering strategy (see Fig. 5.2 for an illustration of selected filters):

Exonic regions Discard all reads that either map to a region that is intronic in any transcript or disagree with the gene structure. This filter prevents changes in the relative transcript abundance to influence the read counts. This is because changes in relative transcript abundance only lead to changes in informative regions, i.e. regions that are intronic in at least one transcript. When comparing different individuals or species with differing gene annotations we suggest extending this filter. In this case, we suggest to discard all reads that that map to regions that are intronic in any of the genomes.

Overlapping regions Discard all reads that map completely into a region that overlaps another gene in order to prevent mixing expression estimates of overlapping genes.

Ambiguous mapping Only retain reads that map non-ambiguously, in the sense that their second best alignment has at least two insertions or deletion (indels) or mismatches more then the best alignment. This prevents reads that originate from another location to map wrongly to the gene under consideration and thereby affecting the estimation, even when the mis-mapping is caused by sequencing errors. An alternative to this filter is to filter for reads that map into regions that have no repeat of at least a read length in the genome. However, this approach fails to account for repetitive spliced reads and sequencing errors in reads.

Deletion and Insertions When estimating the gene expression of multiple individuals or species, where deletions and insertion with respect to the reference are known, we furthermore propose to discard all reads that start in insertions. This eliminates the effect of these surplus reads on the expression estimation.

Genomic Uncertainty When reconstructed genomes are available we propose to discard any reads that map to regions where in any of genomes the assembly is not reliable. This filter prevents the inclusion of reads that map to regions that are potentially deleted in some genomes but not identified as being so.

Applying these filters before gene expression estimation removes unwanted confounding effects and, thus, increases reliability of the estimate for subsequent analyses. However, this gain in reliability comes at the cost of loosing reads. Therefore, depending on the intended use of the estimates, a trade-off between the strictness of the filters and the number of remaining reads can be reasonable.

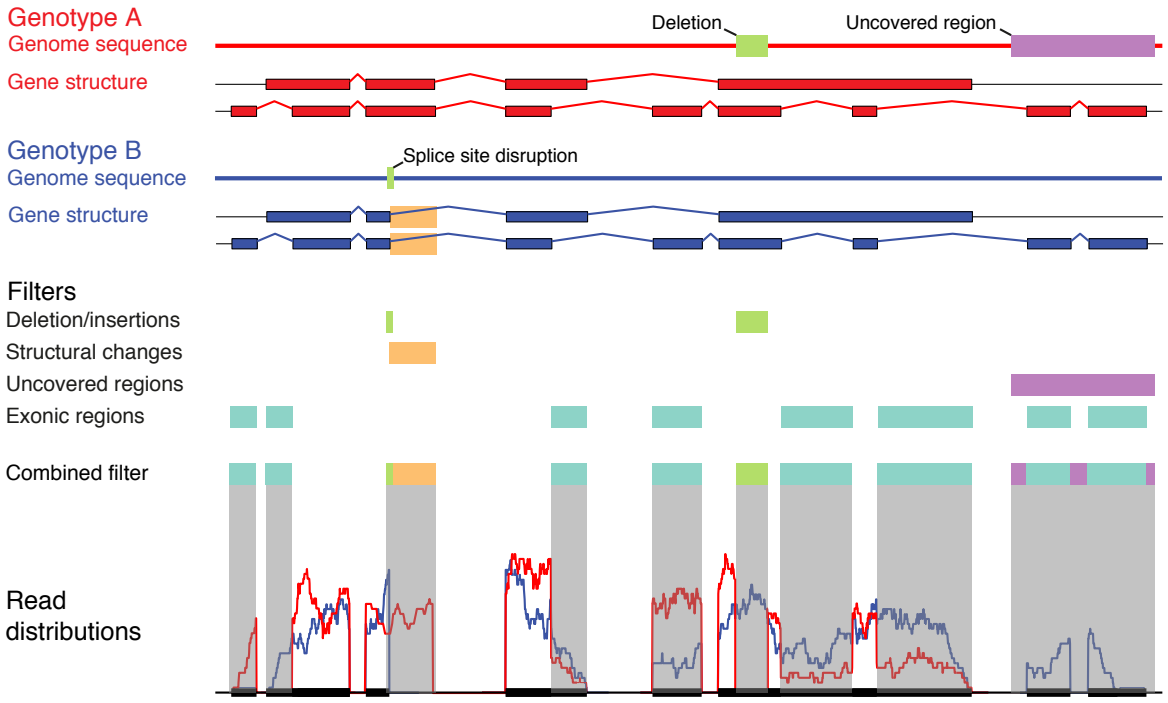

Figure 5.2.: Illustration of read filtering. Shown on top are the two genomes and predicted gene structures for two individuals A (red) and B (blue). In this example there is a deletion (light green) and an uncovered region (light purple) in the genome A. In genome B there is a disruption of the splice site (light green) that leads to a gene structure change. Shown in the middle are the individual filters: The filter for regions that are deleted or inserted (Deletions/Insertions). The filter for regions, where the gene structure changes (Structural changes). The filter for regions, where the genome assembly is unreliable (Uncovered regions) and the filter for regions that are not constitutively exotic (Exonic regions). Shown below the individual filters is the combination of the filters (Combined filter) and the read distribution from A (red) and B (blue). The quantification is based on reads that map to regions that are not filtered for (not shown as grey).

Statistical Test for Gene Expression

When analysing a transcriptome, an immediate question is, which genes are expressed. To answer this question, it is a common approach to define a minimal number of reads a gene needs to have. However, this cutoff is often arbitrary and it is not clear how to choose the cutoff when comparing libraries of different sizes. To provide a statistically sound definition of expressed genes, we therefore propose the following statistical test to determine these genes.

For this test, we define as Null hypothesis that the gene is not expressed. We assume that the number of reads that map to a non-expressed gene follows a Poisson distribution $\mathcal{P}_\lambda$ that models sequencing noise. Therefore, to construct the p-value for a gene we need to estimate the intensity λ in order to determine the Null distribution. For this we propose to assume that the number of expressed gene that have zero read counts is negligible.

We can then estimate λ using the fraction f_0 of genes that have zeros read counts in the not-expressed genes:

$$\mathcal{P}_\lambda(0) = e^{-\lambda} = f_0$$

This leads to the estimate $\hat{\lambda}$ of λ:

$$\hat{\lambda} := -\log f_0$$

With the estimate $\hat{\lambda}$ we can finally compute the p-value p for a gene with n read counts to be expressed:

$$p = 1 - \sum_{i=0}^{n-1} \mathcal{P}_{\hat{\lambda}}(i)$$

Since, however, the unexpressed genes are not known, we estimate f_0 as fraction f_0 of genes that have zeros read counts in all genes. In the case where many genes are not expressed, this leads to a slightly more conservative test for the expression of genes. In the case where almost all genes are expected to be expressed, this approach is overestimating λ and thus leads to overly conservative p-values. In this case, we propose to estimate f_0 on read counts of random intergenic regions that have a similar length as the genes that are tested.

Overall, the proposed test can be used to determine expressed genes while providing a statistical estimate of the significance. Furthermore, the number of expressed genes is also robust to changes of library sizes. The test we propose is therefore particularly well suited when the number of expressed genes are to be compared between libraries.

5.2.3. Detection of Differential Gene Expression

To test for differential gene expression between libraries we used DESeq [5]. For the testing, this software first fits two variance functions on the observed read counts for each of the two sample that are to be compared, in order to estimate the biological variance (similar to the approach described in Sec. 4.2.3). Subsequently, the variance functions are then used to calibrate the variance of the Null distribution during testing. In order to obtain a stable estimate of biological variance for our analysis we estimated the variance functions on the protein coding genes only, as these typically they have a better genome assembly and annotation. We then used these variance functions for differential testing.

The estimated variance function were furthermore used in order to compute variance stabilised counts. This is a transformation of the counts that results in counts whose variance across the replicates is independent of the mean expression. The variance stabilised counts allow their modelling with methods that assume the same variance for all genes, as it is often the case for linear models or mixed models.

5.2.4. Transcription Factor Binding Site Prediction

To examine the extent to which genetic variants influence gene expression by changing transcription factor binding sites (TFBS), we predicted these binding sites in the promoters of all genes in all 19 MAGIC founder strains (see Sec. 5.2.1). The prediction was performed on the first $2,000$ bp upstream of the TSS using FIMO [60], using the default parameters of FIMO. To avoid biasing the prediction of TFBS towards the core promoter we accounted for the high CG-content near the TSS. For this, we used the nucleotide frequencies in the first 200 bp upstream of the TSS from all genes as background distribution as a background model in TFBS calling.

For the MAGIC lines (see Sec. 5.2.1), we predicted the TFBS in the following way: As the genome of the MAGIC lines is a mosaic of the genomes of the MAGIC founder strains (see Fig. 5.1), we imputed the TFBS for the MAGIC lines from the predicted TFBSs of the respective MAGIC founders.

5.2.5. Binding Affinity Computation

We quantified the overall binding affinity of transcription factors to promoters in order to investigate how the affinity changed between different genomes. This was done by first computing the binding affinity of a TF to a TFBS and then deriving from these affinities for single TFBSs the affinity of a TF to a promoter. We did this in the following manner:

Let $S = s_1, \ldots, s_L$ be a promoter sequence of length L, where $s_i \in \{A, C, G, T\}$ for all $i \in \{1, \ldots, L\}$. Let T be a transcription factor PWMs that is given by $T \in [0, 1]^{4 \times l}$, where l is the length of the PWM T (see Sec. 2.4.2). Finally, let $P, P \in [0, 1]$ be a pseudo count and b_A, b_C, b_G and b_T the nucleotide frequencies of a background model.

First, we defined the affinity for a single putative TFBS similar to [191]. Assume for this that $s = s_j, \ldots, s_{l+j-1}$ is a putative binding site of length l starting at position $j \leq L - l + 1$ in S. We then defined the affinity $a^+(T, s_j^l)$ of T and the putative TFBS s_j^l on the forward strand to be:

$$\begin{aligned} a^+(T, s_j^l) &:= \exp\left(\frac{1}{l}\sum_{p=1}^{l} \log \frac{T(s_{j+p-1}, p) + P\, b_{s_{j+p-1}}}{b_{s_{j+p-1}} + P}\right) \\ &= \left(\prod_{p=1}^{l} \frac{T(s_{j+p-1}, p) + P\, b_{s_{j+p-1}}}{b_{s_{j+p-1}} + P}\right)^{\frac{1}{l}} \end{aligned}$$

We defined the affinity $a^-(T, s_j^l)$ for the negative strand analogously, i.e. by computing the affinity of the reverse complement of s_j^l. With these two affinities for the both strands we define the combined affinity of both strands $a(T, s_j^l)$ as the maximum of both scores:

$$a(T, s_j^l) := \max(a^+(T, s),\, a^-(T, s))$$

Based on this affinity for a TFBS, we finally computed the affinity $A(S, T)$ of a PWM T for a promoter S as:

$$A(S, T) := \sum_{p=1}^{L-l+1} a(T, s_p^l),$$

where $s_p^l = s_p, \ldots, s_{p+l-1}$ denotes the subsequence of length l starting at position p of S.

5.2.6. Gene Expression Variance Decomposition

To determine the fraction of the gene expression variance that can be explained by changes in the promoter affinity, we performed a variance decomposition of the gene expression variance into the variance that can be explained by the changes in promoter affinity and into a noise component. For this, we fitted a mixed model (see Sec. 2.6.3) on the gene expressions. The fitting of the model was performed for each gene in the following way:

We first computed the covariance matrix of the promoter affinities in the MAGIC lines. To simplify notation we again omitted the index for the gene whenever possible. We assumed that we had n transcription factors $T_1, \ldots, T_n$ and m strains $M_1, \ldots, M_m$. If we denote by $A_j \in \mathbb{R}^n$ the vector of promoter affinities of the transcription factors for the promoter in strain j, then we defined the unnormalised affinity covariance matrix $\mathbf{A}' \in \mathbb{R}^{m \times m}$ by:

$$\mathbf{A}' := \begin{pmatrix} cov(A_1, A_1) & \cdots & cov(A_1, A_m) \\ \vdots & \ddots & \vdots \\ cov(A_m, A_1) & \cdots & cov(A_m, A_m) \end{pmatrix}$$

We further performed a trace normalisation of the unnormalised affinity covariance matrix:

$$\mathbf{A} := \frac{m}{tr(\mathbf{A}')} \mathbf{A}'$$

This normalisation causes the sum of eigenvalue $\mathbf{A}$ to be the same for all genes, thus making their scales comparable.

Finally, we fitted the following mixed model to the vector of standardised (i.e. having mean zero and variance one) variance stabilised read counts $\mathbf{Y} \in \mathbb{R}^m$ (see Sec. 5.2.3):

$$\mathbf{Y} = \alpha \mathbf{X} + \beta \epsilon \tag{5.1}$$

For this model we assumed that $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{A})$ is a random effect that is normally distributed with mean 0 and covariance $\mathbf{A}$. Furthermore, we assumed that the noise $\epsilon \sim \mathcal{N}(0, I)$ is also normally distributed with mean 0, the identity matrix as covariance matrix and that α and β are the coefficients of the two random effects. The fraction of variance that is explained by the covariate $\mathbf{X}$ is then given by $\frac{\alpha}{\alpha+\beta}$. For the fitting of the model we used the LIMIX package (pers. comm. Oliver Stegle) with default parameters, a method that maximises the log-likelihood of the parameters α and β. We refer to this model in the following as the *affinity variance model*.

We furthermore investigated the genetic variants alone has the same explanatory power for gene expression as the promoter affinities. For this we fitted a second model on the gene expressions $\mathbf{Y}$ that was based on the similarities between the promoters. To derive this model, we first represented for each strain j the genetic variants as a vector $s_j \in \mathbb{R}^{2000}$, where an entry $s_j(k)$ of s_j that was non-zero represented a genomic variant at position k in the promoter with respect to *Col-0*. We then computed, analogously as for the affinity, a normalised promoter covariance matrix $\mathbf{S}$. Finally, we fitted the model 5.1, except that we used covariance matrix $\mathbf{S}$ for the random effect $\mathbf{X}$. We refer to this model in the following as the *promoter variance model*. For both models we did not account for the population structure in the model. We refrained from this as in the MAGIC founders the effect of population structure is only minor [52].

5.3. Results and Discussion

5.3.1. MAGIC Founder Transcriptome Variability

Gene expression is a fundamental molecular phenotype that underlies many other phenotypes. Therefore, characterisation of this molecular phenotype in a natural population allows to shed light on the genetic architecture that underlies phenotypic variation in natural populations. Here, we analysed and characterised the variability of gene expression in the 19 *A. thaliana*

MAGIC founders. For this, we first filtered the reads to remove the ones that could confound the gene expression estimation. After this, we estimated the gene expression and characterised the variability of it in the MAGIC founders. Finally, we analysed the differences in gene expression patterns for genes of different functional roles.

Gene Expression Quantification with Genomic Variation and Uncertainty

We estimated the gene expression for the 19 *A. thaliana* MAGIC founders (see Sec. 5.2.1). For this, we first filtered the reads to remove the ones that could confound the analyses described in Sec. 5.2.2. We applied the filters as follows:

(1) We discarded all reads that mapped ambiguously, meaning that their second best alignment had fewer than two mismatches or indels than the best alignment. On average, this removed 19% of the reads per sample that could potentially confound the estimation.

(2) To minimise the direct effects of polymorphisms on the number of reads, we excluded all reads that started in regions that were insertions or deletions in any of the 19 strains. We furthermore, removed all reads from the expression estimation that mapped to regions where the genome assembly was not reliable, i.e. for regions where there was a lack of reads to support the assembly (*uncovered regions*). This effectively removed 20.425 Mb, that is 26.0%, from the total 75.476 Mb of exonic regions that were considered for gene expression quantification.

(3) We also accounted for the effects of polymorphisms that act indirectly by changing the gene structure. We did this by discarding reads that map to regions where the gene annotation differed between strains. This further reduced the size of the considered exonic regions to 54.870 Mb (99.7%). Overall, the filtering reduced the considered exonic region for 457 protein-coding genes (4,550 when including transposable elements) to zeros.

