# Computational studies on RNA processing in higher eukaryotes

Dissertation
zur Erlangung des Doktorgrades
der Naturwissenschaften

vorgelegt beim
Fachbereich Biowissenschaften (FB15) der
Goethe-Universität Frankfurt am Main

von
**Mirko Brüggemann**
geboren in Hanau (Hessen)

Frankfurt am Main 2021
(D30)

Vom Fachbereich Biowissenschaften (FB15) der Goethe-Universität als Dissertation angenommen.

Dekan: Prof. Dr. Sven Klimpel

Gutachter: Dr. Kathi Zarnack
Gutachter: Prof. Dr. Ingo Ebersberger

Datum der Disputation: ______________________________

# Contents

# List of Figures

# List of Tables

# List of Abbreviations

**Table 1: List of abbreviations.**

| | |
|---|---|
| % | Percentage |
| 3'SS | 3' splice site |
| 3'UTR | 3' Untranslated region |
| 4SU | 4-thiouridine |
| 5'SS | 5' splice site |
| 5'UTR | 5' Untranslated region |
| A3SS | Alternative 3' splice site |
| A5SS | Alternative 5' splice site |
| AS | Alternative splicing |
| BAM | Binary Alignment Map |
| BMLS | Buchmann Institute for Molecular Life Sciences |
| bp | Base pairs |
| BP | Branch point |
| BPS | Branch point sequence |
| C | Cytosine |
| cDNA | complementary DNA |
| CDS | Coding sequence |
| CE | Cassette exon |
| CLIP | Cross linking and immunoprecipitation |
| CLIP-Seq | CLIP sequencing |
| ChIP | Chromatin immunoprecipitation |

*Continued on next page*

Table 1 – *Continued from previous page*

| | |
|---|---|
| ChIP-Seq | ChIP sequencing |
| $cm^2$ | square centimetre |
| DEG | Differentially expressed gene |
| DGE | Differential gene expression |
| DNA | Desoxyribonucleic acid |
| EBI | European Bioinformatics Institute |
| eCLIP | enhanced CLIP |
| ESE | Exonic splicing enhancer |
| ESS | Exonic splicing silencer |
| FC | Fold change |
| FDR | False discovery rate |
| G | Guanine |
| GEO | Gene Expression Omnibus |
| GLIS3 | Kruppel-like zinc finger protein Gli-similar 3 |
| GLM | Generalized linear model |
| GO | Gene Ontology |
| GTF | General transfer format |
| GWAS | Genome wide association study |
| hg38 | Human reference genome 38 |
| HITS-CLIP | High-throughput sequencing of RNA isolated by crosslinking immunoprecipitation |
| HMM | Hidden Markov Model |
| hnRNP | Heterogeneous nuclear ribonucleoproteins |
| iCLIP | Individual nucleotide resolution cross linking and immunoprecipitation |
| IMB | Institute of Molecular Biology |
| IR | Intron retention |
| ISE | Intronic splicing enhancer |
| ISS | Intronic splicing silencer |
| JNK | cJun N-terminal kinase |
| KD | Knock down |
| KH | K Homology domain |
| LFC | Log$log_2$ fold change |
| mRNA | Messenger RNA |
| nm | Nanomolar |
| MXE | Mutually exclusive exon |

*Continued on next page*

Table 1 – *Continued from previous page*

| | |
|---|---|
| NCBI | National Center for Biotechnology Information |
| nt | Nucleotide |
| PAR-CLIP | Photoactivatable ribonucleoside-enhanced crosslinking and immunoprecipitation |
| PCR | Polymerase chain reaction |
| PPT | Polypyrimidine tract |
| pre-mRNA | Precursor messenger RNA |
| PSI | Percent spliced in |
| R | Arginine |
| RBP | RNA-binding protein |
| RI | Retained intron |
| RIP | RNA immunoprecipitation |
| RIP-Seq | RNA immunoprecipitation sequencing |
| RNA | Ribonucleic acid |
| RNA-Seq | RNA sequencing |
| RNase | Ribonuclease |
| RNP | Ribonucleoprotein |
| RRM | RNA recognition motif |
| RRMH | RNA recognition motif homolog |
| rRNA | Ribosomal RNA |
| RT | Reverse transcription |
| RT-qPCR | Reverse transcription-quantitative polymerase chain reaction |
| S | Serine |
| SAM | Sequence Alignment Map |
| SDS-PAGE | Sodium dodecyl sulphate-polyacrylamide gel electrophoresis |
| SELEX | Systematic evolution of ligands by exponential enrichment |
| SNP | Single nucleotide polymorphism |
| snRNP | Small nuclear ribonucleoprotein |
| SR proteins | Serine, arginine-rich proteins |
| SRSF6 | Serine-, arginine-rich splicing factor 6 |
| siCTL | Control siRNA |
| siSRSF6 | SRSF6 siRNA |
| siRNA | small interfering RNA |

*Continued on next page*

Table 1 – *Continued from previous page*

| | |
|---|---|
| T1D | Type 1 diabetes |
| T2D | Type 2 diabetes |
| tRNA | Transfer RNA |
| U | Uridine |
| U2AF65 | U2 auxiliary factor 65 kDa subunit |
| UMI | Unique molecular identifier |
| UTR | Untranslated region |
| UV | Ultraviolet |
| WT | Wild type |
| ZnF | Zinc finger |