(4) We discarded all reads that mapped to intronic regions, disagreed with the gene structure or that mapped completely to regions where two genes overlapped.

We finally counted in each replicate for all strains the numbers of filtered reads that mapped to the genes. For this, in total between 1.2×10^6 and 4.9×10^6 reads per replicate were used to estimate gene expression.

Expression Analysis

To determine the number of expressed genes in each strain, we summed the read counts of both replicates and then applied our statistical test to detect the expressed genes (see Sec. 5.2.2). As a significance threshold, we used an FDR ≤ 0.05. By this, we found that between 18,598 and 19,593 protein coding genes were expressed in the individual strains and 20,173 (73.6%) genes were expressed in at least one strain (see Tab. 5.1). This was slightly less than the number of expressed genes reported in [149]. This slight discrepancy was likely due to the larger number of tissues used in [149]. From the 1,167 non-coding RNAs and 914 pseudogenes were 215 (21.7%) respectively 147 (21.2%) expressed. For the newly predicted genes, the analysis revealed that 314 of 447 (70.2%) of them were expressed.

Next, we analysed the gene expression variability in the MAGIC founder population. For this, we used the gene expression estimates in order to detect genes, whose expression changes significantly (see Sec. 5.2.3). With an FDR≤ 0.05 we found that 9,015 (44.7%) expressed protein-coding genes were differentially expressed between at least one pair of strains and that

Table 5.1.: Effect of filtering on genes of different types.

Gene type	Expressed genes		Difference
	without filtering	with filtering	
Protein-coding genes	20,550	20,173	377 (1.8%)
ncRNA genes	253	215	38 (15.0%)
Novel genes	314	274	40 (12.7%)
Transposable elements	452	257	195 (43.1%)
Transposable element genes	88	36	52 (59.1%)
Pseudogenes	196	147	49 (25.0%)

95 (1.1%) of them had more than a 100-fold change (see Tab. 5.2). Furthermore, we observed that $\sim$ 60% of the differentially expressed genes had more than five strains contributing to the differential expression, suggesting that for these genes variability in gene expression is not under negative selection.

Furthermore, we examined the effect of filtering on gene expression quantification and the detection of differentially expressed genes. For this, we performed the gene expression without prior read filtering. We first investigated the effect of filtering on gene expression quantification. Here, we focused on two aspects: First the number of genes that were lost due to filtering and second, the overall difference in the gene expression estimates.

We examined how many genes could not be detected anymore due to filtering. We observed that the filtering of the reads affected the gene expression of different categories of genes to different degrees (see Tab. 5.1). For protein-coding genes, the filtering lead to a decrease of 377 (1.8%) genes being expressed. The effect of filtering was more pronounced for non-coding RNAs and pseudogenes. Here, 38 (15.0%) respectively 49 (25.0%) genes were not significantly expressed any more. For the novel genes, filtering reduced the number of expressed genes by 38 (15.0%).

We then quantified the changes in gene expression estimates that were caused by the filtering. For this, we tested for differential gene expression between the quantifications obtained when using the filtered and the unfiltered reads. Overall, the filtering for reads that map to multiple locations, multiple genes or non-exonic regions prior to quantification, lead to significant changes in gene expression in at least one strain for 631 genes (FDR ≤ 0.05). Further filtering for structural variation lead to a change in at least one strain for another 425 genes (264 protein-coding genes) at an FDR ≤ 0.05.

Next, we investigated the effect of filtering on the detection of differential gene expression (see Sec. 5.2.3). Here, we found that using the filtered reads decreased the number of detected protein coding genes by 345 (3.7%). For the non-coding RNAs and pseudogenes 35 (41.7%) respectively 25 (30.9%) were found fewer. For the novel genes, the number of differentially expressed genes was reduced by 36 (29.8%). Particularly, we found that for genes with a fold change larger than 100 were considerably more often affected by the filtering than the other genes.

Table 5.2.: Effect of filtering for differentially expressed genes and differentially expressed genes with a maximal fold-change larger than 100. The effect is shown for different gene types.

Gene type	Differential genes			Genes with fold-change > 100		
	unfiltered	filtered	Difference	unfiltered	filtered	Difference
Protein-coding genes	9,360	9,015	345 (3.7%)	142	95	47 (33.1%)
ncRNA genes	84	49	35 (41.7%)	5	3	2 (40.0%)
Novel genes	121	85	36 (29.8%)	2	1	1 (50%)
Transposable elements	85	32	53 (62.4%)	3	2	1 (33.3%)
Transposable element genes	47	18	19 (40.4%)	3	1	2 (66.7%)
Pseudogenes	81	56	25 (30.9%)	6	5	1 (16.7%)

In summary, filtering of the reads prior to quantification of gene expression lead to a reduction of a small number of protein-coding genes that could be detected as expressed. In contrast, the reduction in the number of expressed genes was for the other categories more pronounced. The same trend could also be observed for the number of genes that were differentially expressed, with the notable exception that protein-coding genes that had a high maximal fold change in expression were also strongly affected by filtering.

As our filtering strategy was conservative, in the sense that we aimed to exclude all reads that could potentially confound the gene expression estimation, the main concern is that we discarded too many reads from the analysis. Here we have shown, however, that the loss of genes that were detected as expressed or differentially expressed was minor. This showed that the filtering approach is practicable. Furthermore, our findings show that there are big differences between the quantifications obtained using the filtered and unfiltered reads (e.g. in the genes with high fold-changes). As our filtering strategy only slightly affects expression quantifications of genes with no potential source of confounding reads, this suggests that the observed differences were caused by reads of unknown origin. Consequently, this showed that filtering these reads provided more accurate gene expression estimates.

Functional Analysis of Gene Expression

Conservation of gene regulation has been shown to vary for genes with different functional roles (e.g. see [36, 40]). To systematically describe this pattern of conservation for genes of distinct functional roles in *A. thaliana*, we analysed their gene expression variance.

For this, we first analysed the functional role of genes using a functional annotation of genes, the *gene ontologies* (GO) [8]. For this, we determined the significantly enriched GO-categories in the differentially expressed genes. We found a significant enrichment of differentially expressed genes ($p \leq 0.001$ using Fisher's exact test) in 18 GO-terms (see Tab. A.2). Notably, from these GO-terms all but one were related to response to the biotic environment (e.g. pathogen defence). We hypothesise, that variability in gene expression in these categories

reflects the need of the population to adapt to environmental changes.

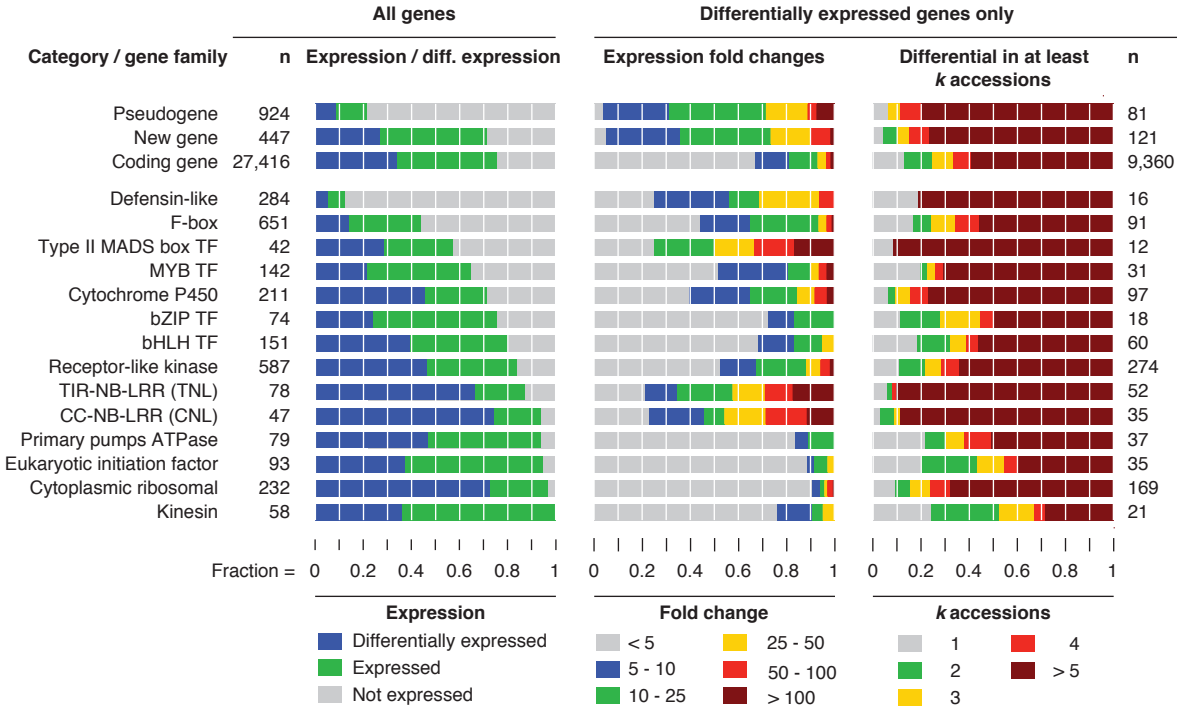

Figure 5.3.: Pattern of gene expression for different gene types and selected gene ontology categories. Shown on the left is the fraction of expressed and differentially expressed genes, in the middle their fold change distribution and on the right the fraction of accessions in which they are differential with respect to the strain *Col-0*. The numbers on the sides indicated the number of genes of each type and category. This figure has been adapted from our publication [52].

We furthermore analysed for different GO-term categories the distribution of gene expression variability (see Fig. 5.3). Here, we analysed three main aspects of the distribution of this variability: (1) The fraction of expressed and differentially expressed genes for a category. (2) The strengths of changes and (3) in how many strains the expression changed. We found that the patterns of variability were distinct for different gene functions. For example, 74% of the defence related NB-LRR genes were differentially expressed and had up to a 400-fold change in expression. Furthermore, we found that the expression of the NB-LRR was in general also variable in a large fraction of the population. In contrast, the housekeeping genes (e.g. ribosomal proteins, transcript factors or kinesins) showed different characteristics: Although a large fraction was expressed, gene expression was mostly stable and changes were generally restricted to a few strains. This showed that there are distinct patterns of gene expression variability for genes in different functional roles.

5.3.2. Dissection of *A. thaliana* Gene Regulation Variance

The *cis*-regulation of gene expression is one of the main factors of transcriptional regulation. In order to understand the mechanisms that drive the variation of gene expression, we investigated the extent to which genetic variation affects *cis*-regulation of gene expression For this, we first analysed the distribution of transcription factor binding sites in *A. thaliana* promoter

regions. We then showed that a large fraction of genetic variants that had a significant association with gene expression variation was located in TFBSs. We showed, how this fact helped to derive functional interpretations of association studies. Finally, we quantified the extent of gene expression variation that can be attributed to genetic perturbations of *cis*-regulation.

Transcription Factor Binding Landscape

To picture the binding landscape of transcription factors in *A. thaliana*, we first predicted potential TFBS (see Sec. 5.2.1). For this we used a set of 254 experimentally determined binding profiles of *A. thaliana* transcription factors. This set contained more than twice as many binding profiles than previously available and thus allowed us to get a picture of the binding landscape of TFs at an unprecedented level of detail. We then used the binding profiles to predict potential TFBSs in the promoters of all genes (see Sec. 5.2.4). As typically the majority of promoters is included in the 2 kb upstream of the TSS [93], we used these regions as promoter regions. We found that we could predict for 220 TF binding profiles with high specificity TFBSs ($p \leq 10^{-5}$, provided by FIMO). For the remaining 34 TFs the TFBSs were not specific enough to provide a basis for subsequent analyses.

Using the predicted binding sites, we then analysed the density of the regulatory network (see Fig. 5.4). For this, we first determined the number of TFBS in the promoters and the number of genes the TFs regulate. We found on average 7.3 TFs bound a promoter. This is less than the numbers reported in [89]. In their work, the authors predicted on average 16.6 TFBS in 500 bp promoter regions. For this they used 144 TF consensus sequences. We believe that this difference in the number of predicted binding sites can be explained by two reasons: Firstly, that our set of motifs does not contain some core promoter regulatory elements (e.g the TATA-binding protein), which are typically predicted in most genes. The second reason is that for their predictions, the authors used consensus sequences as short as 5 bp length. This leads to many unspecific predictions, which is reflected by the fact that in their predictions, the 5 consensus sequences with the highest numbers of predicted TFBS are all 5 or 6 nucleotides long and account for 37.2% of all predictions. We therefore believe that the lower number of predictions we obtained results from a higher specificity of our predictions and the focus on non-core-promoter elements. Besides investigating the numbers of motifs per promoter, we also examined the number of promoters a TF regulates. Here, we found that each of the TFs, for which binding sites could be predicted, bound on average to 741.4 promoters.

Transcription Factor Binding Spacial Preferences

We investigated the spatial binding preference for the TFs. For this, we determined for each TF its binding frequency at all positions in the promoter, i.e. how often the TF bound at a certain distance of the TSS. We then performed a spatial smoothing of the binding frequencies by sliding a window of 25 bp width along the vector of binding frequencies and averaging the frequencies in the window. Next, in order to uncover the predominant classes of binding preferences we performed a hierarchical clustering of these binding frequencies. For this, we applied average linkage clustering, using the Euclidean distance between the binding frequency estimates. Finally, we analysed the top 5 clusters.

We found that these clusters varied substantial (see Fig. 5.5). While the largest cluster was composed of more than 100 binding profiles, the smallest cluster contained only two profiles. Furthermore, we observed that although most clusters had the strongest enrichment of TFBS

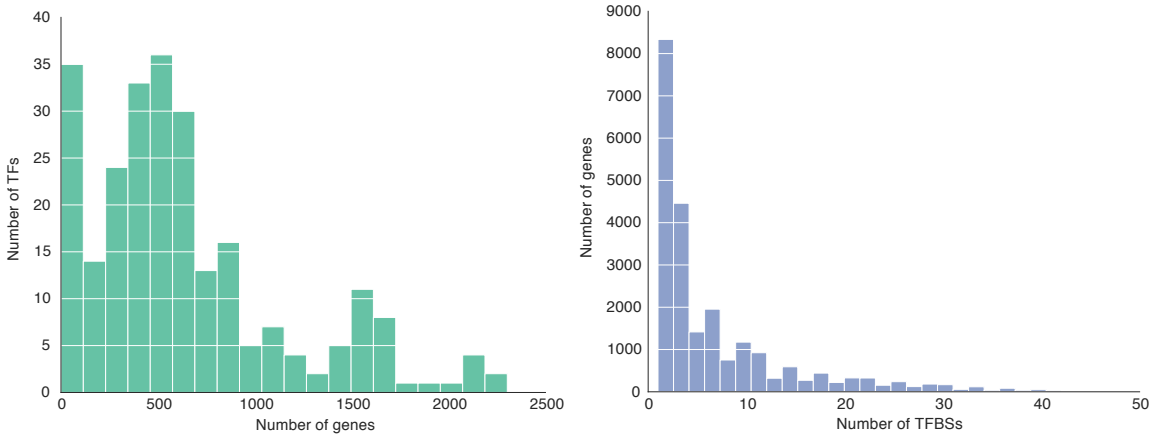

Figure 5.4.: Shown on the left is the histogram of the number genes that are regulated by each transcription factor. Shown on the right is the histogram of the number of TFs that bind the promoters.

in the first 200 bp upstream of the TSS, the profiles showed distinct spatial preferences. One profile had the highest density immediately next to the TSS while the others either had it further upstream or had no apparent spatial preference.

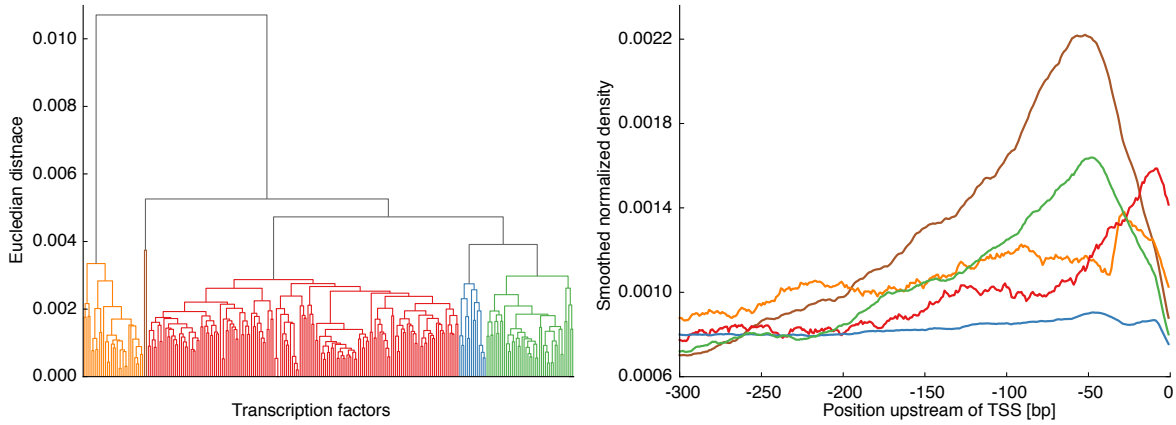

Figure 5.5.: Shown on the left is the hierarchical clustering of the spatial binding distribution of the TFs. The five colours indicate the top-five clusters. Shown on the right is the average spatial preference of the five clusters for the first 300 bp upstream of the TSS.

In summary, we predicted TFBS for the promoters. We found that the there is a dense network of TF-promoter interactions. Furthermore, we found that the TFBS clustered as expected close to the TSS and that they had distinct spatial binding preferences. This is well in line with previous observations. We are, therefore, convinced that our predictions contain many functional TFBS.

Association Enrichment

The mechanism by which genetic variants affect the expression of genes are still a topic of intense research. Due to the central role of transcription factors in the regulation of gene expression, one possible mechanism is that genetic variants act on transcription by reshaping regulatory elements in the promoters of genes. We hypothesise that if alteration of regulatory elements is underlying the observed phenotypic variation, then there should be an enrichment of significantly associated genetic variants in regulatory elements. Here, we examined, whether genetic variants act on transcription by altering TFBSs. For this, we studied the enrichment of significant associations in the TFBSs that we predicted.

We obtained significant genetic associations from an expression QTL (eQTL) study that determined the genetic variants that affected gene expression in the MAGIC founder strains [52]. For our enrichment analysis we only considered genetic variants where we could localise the association in the promoter regions. For this we restricted the set of genetic variants from eQTL study in the following way:

1. We only considered genetic variants where the genome was reliable for all 19 MAGIC founder strains, i.e. that were not uncovered in any of the strains.
2. We only considered variants in genes, where there was a strong association, compared to the associations in a 30 kb window around the TSS that excludes the promoter, was in first 1 kb of promoter regions. More specifically, if $p_{1\,kb}$ is the p-value of the strongest association in the promoter and $p_{30\,kb}$ is the strongest association in a 30 kb window around the TSS that excludes the promoter, then we only included variants in our analysis for which $0.9 * \log_{10}(p_{30\,kb}) > \log_{10}(p_{1\,kb})$.
3. We only considered genetic variants in promoters with no more than five genetic variants fulfilling criteria (1) and (2).

This resulted in a set of $8,063$ variants in $2,065$ genes that we used for our analysis.

Next, we examined the dependence between the significance of the association and the enrichment in TFBSs. For this, we computed for all p-value cutoffs the fraction of significant genetic variants that overlapped with predicted TFBS. We, furthermore, estimated the enrichment that would be expected only due to the common enrichment of significant associations and the TFBSs near the TSS [52]. We estimated the enrichment that is expected by chance by computing for all p-value cutoffs the fraction of significant genetic variants that overlapped TFBS when the promoters are permuted among genes, i.e. computing the enrichment of significant associations in TFBS of a randomly chosen promoter. For the estimate, we further computed the variance using a bootstrapping approach with 100 permutations.

We then analysed, how the enrichment depends on the significance of the observations. We found that as the associations became more significant they were stronger enriched in TFBSs (see Fig. 5.6). For the 10 most significant associations we found that more than 40% were overlapping TFBSs. The enrichment that we observed for highly significant associations ($p \leq 10^{-10}$) was significantly higher ($p \leq 0.05$) than expected by chance. This high enrichment suggest that many of the highly significant genetic variants act on gene expression by disrupting TFBS.

An example of how a genetic variant can act on gene expression is shown in Fig. 5.7. In the promoter of the AT5G47250 gene we found a SNP that changed in two strains a guanine (G) to an adenine (A). This lead to a loss of a TFBS of VNI2, a transcription factor that is known to act negatively on gene expression (see e.g. [189]). Interestingly, the strains that have lost

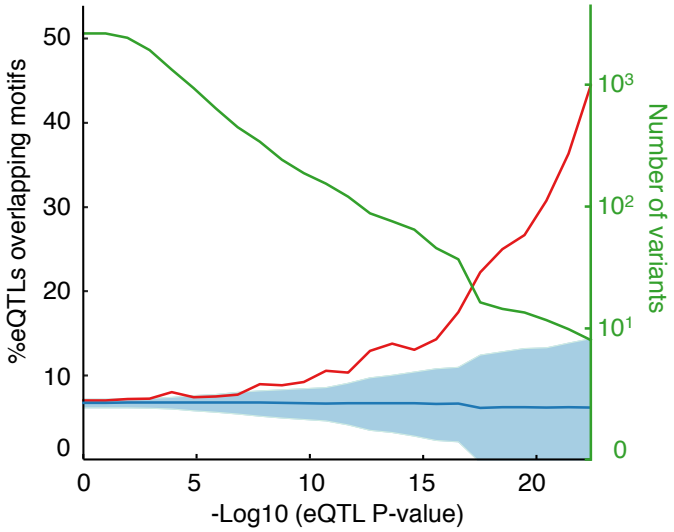

Figure 5.6.: Shown is the overlap of significant associations in TFBS. The enrichment for different significance cutoffs is shown in red. The number of significant associations for a given significance cutoff is shown in green. The blue line indicates the enrichment that is expected by chance and the shaded blue area indicates the 95% quantile of this estimate. This figure has been adapted from our publication [185].

this TFBS show an increased gene expression compared to the strains with the reference allele. This example suggested that TFBSs could also be useful in obtaining a functional annotation of genetic variants. This could provide an interpretation of possible effects of genetic variants and therefore help in identifying and thereby can help identifying causal variants.

In summary, we have shown that significant associations are enriched in TFBS. For strongly significant association more than 40% were overlapping TFBSs. This suggests that indeed alteration of TFBS is a main mechanisms of promoting phenotypic variance. Furthermore, we have shown that TFBS-information can help to understand the mechanisms by which genetic variation act on gene expression as shown for the gene AT5G47250.

Variance Decomposition

In the previous section, we have showed that significant associations are enriched in TFBSs, indicating that genetic variants act on gene expression by changing the transcriptional regulation. This observation motivates the question, what fraction the total variance of gene expression (*total variance*) can be explained by genetic variation in the promoter region and what fraction is caused by alterations of TFBSs.

To investigate this, we formulated two models for the gene expression variance: One that used only information from genetic variants TFBSs in promoters (affinity variance model) and a standard model that used information from all genetic variants in the promoters (promoter variance model). Specifically, the first model describes the total variance using the affinity covariance of the promoters and a noise (see Sec. 5.2.6). This model effectively uses on average only 8.4% of the promoter positions. The second model considers all positions in the promoter as equally informative. This model describes the total variance by the promoter covariance

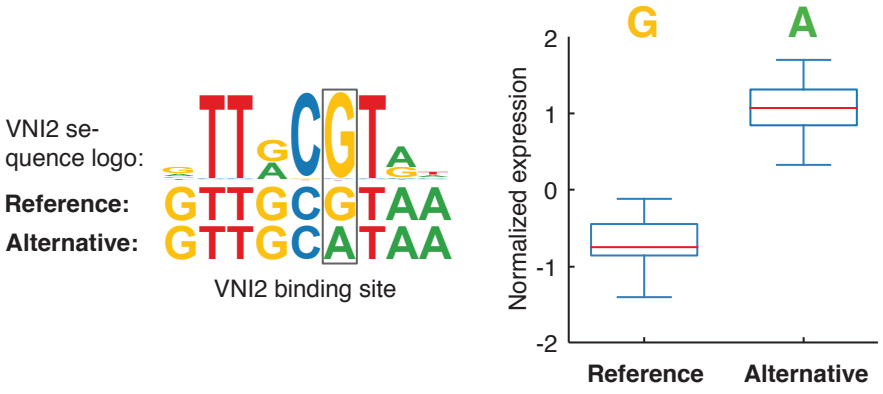

Figure 5.7.: Shown on the left is the sequence logo of the transcription factor VNI2. Shown below the logo is its binding site for the two alleles. Shown on the right is a boxplot of the gene expression of AT5G47250 according to these two alleles. This figure has been adapted from our publication [185].

and a noise (see Sec. 5.2.6).

We fitted these two models for all 14,362 genes of the 208 MAGIC lines (see Sec. 5.2.1) that had no uncovered region their promoters. The gene expressions estimates that were used to fit these models were obtained as described in Sec. 5.2.2.

In order to compare how informative the two models are for the gene expression variation we first computed their predictive power. For this, we determined how well the two models can discriminate, based on the fraction of the total variance they explain, genes with a known genetic *cis*-association in a 30 kb window around their TSS from genes without association. The associations for this experiment, were obtained from an association study that used the same MAGIC lines and gene expression estimates as we did (pers. comm. Oliver Stegle). For the comparison, we determined four sets of genes with significant associations; those with associations that had a p-value less than $p < 10^{-5}$ ($n = 2,755$), $p < 10^{-10}$ ($n = 1,064$), $p < 10^{-15}$ ($n = 516$) and $p < 10^{-20}$ ($n = 292$). Using these four sets of genes we then compared for the two models the enrichment of genes with an association in the genes where the largest fraction of variance could be explained. To quantify the enrichment we computed for both variance models and for each of the four significances thresholds the *area under the precision recall curve* (auPRC), a measure of the predictive power that is commonly used for unbalanced datasets. For the computation we defined the significant genes as the positive sets and the remaining genes as the negative set.

We found a higher auPRC for the affinity variance model than for the promoter affinity model for genes with a p-value that is smaller than 10^{-10}, whereas for less significant p-values the opposite was true (see Tab. 5.3). Furthermore, we noted that the precision for low recall thresholds (recall≤ 0.1) was always higher for the affinity model than for the promoter model (see Fig. 5.8). This shows that the affinity variance model is informative for predicting the gene expression variance. Since the number of positions that are considered in the affinity variance model is almost 12 times less than the number of positions that are considered in the promoter variance model, this also suggest that genetic variants in TFBSs are highly informative for predicting gene expression changes.

Table 5.3.: Area under the precision recall curve (auPRC) for the promoter variance model and the affinity variance model.

Variance model	auPRC			
	$p < 10^{-5}$	$p < 10^{-10}$	$p < 10^{-15}$	$p < 10^{-20}$
Promoter	**0.75**	**0.54**	0.36	0.28
Affinity	0.65	0.52	**0.40**	**0.33**

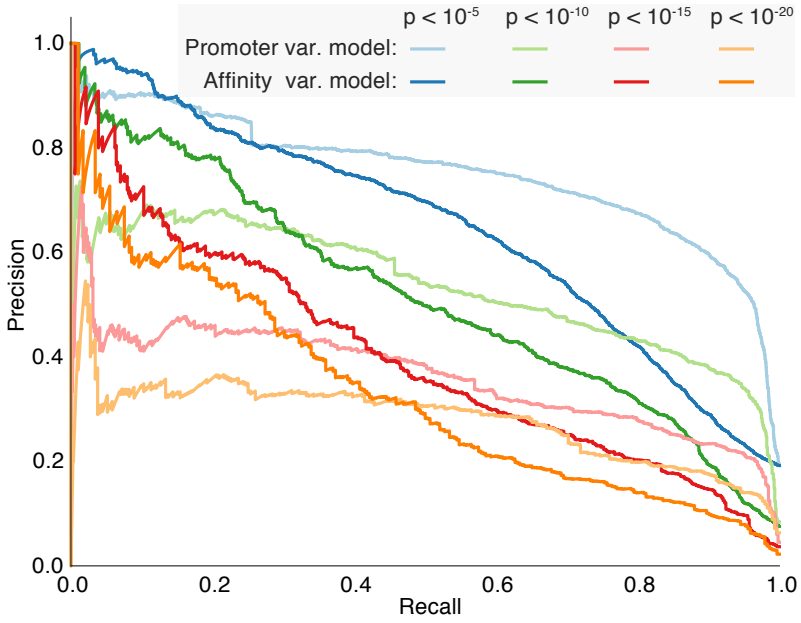

Figure 5.8.: Shown are the precision recall curves (auPRC) for the promoter variance model and the affinity variance model for four different significance thresholds. For the computation of the auPRC, genes with an association that had a p-value less than $p < 10^{-5}, 10^{-10}, 10^{-15}$ and 10^{-20} were considered as positive set and the other genes as negative sets.

After investigating the predictive power of the models, we compared the fraction of the total variance that could be explained by the two models. We found that both, the promoter variance model and the affinity variance model could explain less than 10% of the total variance for the majority of genes (for 10,198 (71.0%) respectively 11,361 (79.1%) genes). Overall, the average percentage of the total variance that could be explained for all genes was similar for the promoter variance model (11.1%) and the affinity variance model (10.8%).

When we compared, however, the distributions of the fraction of total variance explained, we found that they differed substantially between the two models (see Fig. 5.9). The density for the promoter variance model was decreasing for an increasing fraction of variance explained. In contrast, the density for the affinity variance model was bimodal, with one mode at 0 and the other close to 1. This was reflected by the higher fraction of variance that could be explained by the affinity variance model for the genes where at least 10% of the total variance was explained. We found that for these genes, the average percentage of the explained total variance was 47.1% for the affinity variance model and only 33.7% for the promoter variance model. We also observed that the number of genes for which a large percentage ($> 90\%$) of the total variance can be explained differed: The affinity variance model could explain for only 426 genes more than this fraction of the variance, whereas the promoter variance model could explain this for 106 genes.

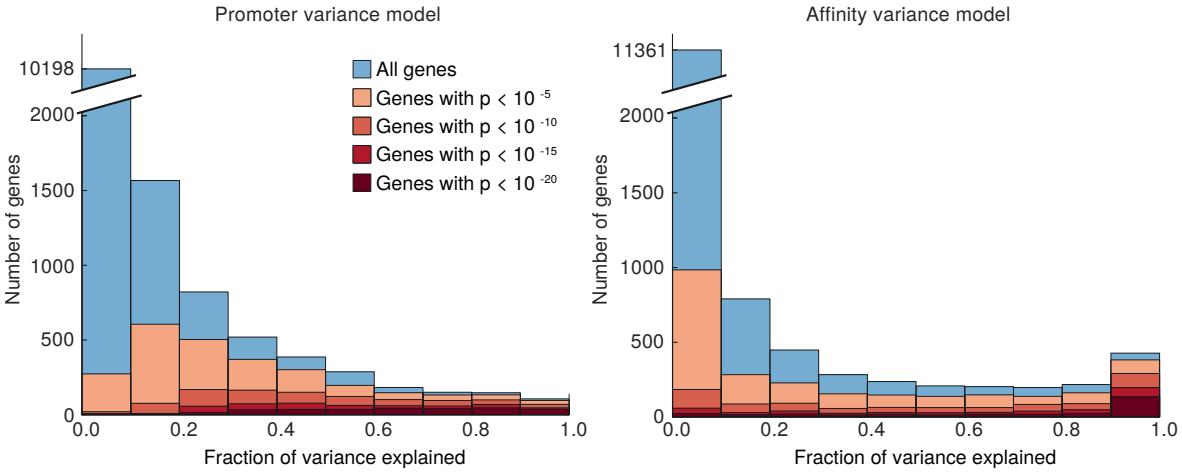

Figure 5.9.: Shown is the distribution of the fraction of the explained total variance. This is shown on the left for the promoter variance model and on the right for the affinity variance model. Shown in different shades of red is the distribution of the fraction of the total variance that is explained for genes with significant associations.

To further investigate the differences between the two models, we compared how they explain the variance of genes with a known genetic *cis*-association in a 30 kb window around their TSSs. For this analysis, we used again the sets of genes having at least one association with a p-value small than $p < 10^{-5}$, $p < 10^{-10}$, $p < 10^{-15}$ and $p < 10^{-20}$. For these variants, we found that the affinity variance model explained more than 90% of the variance for 134 (45.0%) of 292 genes with a p-value smaller than 10^{-20} (see Fig. 5.9). This number is in accordance with the number that we have observed for the association enrichment in Sec. 5.3.2. In contrast, the promoter affinity model could only explain that fraction of the variance for 36 (12.7%) of these significant genes. Overall, we observed that the fraction of genes with an association that had more than 50% respectively 90% of the total variance explained was higher for the

affinity variance model than for the promoter variance model (see Tab. 5.4). The opposite was only the case for the percentages of genes that had more than 10% of the total variance explained.

Table 5.4.: Percentage of genes with a significant association for different fractions of variance explained. The percentages are shown for the affinity variance model and the promoter variance model.

Significance	Method	Fraction of variance explained		
		> 0.1	> 0.5	> 0.9
$p < 10^{-5}$	Promoter	**90.1%**	25.4%	3.4%
	Affinity	64.3%	**35.0%**	**13.8%**
$p < 10^{-10}$	Promoter	**98.1%**	45.7%	6.5%
	Affinity	82.8%	**55.0%**	**27.6%**
$p < 10^{-15}$	Promoter	**98.6%**	57.0%	8.9%
	Affinity	88.8%	**65.3%**	**37.8%**
$p < 10^{-20}$	Promoter	**99.7%**	70.2%	12.7%
	Affinity	92.5%	**70.9%**	**45.9%**

Overall, this showed that the fraction of variance that can be explained by alterations of TFBSs is in the same order as the fraction that can be explained when considering the entire promoter. Since the number of positions that are considered when modelling the variance in the affinity variance model is much smaller than the number considered for the promoter variance model, this suggests that a large extent of the gene expression variation in *A. thaliana* that is caused by variations in the promoter, is due to alterations of TFBSs. Moreover, the fact that a large fraction of the total variance could be explained by the variance affinity model for many genes that have a significant association in a 30 kb window around the TSS, suggest that in general a large extent of the *cis*-variation in *A. thaliana* is caused by alterations of TFBSs. Since we used only a subset of all transcription factors of *A. thaliana* in our analysis and also did not consider tissue specific binding, we believe that these estimates of the fraction of the explained total variance is likely an underestimation of the true extent.

5.4. Summary

In this chapter, we have studied the gene expression in a natural population of *A. thaliana*. For this, we have developed a robust quantification approach that accounts for the various effects of genetic variants can have on gene expression estimation. We have shown that this approach provides stable quantifications of gene expression and thus allowed getting a clear picture of gene expression variation and the regulators of gene expression. Using the quantification strategy, we have performed a comprehensive analysis of the expressed genes, how gene expression varies in the population and discussed potential evolutionary constraints of the *A.thaliana* population.

Furthermore, we have investigated the role that the transcriptional machinery plays in translating genetic variability into changes in gene expression. For this, we have used an extensive collection of transcription factor binding profiles in order to understand how TFBSs are distributed in the promoter regions of genes. By doing this, we obtained a description of TF binding preferences in *A. thailana*. We have, furthermore, shown that this information could be used in order to get a functional interpretation of association studies and that it can be used to get an explanation for the gene expression changes. Additionally, we have shown that alterations of TFBSs explain the variance better than the commonly used similarity of the promoter regions for genes with strongly significant associations. In summary we have shown that the integration of transcription factor binding information and genetic variation is a promising approach in revealing the mechanisms that underly phenotypic variation.

6. Conclusion

In this thesis, we have derived novel approaches to detect and characterise changes in RNA processing using high-throughput sequencing data, by applying state-of-the-art techniques from machine learning, statistics and bioinformatics.

We have first established how information can be extracted from RNA-Seq data for tasks such as transcript identification or detection of differential splicing (Ch. 3). Based on these insights, we have formulated for the identification of transcripts and the detection of differential splicing two probabilistic models to systematically assess the influences of various experimental parameters on the expected information gain of a particular experiment. Application of our models provided insights into the optimal choice of experimental parameters in order to maximise the utility of the experiment. The application of our models further revealed limitations of the commonly used experimental design and consequently we have proposed alternative designs to remedy these shortcomings.

Based on the insights obtained from modelling the information gain of RNA-seq experiments, we have derived statistical tests to detect various aspects of differential post-transcriptional RNA processing from high-throughput sequencing data (Ch. 4). These include the detection of changes in alternative splicing, secondary structure or translation. Furthermore, we have derived a statistical test that allows associating changes in RNA processing to genetic variants. To prove that our methods are of practical relevance, we have evaluated our methods on realistically simulated as well as on experimental data and showed that they outperform various existing state-of-the-art methods. Additionally, we have applied our methods to study different aspects of RNA processing such as alternative splicing, RNA degradation and translation.

Finally, we have characterised the variation of gene expression in an *A. thaliana* population (Ch. 5). To this end, we have combined information on gene expression, genetic variation and transcription factor binding preferences in a linear mixed model to describe the variation in gene expression that is due to alterations in regulatory elements. By doing this, we were able to show that a large fraction of genetic variants that can be linked to changes in gene expression act by perturbing regulatory elements in promoters.

The contributions of this thesis show that while the large amount of data generated in current biological research poses challenges for its analysis, it also allows to generate and validate general hypotheses on the regulation of RNA processing and thus can inspire new insights into the functioning of cells. We have demonstrated that approaches that combine ideas from bioinformatics, machine learning and statistics can be utilised to tap the full potential of the increasing amount of available data for understanding the fundamentals of RNA regulation. There are, however, still some limitations and open problems that we have not yet had the time to address. In the following we will discuss some of these:

Statistical models for count data The discrete Poisson or negative binomial distributions have been shown to more accurately describe RNA-Seq data than continuous Gaussian distributions. We, therefore, have used these two discrete read distributions in this work to derive statistical tests for differential RNA processing. However, the theoretical framework for working with these distributions is less developed than for Gaussian dis-

tributions. As a consequence, fitting complex models is non-trivial and computationally challenging. Therefore, these distributions are restricted to modelling problems of modest complexity. This motivates research on efficient fitting and inference techniques for the Poisson or negative binomial distributions. This extension to more complex applications would thus allow the construction of more accurate models of RNA processing.

Integration of heterogeneous data Besides deriving better techniques to work with discrete data another point that would advance understanding of RNA processing is the integration of heterogeneous data. In this thesis we have shown, for example, that an integrative model of gene expression, genetics and transcription factor binding reveals how genetic variants act on gene expression through modification of transcriptional regulation. In the light of the amount of new high-throughput analysis techniques that have been and still are published (e.g., to investigate polyadenylation, RNA secondary structure or protein occupancy, to name only a few), general approaches to jointly analyse these distinct information sources therefore promise to provide novel insights into RNA processing. As we have shown in Ch. 3, determining appropriate representations of the data is non-trivial and time-consuming. With the increasing number of different data types this is likely to become a bottleneck. Therefore, for integrative approaches to be successful in general, it will be necessary to develop methods that can learn representations for these different data types. For this purpose, a nonparametric extension of our RKHS embeddings (Ch. 4) may be a viable option.

Structured data In most parts of this thesis we have considered the typical case-control type of experiments, where two conditions were compared. This approach is promising when a particular biochemical process is investigated under controlled conditions. However, in other applications the assumption of having only two conditions is too restrictive and does not allow capturing the full complexity of the experiment. For example, complex experimental designs such as time series cannot be represented in this setup. In addition, this setup does also not allows taking into account the structure of the samples being studied (such as population structure, batch effects or other phenomena that confound the analysis). Therefore, there is a need for approaches that allow the analysis of more complex experimental setups and are able to account for different confounding factors.

Quantitative models for RNA processing Our methods to detect changes in RNA processing have been shown to be particularly useful for exploratory data analysis. In particular, we have shown how these methods can be used to reveal basic mechanisms of RNA processing regulation and provide a qualitative assessment of their effect sizes. In order to obtain a characterisation of complex changes, however, our models are of limited use. To address this limitation of our methods we, therefore, suggest further research on quantitative models of RNA processing. This could be done for example by developing predictive models for RNA regulation, based on features such as the RNA sequences and molecular markers. These models would help to derive testable hypotheses and to get a quantitative understanding of the regulatory mechanisms and their functioning.

Overall, we have developed new approaches to analyse RNA-Seq data. Our new methods have enabled novel insights into the regulation of RNA processing, have identified several open questions and, thereby, stimulated future studies.

A. Appendix

A.1. Theorems

Lemma A1. : Let $\mathcal{H}$ be a reproducing kernel Hilbert space that is induced by the kernel $k(\cdot,\cdot)$. Assume furthermore that four distributions V^A, V^B, S^A, S^B are given and let $\mu_{V^A}, \mu_{V^B}, \mu_{S^A}, \mu_{S^B}$ be their mean embedding in $\mathcal{H}$. Then a kernel expansion for sDiff is given by:

$$\begin{aligned}
\text{sDiff}(\mu_{V^A}, \mu_{V^B}, \mu_{S^A}, \mu_{S^B})^2 = \mathbb{E}_{v^A, v'^A \sim V^A,} k(v^A, v'^A) - 2\,\mathbb{E}_{v^A \sim V^A,\, s^A \sim S^A} k(v^A, s^A) \\
+ \mathbb{E}_{s^A, s'^A \sim S^A} k(s^A, s'^A) + \mathbb{E}_{v^B, v'^B \sim V^B} k(v^B, v'^B) \\
- 2\,\mathbb{E}_{v^B \sim V^B,\, s^B \sim S^B} k(v^B, s^B) + \mathbb{E}_{s^B, s'^B \sim S^B} k(s^B, s'^B) \\
+ 2[\mathbb{E}_{v^A \sim V^A,\, s^B \sim S^B} k(v^A, s^B) - \mathbb{E}_{v^A \sim V^A,\, v^B \sim V^B} k(v^A, v^B) \\
- \mathbb{E}_{s^A \sim S^A,\, s^B \sim S^B} k(s^A, s^B) + \mathbb{E}_{v^B \sim V^B,\, s^A \sim S^A} k(v^B, s^A)]
\end{aligned}$$

Proof: We take the square of sDiff and use a linear expansion of the calar product. This yields:

$$\begin{aligned}
\text{sDiff}(\mu_{V^A}, \mu_{V^B}, \mu_{S^A}, \mu_{S^B})^2 &= \|(\mu_{V^A} - \mu_{S^A}) - (\mu_{V^B} - \mu_{S^B})\|^2_{\mathcal{H}} \\
&= < (\mu_{V^A} - \mu_{S^A}) - (\mu_{V^B} - \mu_{S^B}), \\
&\quad (\mu_{V^A} - \mu_{S^A}) - (\mu_{V^B} - \mu_{S^B}) >_{\mathcal{H}} \\
&= < \mu_{V^A}, \mu_{V^A} >_{\mathcal{H}} - 2 < \mu_{V^A}, \mu_{S^A} >_{\mathcal{H}} + < \mu_{S^A}, \mu_{S^A} >_{\mathcal{H}} \\
&\quad + < \mu_{V^B}, \mu_{V^B} >_{\mathcal{H}} - 2 < \mu_{V^B}, \mu_{S^B} >_{\mathcal{H}} + < \mu_{S^B}, \mu_{S^B} >_{\mathcal{H}} \\
&\quad + 2(< \mu_{V^A}, \mu_{S^B} >_{\mathcal{H}} - < \mu_{V^A}, \mu_{V^B} >_{\mathcal{H}} \\
&\quad - < \mu_{S^A}, \mu_{S^B} >_{\mathcal{H}} + < \mu_{V^B}, \mu_{S^A} >_{\mathcal{H}}) \\
&= \mathbb{E}_{v^A, v'^A \sim V^A,} k(v^A, v'^A) - 2\,\mathbb{E}_{v^A \sim V^A,\, s^A \sim S^A} k(v^A, s^A) \\
&\quad + \mathbb{E}_{s^A, s'^A \sim S^A} k(s^A, s'^A) + \mathbb{E}_{v^B, v'^B \sim V^B} k(v^B, v'^B) \\
&\quad - 2\,\mathbb{E}_{v^B \sim V^B,\, s^B \sim S^B} k(v^B, s^B) + \mathbb{E}_{s^B, s'^B \sim S^B} k(s^B, s'^B) \\
&\quad + 2[\mathbb{E}_{v^A \sim V^A,\, s^B \sim S^B} k(v^A, s^B) - \mathbb{E}_{v^A \sim V^A,\, v^B \sim V^B} k(v^A, v^B) \\
&\quad - \mathbb{E}_{s^A \sim S^A,\, s^B \sim S^B} k(s^A, s^B) + \mathbb{E}_{v^B \sim V^B,\, s^A \sim S^A} k(v^B, s^A)]
\end{aligned}$$

□

Lemma A2. : Let $V^A = \{v_A^1, \cdots, v_A^{n_V^A}\}, V^B = \{v_B^1, \cdots, v_B^{n_V^B}\}, S^A = \{s_A^1, \cdots, s_A^{n_s^A}\}$ and $S^B = \{s_B^1, \cdots, s_B^{n_s^B}\}$ be four sets of reads that contain $n_V^A, n_V^B, n_S^A, n_S^B$ reads. Let furthermore $\mathcal{H}$ be a reproducing kernel Hilbert space that is induced by the kernel $k(\cdot,\cdot)$ and denote by $\mu_{V^A}, \mu_{V^B}, \mu_{S^A}, \mu_{S^B}$ the four embeddings of the reads sets in $\mathcal{H}$. Then an unbiased estimator

A. Appendix

$\widehat{\text{sDiff}}$ for sDiff is given by:

$$\begin{aligned}
\widehat{\text{sDiff}}(\mu_{V^A}, \mu_{V^B}, \mu_{S^A}, \mu_{S^B})^2 = &\frac{1}{n_V^A(n_V^A-1)} \sum_{i,j=1, i\neq j}^{n_V^A} k(v_A^i, v_A^j) - \frac{2}{n_V^A n_S^A} \sum_{i,j=1}^{n_V^A, n_S^A} k(v_A^i, s_A^j) \\
&+ \frac{1}{n_S^A(n_S^A-1)} \sum_{i,j=1, i\neq j}^{n_S^A} k(s_A^i, s_A^j) + \frac{1}{n_V^B(n_V^B-1)} \sum_{i,j=1, i\neq j}^{n_V^B} k(v_B^i, v_B^j) \\
&\frac{2}{n_V^B n_S^B} \sum_{i,j=1}^{n_V^B, n_S^B} k(v_B^i, s_B^j) + \frac{1}{n_S^B(n_S^B-1)} \sum_{i,j=1, i\neq j}^{n_S^B} k(s_B^i, s_B^j) \\
&+ 2\Big(\frac{1}{n_V^A n_S^B} \sum_{i,j=1}^{n_V^A, n_S^B} k(v_A^j, s_B^j) - \frac{1}{n_V^A n_V^B} \sum_{i,j=1}^{n_V^A, n_V^B} k(v_A^j, v_B^j) \\
&\quad - \frac{1}{n_S^A n_S^B} \sum_{i,j=1}^{n_S^A, n_S^B} k(s_A^j, s_B^j) + \frac{1}{n_V^B n_S^A} \sum_{i,j=1}^{n_V^B, n_S^A} k(v_B^j, s_A^j)\Big)
\end{aligned}$$

Proof: The proof is analogous to the proof of ([62], Lemma 6). The estimator $\widehat{\text{sDiff}}$ can be obtained by replacing the population expectations in Lemma A1 with their corresponding U-statistics and sample averages.

$$\begin{aligned}
\widehat{\text{sDiff}}(\mu_{V^A}, \mu_{V^B}, \mu_{S^A}, \mu_{S^B})^2 &= \|(\mu_{V^A} - \mu_{S^A}) - (\mu_{V^B} - \mu_{S^B})\|_{\mathcal{H}}^2 \\
&= \frac{1}{n_V^A(n_V^A-1)} \sum_{i,j=1, i\neq j}^{n_V^A} k(v_A^i, v_A^j) - \frac{2}{n_V^A n_S^A} \sum_{i,j=1}^{n_V^A, n_S^A} k(v_A^i, s_A^j) \\
&\quad + \frac{1}{n_S^A(n_S^A-1)} \sum_{i,j=1, i\neq j}^{n_S^A} k(s_A^i, s_A^j) + \frac{1}{n_V^B(n_V^B-1)} \sum_{i,j=1, i\neq j}^{n_V^B} k(v_B^i, v_B^j) \\
&\quad \frac{2}{n_V^B n_S^B} \sum_{i,j=1}^{n_V^B, n_S^B} k(v_B^i, s_B^j) + \frac{1}{n_S^B(n_S^B-1)} \sum_{i,j=1, i\neq j}^{n_S^B} k(s_B^i, s_B^j) \\
&\quad + 2\Big(\frac{1}{n_V^A n_S^B} \sum_{i,j=1}^{n_V^A, n_S^B} k(v_A^j, s_B^j) - \frac{1}{n_V^A n_V^B} \sum_{i,j=1}^{n_V^A, n_V^B} k(v_A^j, v_B^j) \\
&\qquad - \frac{1}{n_S^A n_S^B} \sum_{i,j=1}^{n_S^A, n_S^B} k(s_A^j, s_B^j) + \frac{1}{n_V^B n_S^A} \sum_{i,j=1}^{n_V^B, n_S^A} k(v_B^j, s_A^j)\Big)
\end{aligned}$$

Following the argumentation in [62], this shows that $\widehat{\text{sDiff}}$ is an unbiased estimator of sDiff. □

A.2. Tables

Table A.1.: Area under the ROC-curve in the interval $[0, 0.2]$ (auROC20) and auROC for rDiff, CuffDiff and MISO. The comparison is shown on the two artificial dataset with a small and large biological variance.

Method	auROC20		auROC	
	small biological variance	large biological variance	small biological variance	large biological variance
rDiff.nonparametric	0.077	0.073	0.686	0.677
rDiff.parametric	0.101	0.093	0.763	0.734
rDiff.poisson	0.099	0.082	0.752	0.719
rDiff.mmd	0.062	0.054	0.652	0.627
CuffDiff	0.085	0.055	0.669	0.585
MISO	0.089	0.061	0.692	0.614

Table A.2.: GO-term enrichment. GO-terms tat are written in Bold are defense related. This table has been adapted from our publication [52].

GO-Term	P-value	Number of genes associated with GO-term
Defense response	8.05×10^{-15}	68
Response to stress	5.50×10^{-13}	87
Response to stimulus	8.69×10^{-11}	168
Apoptosis	9.53×10^{-9}	34
Programmed cell death	1.54×10^{-8}	35
Cell death	2.16×10^{-8}	35
Death	2.16×10^{-8}	35
Immune system process	1.04×10^{-6}	26
Immune response	3.00×10^{-6}	25
Innate immune response	3.00×10^{-6}	25
Response to biotic stimulus	5.68×10^{-6}	29
Multi-organism process	6.30×10^{-6}	27
Response to other organism	1.14×10^{-5}	26
S-glycoside metabolic process	8.58×10^{-5}	5
Glucosinolate metabolic process	8.58×10^{-5}	5
Glycosinolate metabolic process	8.58×10^{-5}	5
Defense response to fungus	5.74×10^{-4}	7
Response to fungus	5.74×10^{-4}	7

Table A.3.: auPRC for two models for genes where the fraction of explained variance $> 10\%$ in both models

Variance model	auPRC			
	$p < 10^{-5}$	$p < 10^{-10}$	$p < 10^{-15}$	$p < 10^{-20}$
Promotor	**0.86**	0.62	0.41	0.31
Affinity	**0.86**	**0.63**	**0.46**	**0.37**

A.3. Figures

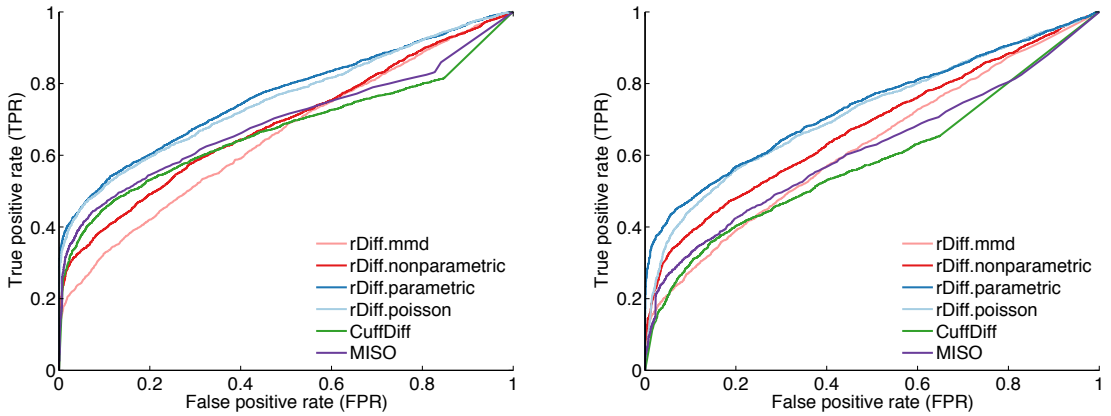

Figure A.1.: ROC curves for rDiff, CuffDiff and Miso for the dataset described in Sec. 4.2.7. Shown on the left are the ROC curves for the dataset with small biological variance and on the right for the dataset with the larger biological variance.

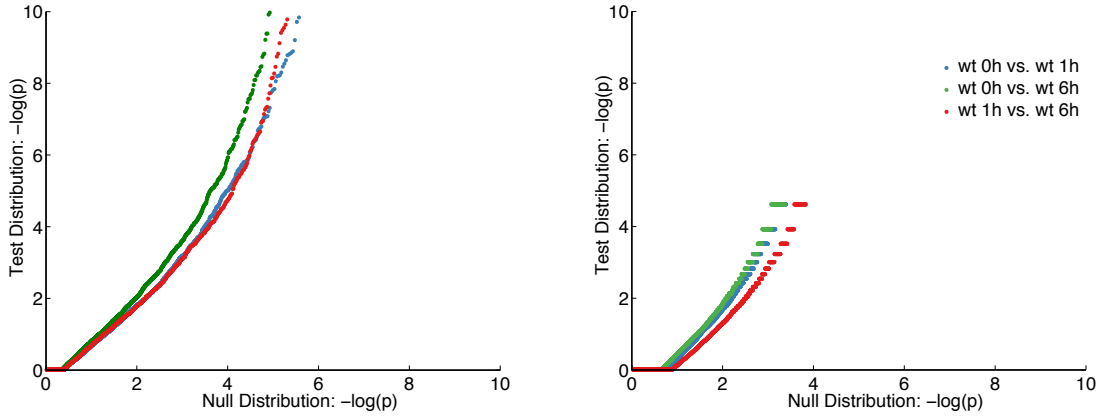

Figure A.2.: QQ plots for rDiff on the dataset described in Sec. 4.2.8.

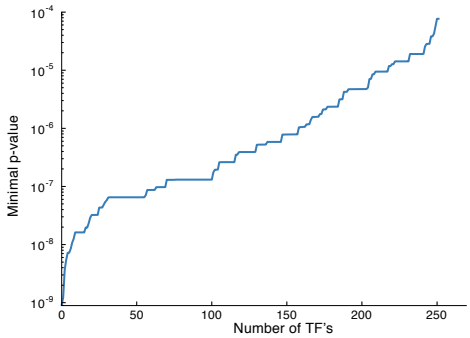

Figure A.4.: Shown is the cumulative distribution of minimal p-values of different TFs.

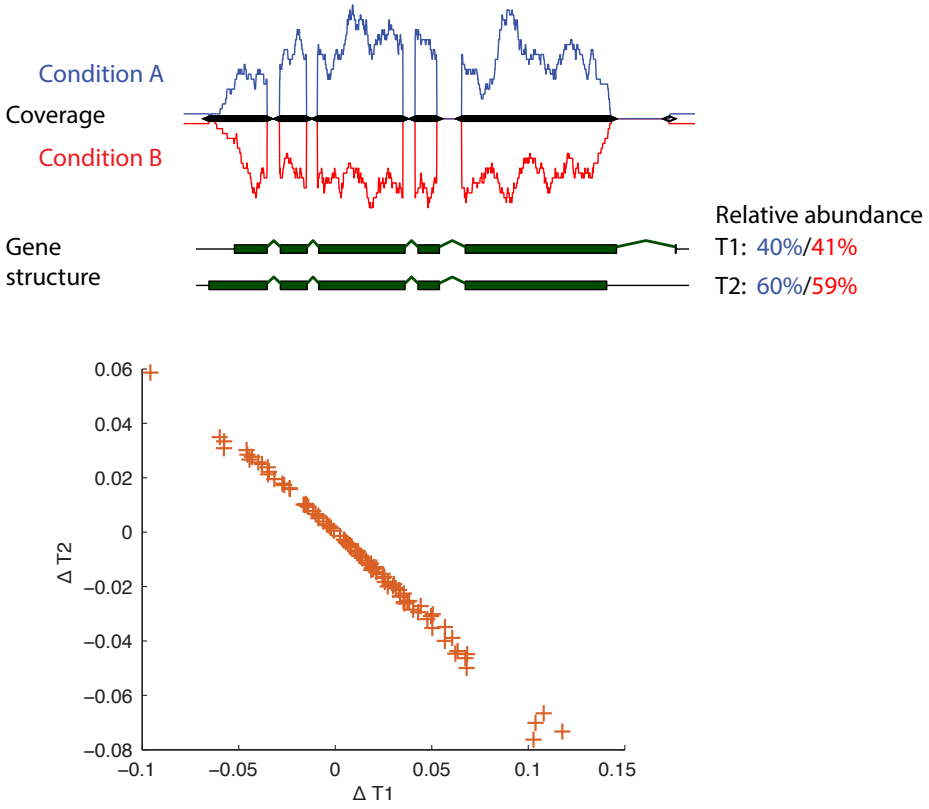

Figure A.3.: Error amplification of quantification. Shown on top are the coverages of a gene with two transcripts T1 and T2 in two conditions A (red) and B (blue). Between the two conditions the differences of the relative transcript abundance of T1 and T2 is 1%. Shown below is the delta-plot of the transcript quantifications when simulating these relative abundances 100 times. It can be seen that the quantifications change up to 10%. This plot has been adapted from a personal communication of Oliver Stegle.

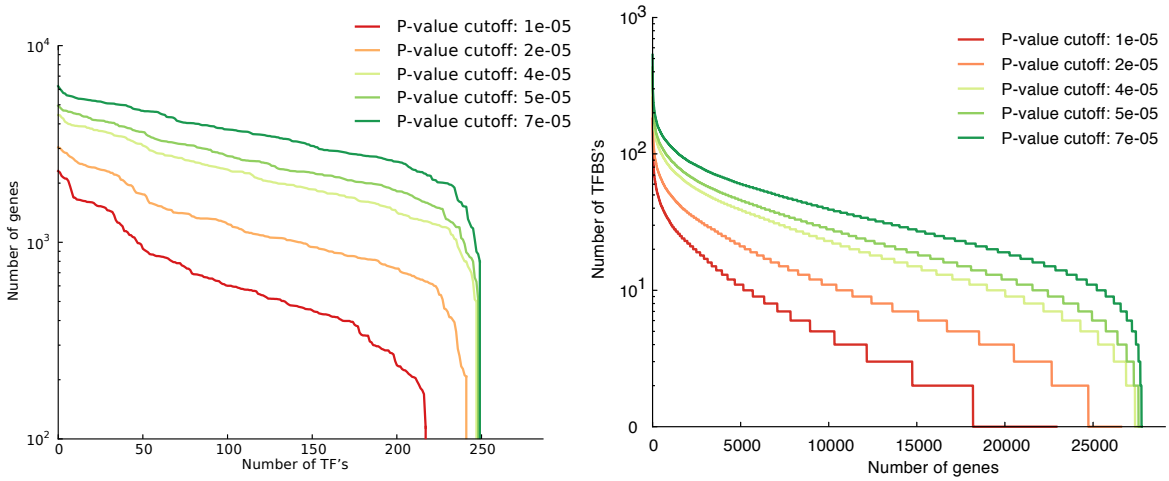

Figure A.5.: TFBS statistics for different minimal p-values of TFBSs. Shown on the left is the cumulative distribution function of the number genes that are regulated by each transcription factor. Shown on the right is the cumulative distribution function of the number of TFs that bind the promoters.

A.3.1. Combinations of kernels

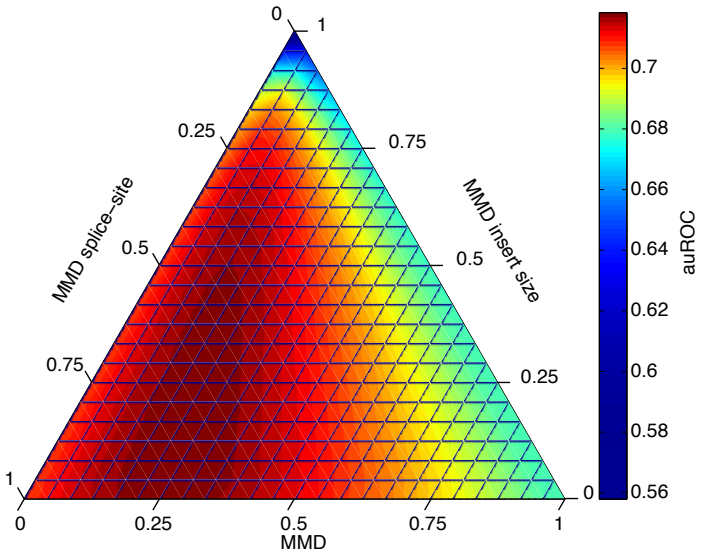

Figure A.6.: Ternary plot of the performance (auROC) of linear combinations of MMD, MMD one splice sites and MD on insert sizes.

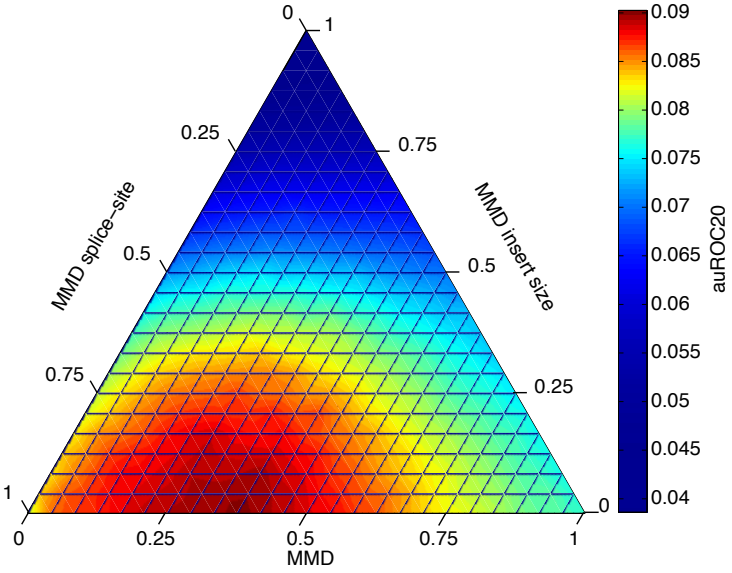

Figure A.7.: Ternary plot of the performance (auROC20) of linear combinations of MMD, MMD one splice sites and MD on insert sizes.

A.3.2. Running rDiff

Requirements

rDiff requires a UNIX environment as well as the following packages:

- Octave [128] (64 bit version, version 3.4 or higher) or Matlab [112] (version 7.6 or higher)
- Python (version 2.6.5 or higher)
- Scipy [82] (version 0.7.1 or higher)
- SAMTools [99] (version 0.1.7 or higher)
- wget

Command line options

The rDiff standalone software package provides many options to adapt the parameters to the task at hand. The available parameters are:.

-h Display the help

-o This option takes as argument the output directory where the input files are saved. This is also where rDiff will save the other output files.

-d This specifies the directory where the bam-files are located. If they are in different directories this can be also / . The the path to the BAM files can then be given as part of the bam-file names.

-a This argument specifies which sample should be used for sample 1. It takes as argument a comma-separated list of bam-files for sample 1. It is important not to have spaces between the files. The input should be of the form: File1.bam,File2.bam,...

-b This argument specifies which sample should be used for sample 2. It takes as argument a comma-separated list of bam-files for sample 2. It is important not to have spaces between the files. The input should be of the form: `File1.bam,File2.bam,...`

-g Path to GFF3 gene structure

-L Read length used for rDiff.parametric to compute the alternative regions. The default is 75 bp. For reads that are longer or shorter than the specified length rDiff will try to find the best match to an alternative region.

-m This option takes as argument the method that should be used for testing. The options for this parameter are:

param for rDiff.parametric (default)

nonparam for rDiff.nonparametric

poisson for rDiff.poisson

mmd for MMD

-M Minimal read length required for a red to be considered. The default is 30 bp. The reads that are shorter are not used for the analysis.

-e Skip the gene expression estimation. If the gene expression estimation step should be skipped enter 0. The default is 1.

-E Only estimate the gene expression and variance function estimation but do not perform testing. If you want to exit after the variance function estimation enter 0. The default is 1.

-A This parameter takes as argument the path to variance function for sample 1. This option can be used for example, if a previously computed variance function should be used.

-B This parameter takes as argument the path to variance function for sample 2. This option can be used for example, if a previously computed variance function should be used.

-S Filename under which variance function for sample 1 will be saved.

-T Filename under which variance function for sample 2 will be saved.

-P Using this option, one can specify a given parametric variance function for sample 1 of the form $f(x) = a + b * x + b * x^2$. The argument for this option is `a,b,c`.

-Q Using this option one can specify a given parametric variance function for sample 2 of the form $f(x) = a + b * x + b * x^2$. The argument for this option is `a,b,c`.

-y This parameter allows to use only the gene start and stop for the rDiff.nonparametric variance function estimation. Enter 1 if this should be done and 0 otherwise.

-s This option allows to sample the reads down to a certain number. This increases the speed for highly covered genes. The argument is number of reads per gene to which should be sampled. The default is $10,000$.

-C Number of bases to clip from each end of each read. This reduces the false mappings of spliced read ends. The default is 3 bp.

-p Number of permutations performed for rDiff.nonparametric. The default is $1,000$.

-x Merge sample 1 and sample 2 for variance function estimation. Type 1 to merge the samples. The default is 0.

Examples

rDiff.parametric

When the gene structure is known we recommend using rDiff.parametric. This method tests for difference in the relative abundance of annotated transcripts. rDiff.parametric requires as input the BAM files for both sample, as well as a GFF3 gene structure. In the following example we show how to apply rDiff.parmetric in order to test for differences between the two samples "1" and "2", which have replicates `bam1_r1.bam, bam1_r2.bam` resp. `bam2_r1.bam, bam2_r2.bam`. In our example we assume that the BAM files are located in the directory bamdir and that the reads are 75 bp long. Furthermore, we assume that our gene structure is saved in the file `genes.gff3` in the GFF3 format. The test can then be started by first changing into the directory `bin`:

```
cd bin
```

and then typing:

```
./rdiff -o outdir -d bamdir -a bam1_r1.bam,bam1_r2.bam -b bam2_r1.bam,bam2_r2.bam
-g genes.gff3 -m param -L 75 -m 30
```

Here, we required furthermore that a read has to be at least 30 bp long in order to be included in the analysis. A

The output files can be then found in `outdir` after rDiff is completed. The output files are

described in the following list:

P_values_rDiff_parametric.tab This file contains the p-values of rDiff.parametric. The file is tab-delimited and has three columns. The first column contains the gene names, the second the p-values and the third the test status.

Gene_expression.tab This file contains the gene expression estimations for all the replicates. The file is tab-delimited. The first column contains the gene names and the remaining columns the read counts for each gene for all replicates.

Alternative_region_counts.mat This file contains the counts for the alternative regions. The format is the binary mat format.

genes.mat This file contains the gene structure. The format is the binary mat format.

variance_function_1.mat This file contains the saved variance function for sample "1". It is a locfit-structure saved in the binary mat format.

variance_function_2.mat This file contains the saved variance function for sample "2". It is a locfit-structure saved in the binary mat format.

rDiff.nonparametric

When the gene structure is incomplete we recommend using rDiff.nonparametric. This test determines significant differences in read coverages between two samples. To run rDiff.nonparametric requires as input the BAM files for both samples as well as a GFF3 gene structure. rDiff.nonparametric tries to estimate the biological variance on the annotated gene structure. Therefore, it is of advantage to have an as complete gene structure as possible. For the testing rDiff.nonparametric uses only the gene starts and gene stops. In the following example we show how rDiff.nonparametric can be used to test for differences between the two samples "1" and "2", which have replicates `bam1_r1.bam, bam1_r2.bam` resp. `bam2_r1.bam, bam2_r2.bam`. In our example we assume that the BAM files are located in the directory bamdir and that the reads are 75 bp long. Furthermore, we assume that our gene structure is saved in the file `genes.gff3` in the GFF3 format. The test can then be started by first changing into the directory `bin`:

```
cd bin
```

and then typing:

```
./rdiff -o outdir -d bamdir -a bam1_r1.bam,bam1_r2.bam -b bam2_r1.bam,bam2_r2.bam
-g genes.gff3 -m nonparam -L 75 -m 30
```

Here, we required furthermore that a read has to be at least 30 bp long in order to be included in the analysis.

The output files can be then found in `outdir` after rDiff is completed. The output files are described in the following list:

P_values_rDiff_nonparametric.tab This file contains the p-values of rDiff.nonparametric. The file is tab-delimited and has three columns. The first column contains the gene names, the second the p-values and the third the test status.

Gene_expression.tab This file contains the gene expression estimations for all the replicates. The file is tab-delimited. The first column contains the gene names and the remaing columns contain the read counts for each gene in all replicates.

Nonparametric_region_counts.mat This file contains the counts for the alternative regions used to estimate the variance functions. The format is the binary mat format.

genes.mat This file contains the gene structure. The format is the binary mat format.

variance_function_1.mat This file contains the saved variance function for sample "1". It is a locfit-structure saved in the binary mat format.

variance_function_2.mat This file contains the saved variance function for sample "2". It is a locfit-structure saved in the binary mat format.

Working without replicates

When there is only one replicate available available in a sample we suggest merging all replicates for the estimation of the variance function. This can be achieved using the option -x when starting rDiff.

Bibliography

[1] HiSeq 1500/2500 Sequencing Systems. `http://res.illumina.com/documents/products/datasheets/datasheet_hiseq2500.pdf`.

[2] 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012.

[3] D. Aird, M. G. Ross, W.-S. Chen, M. Danielsson, T. Fennell, C. Russ, D. B. Jaffe, C. Nusbaum, and A. Gnirke. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology*, 12(2):R18, 2011.

[4] G. P. Alamancos, E. Agirre, and E. Eyras. Methods to study splicing from high-throughput RNA sequencing data. *ArXiv e-prints*, 2013.

[5] S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106, 2010.

[6] S. Anders, A. Reyes, and W. Huber. Detecting differential usage of exons from RNA-seq data. *Genome Research*, 2012.

[7] Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, 408(6814):796–815, 2000.

[8] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.

[9] P. L. Auer and R. W. Doerge. Statistical design and analysis of RNA sequencing data. *Genetics*, 185(2):405–416, 2010.

[10] Y. S. Aulchenko, S. Ripke, A. Isaacs, and C. M. Van Duijn. GenABEL: an R library for genome-wide association analysis. *Bioinformatics*, 23(10):1294–1296, 2007.

[11] O. T. Avery, C. M. MacLeod, and M. McCarty. Studies on the chemical nature of the substance inducing transformation of pneumococcal types induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. *The Journal of Experimental Medicine*, 79(2):137–158, 1944.

[12] T. L. Bailey. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics*, 27(12):1653–1659, 2011.

[13] A. Bashir, V. Bansal, and V. Bafna. Designing deep sequencing experiments: detecting structural variation and estimating transcript abundance. *BMC Genomics*, 11(1):385, 2010.

[14] J. Behr, A. Kahles, Y. Zhong, V. T. Sreedharan, P. Drewe, and G. Rätsch. MITIE: Simultaneous RNA-Seq-based transcript identification and quantification in multiple samples. *Bioinformatics*, 29(20):2529–2538, 2013.

[15] D. R. Bentley, S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, K. P. Hall, D. J. Evers, C. L. Barnes, H. R. Bignell, J. M. Boutell, J. Bryant, R. J. Carter, R. Keira Cheetham, A. J. Cox, D. J. Ellis, M. R. Flatbush, N. A. Gormley, S. J. Humphray, L. J. Irving, M. S. Karbelashvili, S. M. Kirk, H. Li, X. Liu, K. S. Maisinger, L. J. Murray, B. Obradovic, T. Ost, M. L. Parkinson, M. R. Pratt, I. M. Rasolonjatovo, M. T. Reed, R. Rigatti, C. Rodighiero, M. T. Ross, A. Sabot, S. V. Sankar, A. Scally, G. P. Schroth, M. E. Smith, V. P. Smith, A. Spiridou, P. E. Torrance, S. S. Tzonev, E. H. Vermaas, K. Walter, X. Wu, L. Zhang, M. D. Alam, C. Anastasi, I. C. Aniebo, D. M. Bailey, I. R. Bancarz, S. Banerjee, S. G. Barbour, P. A. Baybayan, V. A. Benoit, K. F. Benson, C. Bevis, P. J. Black, A. Boodhun, J. S. Brennan, J. A. Bridgham, R. C. Brown, A. A. Brown, D. H. Buermann, A. A. Bundu,

J. C. Burrows, N. P. Carter, N. Castillo, M. Chiara E Catenazzi, S. Chang, R. Neil Cooley, N. R. Crake, O. O. Dada, K. D. Diakoumakos, B. Dominguez-Fernandez, D. J. Earnshaw, U. C. Egbujor, D. W. Elmore, S. S. Etchin, M. R. Ewan, M. Fedurco, L. J. Fraser, K. V. Fuentes Fajardo, W. Scott Furey, D. George, K. J. Gietzen, C. P. Goddard, G. S. Golda, P. A. Granieri, D. E. Green, D. L. Gustafson, N. F. Hansen, K. Harnish, C. D. Haudenschild, N. I. Heyer, M. M. Hims, J. T. Ho, A. M. Horgan, K. Hoschler, S. Hurwitz, D. V. Ivanov, M. Q. Johnson, T. James, T. A. Huw Jones, G.-D. D. Kang, T. H. Kerelska, A. D. Kersey, I. Khrebtukova, A. P. Kindwall, Z. Kingsbury, P. I. Kokko-Gonzales, A. Kumar, M. A. Laurent, C. T. Lawley, S. E. Lee, X. Lee, A. K. Liao, J. A. Loch, M. Lok, S. Luo, R. M. Mammen, J. W. Martin, P. G. McCauley, P. McNitt, P. Mehta, K. W. Moon, J. W. Mullens, T. Newington, Z. Ning, B. Ling Ng, S. M. Novo, M. J. O'Neill, M. A. Osborne, A. Osnowski, O. Ostadan, L. L. Paraschos, L. Pickering, A. C. Pike, A. C. Pike, D. Chris Pinkard, D. P. Pliskin, J. Podhasky, V. J. Quijano, C. Raczy, V. H. Rae, S. R. Rawlings, A. Chiva Rodriguez, P. M. Roe, J. Rogers, M. C. Rogert Bacigalupo, N. Romanov, A. Romieu, R. K. Roth, N. J. Rourke, S. T. Ruediger, E. Rusman, R. M. Sanches-Kuiper, M. R. Schenker, J. M. Seoane, R. J. Shaw, M. K. Shiver, S. W. Short, N. L. Sizto, J. P. Sluis, M. A. Smith, J. Ernest Sohna Sohna, E. J. Spence, K. Stevens, N. Sutton, L. Szajkowski, C. L. Tregidgo, G. Turcatti, S. Vandevondele, Y. Verhovsky, S. M. Virk, S. Wakelin, G. C. Walcott, J. Wang, G. J. Worsley, J. Yan, L. Yau, M. Zuerlein, J. Rogers, J. C. Mullikin, M. E. Hurles, N. J. McCooke, J. S. West, F. L. Oaks, P. L. Lundberg, D. Klenerman, R. Durbin, and A. J. Smith. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59, 2008.

[16] M. F. Berger and M. L. Bulyk. Protein binding microarrays (pbms) for rapid, high-throughput characterization of the sequence specificities of DNA binding proteins. In *Gene Mapping, Discovery, and Expression*, pages 245–260. Springer, 2006.

[17] S. M. Berget, C. Moore, and P. A. Sharp. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proceedings of the National Academy of Sciences of the United States of America*, 74(8):3171–3175, 1977.

[18] F. Bertucci, S. Salas, S. Eysteries, V. Nasser, P. Finetti, C. Ginestier, E. Charafe-Jauffret, B. Loriod, L. Bachelart, J. Montfort, G. Victorero, F. Viret, V. Ollendorff, V. Fert, M. Giovaninni, J.-R. Delpero, C. Nguyen, P. Viens, G. Monges, D. Birnbaum, and R. Houlgatte. Gene expression profiling of colon cancer by dna microarrays and correlation with histoclinical parameters. *Oncogene*, 23(7):1377–1391, 2004.

[19] F. Besse and A. Ephrussi. Translational control of localized mRNAs: restricting protein synthesis in space and time. *Nature Reviews Molecular Cell Biology*, 9(12):971–80, 2008.

[20] C. M. Bishop. Pattern recognition and machine learning. 2007.

[21] D. L. Black. Mechanisms of alternative pre-messenger RNA splicing. *Annual Review of Biochemistry*, 72:291–336, 2003.

[22] D. Blankenberg, G. Von Kuster, N. Coraor, G. Ananda, R. Lazarus, M. Mangan, A. Nekrutenko, and J. Taylor. Galaxy: a web-based genome analysis tool for experimentalists. *Curr Protoc Mol Biol*, Chapter 19:Unit 19.10.1–Unit 19.1021, 2010.

[23] C. I. Bliss and R. A. Fisher. Fitting the negative binomial distribution to biological data. *Biometrics*, 9(2):176–200, 1953.

[24] R. Bohnert, J. Behr, and G. Rätsch. Transcript quantification with RNA-Seq data. *BMC Bioinformatics*, 10(Suppl 13):P5, 2009.

[25] C. Bonferroni. Teoria statistica delle classi e calcolo delle probabilit. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8:3–62, 1936.

[26] K. Borgwardt, A. Gretton, M. Rasch, H. Kriegel, B. Schölkopf, and A. Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–57, 2006.

[27] O. Bousquet, S. Boucheron, and G. Lugosi. Introduction to statistical learning theory. In *Advanced Lectures on Machine Learning*, pages 169–207. Springer, 2004.

[28] A. N. Brooks, L. Yang, M. O. Duff, K. D. Hansen, J. W. Park, S. Dudoit, S. E. Brenner, and B. R. Graveley. Conservation of an RNA regulatory map between drosophila and mammals. *Genome Research*, 21(2):193–202, 2011.

[29] Cancer Genome Atlas Research Network. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature*, 499(7456):43–49, 2013.

[30] M. Chen and J. L. Manley. Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nature Reviews Molecular Cell Biology*, 10(11):741–54, 2009.

[31] L. Chow. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell*, 12(1):1–8, 1977.

[32] D. Clark and N. Pazdernik. *Molecular Biology*. Academic Press, 2013.

[33] R. M. Clark, G. Schweikert, C. Toomajian, S. Ossowski, G. Zeller, P. Shinn, N. Warthmann, T. T. Hu, G. Fu, D. A. Hinds, H. Chen, K. A. Frazer, D. H. Huson, B. Schölkopf, M. Nordborg, G. Rätsch, J. R. Ecker, and D. Weigel. Common sequence polymorphisms shaping genetic diversity in Arabidopsis thaliana. *Science*, 317(5836):338–342, 2007.

[34] A. Cornish-Bowden. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acids Research*, 13(9):3021, 1985.

[35] F. Crick. On protein synthesis. *The Symposia of the Society for Experimental Biology*, 12:138–163, 1958.

[36] E. H. Davidson and D. H. Erwin. Gene regulatory networks and the evolution of animal body plans. *Science*, 311(5762):796–800, 2006.

[37] A. De Moivre. *The doctrine of chances*. 1738.

[38] S. Djebali, C. A. Davis, A. Merkel, A. Dobin, T. Lassmann, A. Mortazavi, A. Tanzer, J. Lagarde, W. Lin, F. Schlesinger, C. Xue, G. K. Marinov, J. Khatun, B. A. Williams, C. Zaleski, J. Rozowsky, M. Roder, F. Kokocinski, R. F. Abdelhamid, T. Alioto, I. Antoshechkin, M. T. Baer, N. S. Bar, P. Batut, K. Bell, I. Bell, S. Chakrabortty, X. Chen, J. Chrast, J. Curado, T. Derrien, J. Drenkow, E. Dumais, J. Dumais, R. Duttagupta, E. Falconnet, M. Fastuca, K. Fejes-Toth, P. Ferreira, S. Foissac, M. J. Fullwood, H. Gao, D. Gonzalez, A. Gordon, H. Gunawardena, C. Howald, S. Jha, R. Johnson, P. Kapranov, B. King, C. Kingswood, O. J. Luo, E. Park, K. Persaud, J. B. Preall, P. Ribeca, B. Risk, D. Robyr, M. Sammeth, L. Schaffer, L.-H. See, A. Shahab, J. Skancke, A. M. Suzuki, H. Takahashi, H. Tilgner, D. Trout, N. Walters, H. Wang, J. Wrobel, Y. Yu, X. Ruan, Y. Hayashizaki, J. Harrow, M. Gerstein, T. Hubbard, A. Reymond, S. E. Antonarakis, G. Hannon, M. C. Giddings, Y. Ruan, B. Wold, P. Carninci, R. Guigo, and T. R. Gingeras. Landscape of transcription in human cells. *Nature*, 489(7414):101–108, 2012.

[39] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

[40] J. Doebley and L. Lukens. Transcriptional regulators and the evolution of plant form. *The Plant Cell*, 10(7):1075–1082, 1998.

[41] G. Drechsel, A. Kahles, A. K. Kesarwani, E. Stauffer, J. Behr, P. Drewe, G. Rätsch, and A. Wachter. Nonsense-mediated decay of alternative precursor mRNA splicing variants is a major determinant of the Arabidopsis steady state transcriptome. *Plant Cell*, 25(10):3726–3742, 2013.

[42] P. Drewe, O. Stegle, L. Hartmann, A. Kahles, R. Bohnert, A. Wachter, K. Borgwardt, and G. Rätsch. Accurate detection of differential RNA processing. *Nucleic Acids Research*, 41(10):5189–5198, 2013.

[43] R. M. Dudley. *Real analysis and probability*, volume 74. Cambridge University Press, 2002.

[44] B. Efron. Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, pages 1–26, 1979.

[45] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004.

[46] P. Engström, T. Steijger, B. Sipos, G. R. Grant, A. Kahles, The RGASP Consortium, G. Rätsch, N. Goldman, T. J. Hubbard, J. Harrow, R. Guigo, and P. Bertone. Systematic evaluation of spliced alignment programs for RNA-seq data. *Nature Methods*, 2013.

[47] N. A. Faustino and T. A. Cooper. Pre-mRNA splicing and human disease. *Genes & Development*, 17(4):419–437, 2003.

[48] R. A. Fisher. The correlation between relatives on the supposition of mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, 52(02):399–433, 1919.

[49] P. Francis, H. M. Namløs, C. Müller, P. Edén, J. Fernebro, J.-M. Berner, B. Bjerkehagen, M. Åkerman, P.-O. Bendahl, A. Isinger, A. Rydholm, O. Myklebost, and M. Nilbert. Diagnostic and prognostic gene expression signatures in 177 soft tissue sarcomas: hypoxia-induced transcription profile signifies metastatic potential. *BMC Genomics*, 8(1):73, 2007.

[50] N. Friedman, L. Cai, and X. S. Xie. Linking stochastic dynamics to population distribution: an analytical framework of gene expression. *Physical Review Letters*, 97(16):168302, 2006.

[51] M. C. Frith, U. Hansen, J. L. Spouge, and Z. Weng. Finding functional sequence elements by multiple local alignment. *Nucleic Acids Research*, 32(1):189–200, 2004.

[52] X. Gan, O. Stegle, J. Behr, J. Steffen, P. Drewe, K. Hildebrand, R. Lyngsoe, S. Schultheiss, E. Osborne, V. Sreedharan, A. Kahles, R. Bohnert, G. Jean, P. Derwent, P. Kersey, E. Belfield, N. Harberd, E. Kemen, C. Toomajian, P. Kover, R. Clark, G. Rätsch, and R. Mott. Multiple reference genomes and transcriptomes for Arabidopsis thaliana. *Nature*, 477(7365):419–423, 2011.

[53] H.-O. Georgii. *Stochastik: Einführung in die Wahrscheinlichkeitstheorie und Statistik*. Walter de Gruyter, 2009.

[54] B. Giardine, C. Riemer, R. C. Hardison, R. Burhans, L. Elnitski, P. Shah, Y. Zhang, D. Blankenberg, I. Albert, J. Taylor, W. Miller, W. J. Kent, and A. Nekrutenko. Galaxy: a platform for interactive large-scale genome analysis. *Genome Res*, 15(10):1451–1455, 2005.

[55] P. Glaus, A. Honkela, and M. Rattray. Identifying differentially expressed transcripts from RNA-seq data with biological variation. *Bioinformatics*, 28(13):1721–1728, 2012.

[56] J. Goecks, A. Nekrutenko, J. Taylor, and G. T. . Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, 11(8):R86, 2010.

[57] A. Goffeau, B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon, H. Feldmann, F. Galibert, J. D. Hoheisel, C. Jacq, M. Johnston, E. J. Louis, H. W. Mewes, Y. Murakami, P. Philippsen, H. Tettelin, and S. G. Oliver. Life with 6000 genes. *Science*, 274(5287):546–567, 1996.

[58] J. R. González, L. Armengol, X. Solé, E. Guinó, J. M. Mercader, X. Estivill, and V. Moreno. SNPassoc: an R package to perform whole genome association studies. *Bioinformatics*, 23(5):654–655, 2007.

[59] M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, and A. Regev. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, 29(7):644–652, 2011.

[60] C. E. Grant, T. L. Bailey, and W. S. Noble. FIMO: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018, 2011.

[61] E. D. Green. Strategies for the systematic sequencing of complex genomes. *Nature Reviews Genetics*, 2(8):573–583, 2001.

[62] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13:723–773, 2012.

[63] T. Griebel, B. Zacher, P. Ribeca, E. Raineri, V. Lacroix, R. GuigŮ, and M. Sammeth. Modelling and simulating generic RNA-Seq experiments with the flux simulator. *Nucleic Acids Research*, 2012.

[64] M. Hafner, M. Landthaler, L. Burger, M. Khorshid, J. Hausser, P. Berninger, A. Rothballer, M. Manuel Ascano, A.-C. Jungkamp, M. Munschauer, A. Ulrich, G. S. Wardle, S. Dewell, M. Zavolan, and T. Tuschl. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, 141(1):129–141, 2010.

[65] K. D. Hansen, S. E. Brenner, and S. Dudoit. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Research*, 38(12):e131, 2010.

[66] T. Hardcastle and K. Kelly. baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, 11(1):422, 2010.

[67] T. Hastie, R. Tibshirani, and J. J. H. Friedman. *The elements of statistical learning*, volume 1. Springer New York, 2001.

[68] S. Heber, M. Alekseyev, S.-H. Sze, H. Tang, and P. A. Pevzner. Splicing graphs and EST assembly problem. *Bioinformatics*, 18 Suppl 1:S181–S188, 2002.

[69] R. Higuchi, G. Dollinger, P. S. Walsh, and R. Griffith. Simultaneous amplification and detection of specific DNA sequences. *Biotechnology*, 10(4):413–417, 1992.

[70] D. Hiller, H. Jiang, W. Xu, and W. Wong. Identifiability of isoform deconvolution from junction arrays and RNA-Seq. *Bioinformatics*, 25(23):3056–9, 2009.

[71] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, pages 65–70, 1979.

[72] F. C. Holstege, E. G. Jennings, J. J. Wyrick, T. I. Lee, C. J. Hengartner, M. R. Green, T. R. Golub, E. S. Lander, and R. A. Young. Dissecting the regulatory circuitry of a eukaryotic genome. *Cell*, 95(5):717–728, 1998.

[73] N. T. Ingolia, S. Ghaemmaghami, J. R. Newman, and J. S. Weissman. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, 324(5924):218–223, 2009.

[74] R. J. Jackson, C. U. Hellen, and T. V. Pestova. The mechanism of eukaryotic translation initiation and principles of its regulation. *Nature Reviews Molecular Cell Biology*, 11(2):113–127, 2010.

[75] G. Jean, A. Kahles, V. Sreedharan, F. De Bona, and G. Rätsch. RNA-Seq read alignments with PALMapper. *Current Protocols in Bioinformatics*, Chapter 11(December):Unit 11.6, 2010.

[76] H. Jiang and W. Wong. Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics*, 25(8):1026–32, 2009.

[77] J. Jiang. *Linear and generalized linear mixed models and their applications*. Springer, 2007.

[78] L. Jiang, F. Schlesinger, C. A. Davis, Y. Zhang, R. Li, M. Salit, T. R. Gingeras, and B. Oliver. Synthetic spike-in standards for RNA-seq experiments. *Genome Research*, 21(9):1543–1551, 2011.

[79] W. Johannsen. *Elemente der exakten Erblichkeitslehre*. Jena,G. Fischer, 1909.

[80] D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold. Genome-wide mapping of in vivo protein-DNA interactions. *Science*, 316(5830):1497–1502, 2007.

[81] N. Johnson, A. Kemp, and S. Kotz. *Univariate Discrete Distributions*. Wiley Series in Probability and Statistics. Wiley, 2005.

[82] E. Jones, T. Oliphant, P. Peterson, et al. SciPy: Open source scientific tools for Python, 2001–.

[83] Y. Katz, E. Wang, E. Airoldi, and C. Burge. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods*, 7(12):1009–1015, 2010.

[84] M. Kertesz, Y. Wan, E. Mazor, J. L. Rinn, R. C. Nutter, H. Y. Chang, and E. Segal. Genome-wide measurement of RNA secondary structure in yeast. *Nature*, 467(7311):103–107, 2010.

[85] M. Kinsey, R. Smith, and S. L. Lessnick. NR0B1 is required for the oncogenic phenotype mediated by EWS/FLI in Ewing's sarcoma. *Molecular Cancer Research*, 4(11):851–859, 2006.

[86] M. Kircher and J. Kelso. High-throughput DNA sequencing–concepts and limitations. *Bioessays*, 32(6):524–536, 2010.

[87] A. N. Kolmogorov. Sulla determinazione empirica di una legge di distribuzione. *Giornale dellŠ Instituto Italiano degli Attuari*, 4(1):83–91, 1933.

[88] J. Kong and P. Lasko. Translational control in cellular and developmental processes. *Nature Reviews Genetics*, 13(6):383–94, 2012.

[89] P. Korkuć, J. H. Schippers, and D. Walther. Characterization and identification of cis-regulatory elements in Arabidopsis based on single-nucleotide polymorphism information. *Plant Physiology*, 164(1):181–200, 2014.

[90] A. R. Kornblihtt, I. E. Schor, M. Alló, G. Dujardin, E. Petrillo, and M. J. Muñoz. Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nature Reviews Molecular Cell Biology*, 2013.

[91] P. X. Kover and R. Mott. Mapping the genetic basis of ecologically and evolutionarily relevant traits in Arabidopsis thaliana. *Current Opinion in Plant Biology*, 15(2):212–217, 2012.

[92] P. X. Kover, W. Valdar, J. Trakalo, N. Scarcelli, I. M. Ehrenreich, M. D. Purugganan, C. Durrant, and R. Mott. A multiparent advanced generation inter-cross to fine-map quantitative traits in Arabidopsis thaliana. *PLoS Genetics*, 5(7):e1000551, 2009.

[93] E. Kristiansson, M. Thorsen, M. J. Tamás, and O. Nerman. Evolutionary forces act on promoter length: identification of enriched cis-regulatory elements. *Molecular Biology and Evolution*, 26(6):1299–1307, 2009.

[94] V. Lacroix, M. Sammeth, R. Guigo, and A. Bergeron. Exact transcriptome reconstruction from short sequence reads. In *Workshop on Algorithms in Bioinformatics*, pages 50–63, 2008.

[95] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blöcker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowski, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y. J. Chen, and J. Szustakowki. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[96] P. Landgraf, M. Rusu, R. Sheridan, A. Sewer, N. Iovino, A. Aravin, S. Pfeffer, A. Rice, A. O. Kamphorst, M. Landthaler, C. Lin, N. D. Socci, L. Hermida, V. Fulci, S. Chiaretti, R. Foà, J. Schliwka, U. Fuchs, A. Novosel, R.-U. Müller, B. Schermer, U. Bissels, J. Inman, Q. Phan, M. Chien, D. B. Weir, R. Choksi, G. De Vita, D. Frezzetti, H.-I. Trompeter, V. Hornung, G. Teng, G. Hartmann, M. Palkovits, R. Di Lauro, P. Wernet, G. Macino, C. E. Rogler, J. W. Nagle, J. Ju, F. N. Papavasiliou, T. Benzing,

P. Lichter, W. Tam, M. J. Brownstein, A. Bosio, A. Borkhardt, J. J. Russo, C. Sander, M. Zavolan, and T. Tuschl. A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell*, 129(7):1401–1414, 2007.

[97] B. Lawrence, A. Perez-Atayde, M. K. Hibbard, B. P. Rubin, P. D. Cin, J. L. Pinkus, G. S. Pinkus, S. Xiao, E. S. Yi, C. D. Fletcher, and J. A. Fletcher. TPM3-ALK and TPM4-ALK Oncogenes in Inflammatory Myofibroblastic Tumors . *The American Journal of Pathology*, 157(2):377 – 384, 2000.

[98] E. L. Lehmann and J. Romano. *Testing statistical hypotheses*. Springer Texts in Statistics. Springer, New York, third edition, 2005.

[99] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[100] J. Li, H. Jiang, and W. Wong. Method modeling non-uniformity in short-read rates in RNA-Seq data. *Genome Biology*, 11(5):R25, 2010.

[101] R. Li, H. Zhu, J. Ruan, W. Qian, X. Fang, Z. Shi, Y. Li, S. Li, G. Shan, K. Kristiansen, S. Li, H. Yang, J. Wang, , and J. Wang. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Research*, 20(2):265–272, 2010.

[102] M. J. Lindstrom and D. M. Bates. Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, 83(404):1014–1022, 1988.

[103] C. Lippert, J. Listgarten, Y. Liu, C. M. Kadie, R. I. Davidson, and D. Heckerman. FaST linear mixed models for genome-wide association studies. *Nature Methods*, 8(10):833–835, 2011.

[104] C. Loader. locfit: Local regression, likelihood and density estimation, 2007. R package.

[105] H. Lodish, A. Berk, C. A. Kaiser, M. Krieger, M. P. Scott, A. Bretscher, H. Ploegh, and P. Matsudaira. *Molecular Cell Biology*. W. H. Freeman, 7th edition, 2013.

[106] R. Lorenz, S. H. Bernhart, C. H. Zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker. ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, 6(1):26, 2011.

[107] J. B. Lucks, S. A. Mortimer, C. Trapnell, S. Luo, S. Aviran, G. P. Schroth, L. Pachter, J. A. Doudna, and A. P. Arkin. Multiplexed RNA structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Proceedings of the National Academy of Sciences of the United States of America*, 108(27):11063–11068, 2011.

[108] X. M. Ma and J. Blenis. Molecular mechanisms of mTOR-mediated translational control. *Nature Reviews Molecular Cell Biology*, 10(5):307–18, 2009.

[109] M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y.-J. J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. H. Ho, C. H. H. Ho, G. P. Irzyk, S. C. Jando, M. L. Alenquer, T. P. Jarvie, K. B. Jirage, J.-B. B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley, and J. M. Rothberg. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380, 2005.

[110] J. Marioni, C. Mason, S. Mane, M. Stephens, and Y. Gilad. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Resarch*, 18(9):1509–17, 2008.

[111] G. A. Maston, S. K. Evans, and M. R. Green. Transcriptional regulatory elements in the human genome. *Annual Review of Genomics and Human Genetics*, 7(1):29–59, 2006.

[112] MATLAB. version 7.7 (r2008a), 2008.

[113] V. Matys, O. V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer, N. Voss, P. Stegmaier, B. Lewicki-Potapov, H. Saxel, A. E. Kel, and E. Wingender. TRANSFAC® and its module TRANSCompel®: transcriptional gene regulation in eukaryotes. *Nucleic Acids Research*, 34(suppl 1):D108–D110, 2006.

[114] A. M. Maxam and W. Gilbert. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 74(2):560–4, 1977.

[115] M. L. Metzker. Sequencing technologies - the next generation. *Nature Reviews Gentics*, 11(1):31–46, 2010.

[116] L. Michaelis and M. L. Menten. Die Kinetik der Invertinwirkung. *Biochemie*, 49(333-369):352, 1913.

[117] E. Michishita, G. Garcés, J. C. Barrett, and I. Horikawa. Upregulation of the kiaa1199 gene is associated with cellular mortality. *Cancer letters*, 239(1):71–77, 2006.

[118] F. Miescher. Hoppe-seyler's medizinisch-chemische untersuchungen. *Über die chemische Zusammensetzung der Eiterzellen*, 4:441–460, 1871.

[119] T. Morgan. *The Mechanism of Mendelian Heredity*. H. Holt, 1915.

[120] T. H. Morgan. Random segregation versus coupling in Mendelian inheritance. *Science*, 34(873):384–384, 1911.

[121] A. Mortazavi, B. Williams, K. McCue, L. Schaeffer, and B. Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7):621–8, 2008.

[122] S. Motameny, S. Wolters, P. Nürnberg, and B. Schumacher. Next generation sequencing of miRNAs–strategies, resources and methods. *Genes*, 1(1):70–84, 2010.

[123] K. Murphy. *Machine Learning: A Probabilistic Perspective*. Adaptive computation and machine learning series. MIT Press, 2012.

[124] K. P. Murphy. *Machine learning: a probabilistic perspective*. The MIT Press, 2012.

[125] E. W. Myers, G. G. Sutton, A. L. Delcher, I. M. Dew, D. P. Fasulo, M. J. Flanigan, S. A. Kravitz, C. M. Mobarry, K. H. J. Reinert, K. A. Remington, E. L. Anson, R. A. Bolanos, H.-H. Chou, C. M. Jordan, A. L. Halpern, S. Lonardi, E. M. Beasley, R. C. Brandon, L. Chen, P. J. Dunn, Z. Lai, Y. Liang, D. R. Nusskern, M. Zhan, Q. Zhang, X. Zheng, G. M. Rubin, M. D. Adams, and J. C. Venter. A whole-genome assembly of Drosophila. *Science*, 287(5461):2196–2204, 2000.

[126] U. Nagalakshmi, Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, and M. Snyder. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, 320(5881):1344–1349, 2008.

[127] J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231:289–337, 1933.

[128] Octave community. GNU Octave 3.4, 2014.

[129] S. Ossowski, K. Schneeberger, R. M. Clark, C. Lanz, N. Warthmann, and D. Weigel. Sequencing of natural strains of arabidopsis thaliana with short reads. *Genome research*, 18(12):2024–2033, 2008.

[130] F. Ozsolak and P. M. Milos. Rna sequencing: advances, challenges and opportunities. *Nature Reviews Genetics*, 12(2):87–98, 2011.

[131] H. Pearson. Genetics: What is a gene? *Nature*, 441(7092):398–401, 2006.

[132] S. Pepke, B. Wold, and A. Mortazavi. Computation for ChIP-seq and RNA-seq studies. *Nature Methods*, 6(11s):S22–S32, 2009.

[133] R. Perales and D. Bentley. "Cotranscriptionality": the transcription elongation complex as a nexus for nuclear transactions. *Molecular Cell*, 36(2):178–91, 2009.

[134] V. Plagnol, J. Curtis, M. Epstein, K. Y. Mok, E. Stebbings, S. Grigoriadou, N. W. Wood, S. Hambleton, S. O. Burns, A. J. Thrasher, D. Kumararatne, R. Doffinger, and S. Nejentsev. A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics*, 28(21):2747–2754, 2012.

[135] J. Przyborowski and H. Wilenski. Homogeneity of results in testing samples from poisson series: With an application to testing clover seed for dodder. *Biometrika*, 31(3/4):313–323, 1940.

[136] H. Richard, M. H. Schulz, M. Sultan, A. Nurnberger, S. Schrinner, D. Balzereit, E. Dagand, A. Rasche, H. Lehrach, M. Vingron, S. A. Haas, and M. Yaspo. Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments. *Nucleic Acids Research*, 38(10), 2010.

[137] A. Roberts, C. Trapnell, J. Donaghey, J. L. Rinn, and L. Pachter. Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biology*, 12(3):R22, 2011.

[138] G. Robertson, J. Schein, R. Chiu, R. Corbett, M. Field, S. D. Jackman, K. Mungall, S. Lee, H. M. Okada, J. Q. Qian, M. Griffith, A. Raymond, N. Thiessen, T. Cezard, Y. S. Butterfield, R. Newsome, S. K. Chan, R. She, R. Varhol, B. Kamoh, A.-L. Prabhu, A. Tam, Y. Zhao, R. A. Moore, M. Hirst, M. A. Marra, S. J. M. Jones, P. A. Hoodless, and I. Birol. De novo assembly and analysis of RNA-seq data. *Nature Methods*, 7(11):909–912, 2010.

[139] M. Robinson and A. Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3):R25, 2010.

[140] M. D. Robinson and G. K. Smyth. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23(21):2881–7, 2007.

[141] M. D. Robinson and G. K. Smyth. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics*, 9(2):321–32, 2008.

[142] J. Robles, S. Qureshi, S. Stephen, S. Wilson, C. Burden, and J. Taylor. Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing. *BMC Genomics*, 13(1):484, 2012.

[143] M. G. Ross, C. Russ, M. Costello, A. Hollinger, N. J. Lennon, R. Hegarty, C. Nusbaum, and D. B. Jaffe. Characterizing and measuring bias in sequence data. *Genome Biology*, 14(5):R51, 2013.

[144] J.-F. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. F. Berriz, F. D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, N. Klitgord, C. Simon, M. Boxem, S. Milstein, J. Rosenberg, D. S. Goldberg, L. V. Zhang, S. L. Wong, G. Franklin, S. Li, J. S. Albala, J. Lim, C. Fraughton, E. Llamosas, S. Cevik, C. Bex, P. Lamesch, R. S. Sikorski, J. Vandenhaute, H. Y. Zoghbi, A. Smolyar, S. Bosak, R. Sequerra, L. Doucette-Stamm, M. E. Cusick, D. E. Hill, F. P. Roth, and M. Vidal. Towards a proteome-scale map of the human protein–protein interaction network. *Nature*, 437(7062):1173–1178, 2005.

[145] W. Rudin. *Real and Complex Analysis*. Tata McGraw-Hill, 1987.

[146] A. Sandelin, W. Alkema, P. Engström, W. W. Wasserman, and B. Lenhard. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Research*, 32(suppl 1):D91–D94, 2004.

[147] A. Sandelin, P. Carninci, B. Lenhard, J. Ponjavic, Y. Hayashizaki, and D. a. Hume. Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nature Reviews Genetics*, 8(6):424–36, 2007.

[148] F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12):5463–5467, 1977.

[149] M. Schmid, T. S. Davison, S. R. Henz, U. J. Pape, M. Demar, M. Vingron, B. Schölkopf, D. Weigel, and J. U. Lohmann. A gene expression map of Arabidopsis thaliana development. *Nature Genetics*, 37(5):501–506, 2005.

[150] R. J. Schmitz, M. D. Schultz, M. A. Urich, J. R. Nery, M. Pelizzola, O. Libiger, A. Alix, R. B. McCosh, H. Chen, N. J. Schork, and J. R. Ecker. Patterns of population epigenomic diversity. *Nature*, 495(7440):193–198, 2013.

[151] D. Schmucker, J. C. Clemens, H. Shu, C. A. Worby, J. Xiao, M. Muda, J. E. Dixon, and S. L. Zipursky. Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell*, 101(6):671–684, 2000.

[152] T. D. Schneider and R. M. Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research*, 18(20):6097–6100, 1990.

[153] G. Schweikert, B. Cseke, T. Clouaire, A. Bird, and G. Sanguinetti. MMDiff: quantitative testing for shape changes in ChIP-Seq data sets. *BMC Genomics*, 14(1):826, 2013.

[154] G. Schweikert, A. Zien, G. Zeller, J. Behr, C. Dieterich, C. S. Ong, P. Philips, F. De Bona, L. Hartmann, A. Bohlen, N. Krüger, S. Sonnenburg, and G. Rätsch. mGene: accurate SVM-based gene finding with an application to nematode genomes. *Genome Research*, 19(11):2133–2143, 2009.

[155] A. M. Sherwood, C. Desmarais, R. J. Livingston, J. Andriesen, M. Haussler, C. S. Carlson, and H. Robins. Deep sequencing of the human TCRγ and TCRβ repertoires suggests that TCRβ rearranges after $\alpha\beta$ and $\gamma\delta$ T cell commitment. *Science Translational Medicine*, 3(90):90ra61, 2011.

[156] D. Silvera, S. C. Formenti, and R. J. Schneider. Translational control in cancer. *Nature Reviews Cancer*, 10(4):254–66, 2010.

[157] D. Singh, C. F. Orellana, Y. Hu, C. D. Jones, Y. Liu, D. Y. Chiang, J. Liu, and J. F. Prins. FDM: a graph-based statistical method to detect differential transcription using RNA-seq data. *Bioinformatics*, 27(19):2633–2640, 2011.

[158] L. M. Smith, L. Hartmann, P. Drewe, R. Bohnert, A. Kahles, C. Lanz, and G. Rätsch. Multiple insert size paired-end sequencing for deconvolution of complex transcriptomes. *RNA Biology*, 9(5):596–609, 2012.

[159] A. J. Smola and B. Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222, 2004.

[160] F. Spitz and E. E. Furlong. Transcription factors: from enhancer binding to developmental control. *Nature Reviews Genetics*, 2012.

[161] V. T. Sreedharan, S. J. Schultheiss, G. Jean, A. Kahles, R. Bohnert, P. Drewe, P. Mudrakarta, N. Görnitz, G. Zeller, and G. Rätsch. Oqtans: The RNA-seq Workbench in the Cloud for Complete and Reproducible Quantitative Transcriptome Analysis. *Bioinformatics*, 2014.

[162] R. Staden. The staden sequence analysis package. *Molecular Biotechnology*, 5(3):233–241, 1996.

[163] M. Stanke, R. Steinkamp, S. Waack, and B. Morgenstern. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Research*, 32(suppl 2):W309–W312, 2004.

[164] I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *The Journal of Machine Learning Research*, 2:67–93, 2002.

[165] J. D. Storey and R. Tibshirani. Statistical significance for genome-wide experiments. *Proceedings of the National Academy of Sciences of the United States of America*, 100:9440–9445, 2003.

[166] G. D. Stormo and D. S. Fields. Specificity, free energy and information content in protein–DNA interactions. *Trends in Biochemical Sciences*, 23(3):109–113, 1998.

[167] The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. 2012.

[168] W. Thompson, E. C. Rouchka, and C. E. Lawrence. Gibbs Recursive Sampler: finding transcription factor binding sites. *Nucleic Acids Research*, 31(13):3580–3585, 2003.

[169] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[170] A. Tikhnov and V. Y. Arsenin. Solution of ill-posed problems. *Winston, Wasington. DC*, 1977.

[171] C. Trapnell, D. G. Hendrickson, M. Sauvageau, L. Goff, J. L. Rinn, and L. Pachter. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature Biotechnology*, 31(1):46–53, 2012.

[172] C. Trapnell, L. Pachter, and S. L. Salzberg. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105–1111, 2009.

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

[174] J. W. Tukey. Bias and confidence in not quite large samples. *Annals of Mathematical Statistics*, 29(2):614, 1958.

[175] J. G. Underwood, A. V. Uzilov, S. Katzman, C. S. Onodera, J. E. Mainzer, D. H. Mathews, T. M. Lowe, S. R. Salama, and D. Haussler. FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. *Nature Methods*, 7(12):995–1001, 2010.

[176] A. Valouev, J. Ichikawa, T. Tonthat, J. Stuart, S. Ranade, H. Peckham, K. Zeng, J. A. Malek, G. Costa, K. McKernan, A. Sidow, A. Fire, and S. M. Johnson. A high-resolution, nucleosome position map of c. elegans reveals a lack of universal sequence-dictated positioning. *Genome Research*, 18(7):1051–1063, 2008.

[177] V. Vapnik. Estimation of dependencies based on empirical data. *New York*, 1982.

[178] V. N. Vapnik and A. J. Chervonenkis. Theory of pattern recognition. 1974.

[179] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, and R. A. Holt. The sequence of the human genome. *Science*, 291(5507):1304–1351, 2001.

[180] Y. Wan, K. Qu, Q. C. Zhang, R. A. Flynn, O. Manor, Z. Ouyang, J. Zhang, R. C. Spitale, M. P. Snyder, E. Segal, and H. Y. Chang. Landscape and variation of RNA secondary structure across the human transcriptome. *Nature*, 505(7485):706–709, 2014.

[181] E. Wang, R. Sandberg, S. Luo, and I. Khrebtukova. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221), 2008.

[182] E. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. Kingsmore, G. Schroth, and C. Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476, 2008.

[183] Z. Wang, M. Gerstein, and M. Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009.

[184] J. Watson and F. Crick. A structure for deoxyribose nucleic acid. *Nature*, 171, 1953.

[185] M. T. Weirauch, A. Yang, M. Albu, A. Cote, A. Montenegro-Montero, H. S. Drewe, P.and Najafabadi, S. A. Lambert, I. Mann, K. Cook, H. Zheng, A. Goity, H. van Bakel, J. Lozano, M. Galli, M. Lewsey, E. Huang, T. Mukherjee, X. Chen, J. S. Reece-Hoyes, S. Govindarajan, G. Shaulsky, A. J. M. Walhout, F. Bouget, G. Rätsch, L. F. Larrondo, J. R. Ecker, and T. R. Hughes. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, 2014. (in press).

[186] B. T. Wilhelm and J.-R. Landry. RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing. *Methods*, 48(3):249–257, 2009.

[187] A. L. Wolfe, K. Singh, Y. Zhong, P. Drewe, V. K. Rajasekhar, K. J. Mavrakis, J. Van der Meulen, J. H. Schatz, C. M. Rodrigo, M. Jiang, P. Rondou, E. de Stanchina, J. Teruya-Feldstein, F. Speleman, J. A. Porco Jr., J. Pelletier, G. Rätsch, and H.-G. Wendel. The 5' untranslated region (UTR) of many oncogenes and transcription factors encodes a targetable dependence on the eIF4A RNA helicase. (in press), 2014.

[188] Q. Xu, B. Modrek, and C. Lee. Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. *Nucleic Acids Research*, 30(17):3754–3766, 2002.

[189] M. Yamaguchi, M. Ohtani, N. Mitsuda, M. Kubo, M. Ohme-Takagi, H. Fukuda, and T. Demura. VND-INTERACTING2, a NAC domain transcription factor, negatively regulates xylem vessel formation in Arabidopsis. *The Plant Cell*, 22(4):1249–1263, 2010.

[190] J. Yang, N. A. Zaitlen, M. E. Goddard, P. M. Visscher, and A. L. Price. Advantages and pitfalls in the application of mixed-model association methods. *Nature Genetics*, 46(2):100–106, 2014.

[191] Y. Zhao and G. D. Stormo. Quantitative analysis demonstrates most transcription factors require only simple models of specificity. *Nature Biotechnology*, 29(6):480–483, 2011.

[192] Q. Zhou, T. Li, and D. H. Price. RNA polymerase II elongation control. *Annual Review of Biochemistry*, 81:119–143, 2012.