# Zusammenfassung

Das Transkriptom eukaryotischer Zellen befindet sich in einem dynamischen, sich stetig verändernden Prozess. Viele regulatorische Mechanismen werden dabei von so genannten RNA-bindenden Proteinen (RBPs) kontrolliert. Diese interagieren mit spezifischen Bindestellen auf der RNA. Eine wichtige Methode zur Analyse des Bindeverhaltens von RBPs ist „individual-nucleotide resolution UV crosslinking and immunoprecipitation“ (iCLIP). Dabei werden die Protein-RNA-Kontakte in der Zelle mit Hilfe von UV-Strahlen fixiert. Anschließend kann ein bestimmtes RBP durch einen Antikörper extrahiert, und die daran gebunden RNA-Moleküle sequenziert werden. Dabei entstehen Millionen kleiner Sequenzstücke („reads“), welche dann zum Beispiel gegen das menschliche Genom aligniert werden können, um Rückschlüsse auf die Position des RBPs auf der RNA zuzulassen. Ziel meiner Arbeit war es, durch die Auswertung von iCLIP-Daten die RBP-getriebene RNA-Prozessierung in Eukaryoten besser zu verstehen.

Die hier vorgelegte Arbeit lässt sich in zwei Abschnitte unterteilen. Im ersten Teil wird eine Verfahrensweise zur computergestützten Auswertung von iCLIP-Daten beschrieben. Es existieren zwar bereits andere Verfahrensweisen, diese sind allerdings entweder unvollständig oder beschreiben eine Nischenlösung für eine bestimmte Fragestellung. Dabei wird in der Regel für das jeweils zu untersuchende RBP und das verwendete experimentelle Setup eine speziell angepasste Art der Datenauswertung verwendet. Dadurch sind die Ergebnisse verschiedener RBPs nur schwer miteinander zu vergleichen. Dies ist aber gerade bei komplexeren Fragestellungen zwingend erforderlich. Unter der Verwendung bereits publizierter iCLIP-Daten des Spleißfaktor „U2 small

nuclear RNA auxiliary factor 2“ (U2AF2/ U2AF65) wurde deshalb eine standardisierte Verfahrensweise zur iCLIP-Datenauswertung entwickelt.

Im zweiten Abschnitt der Arbeit wird das Bindeverhalten des RBPs „serine and arginine rich splicing factor 6“ (SRSF6) im Kontext von Diabetes analysiert, indem die im ersten Teil definierte Verfahrensweise Anwendung findet. Es ist bereits bekannt, dass die Expression von SRSF6 in Pankreas-Zellen durch das Gen „GLI-similar 3“ (*GLIS3*) reguliert wird. In Typ-1- und Typ-2-Diabetes (T1D, T2D) ist die Expression von GLIS3 gestört, was die Expression von SRSF6 beeinflusst. SRSF6 fungiert in der Zelle als ein Spleißfaktor, sodass eine Veränderung des Expressiosniveaus das gesamte Transkriptom beeinflussen kann. Ziel dieses Teils der Arbeit ist es, die SRSF6-abhängigen Veränderungen im Transkriptom von Beta-Zellen zu beschreiben, um daraus Rückschlüsse über dessen Regulierung und den Einfluss auf Diabetes zu gewinnen. Dazu wurden sowohl iCLIP- als auch RNA-Seq-Daten aus der menschlichen Pankreas-Ziellinie EndoC-$\beta$H1 analysiert.

## Eine Verfahrensweise zur Auswertung von iCLIP-Daten

Die Auswertung von iCLIP-Daten orientiert sich in der Regel an drei Schritten. Zunächst werden die Daten bezüglich ihrer Qualität geprüft. Bei diesem Schritt werden zum Beispiel reads, welche mit Sequenzierfehlern behaftet sind, aussortiert. Anschließend werden im sogenannten „peak calling“ Akkumulationen von reads bestmöglich in einzelne Bindestellen zusammengefasst. Im letzten Schritt wird dann versucht Rückschlüsse über das RBP aufgrund des beobachteten Bindeverhaltens zu ziehen.

Das initiale Filtern der U2AF65 iCLIP-Daten lieferte 134.386.066 einzelne reads, welche sich auf vier biologische Replikate aufteilten. Diese wurden zu einem Meta-Replikat zusammengefasst und Bindestellen wurden mit PureCLIP vorhergesagt. PureCLIP, wie auch andere Programme, sagt Bindestellen mit unterschiedlicher Breite vorher. In unserem Fall erstreckten sich diese von einem Nukleotid, bis hin zu einigen Bindestellen mit über 40 Nukleotiden an Breite. Deshalb habe ich eine Methode zum Zusammenfassen und Vereinheitlichen der Bindestellenbreite entwickelt. Dabei wurden in einem definierten Fenster benachbarte Bindestellen zunächst in Regionen zusammengefasst. Anschließend wurden alle Regionen, welche größer als die Zielbreite waren, iterativ aufgeteilt, wohingegen kleinere Regionen symmetrisch erweitert wurden. Im Falle von U2AF65 konnten so ca. 300.000 Bindestellen mit neun Nukleotiden in der Breite definiert werden.

Ein Hauptproblem aller bisher existierenden peak calling Programme ist die Verwendung von Replikaten. Es ist unbestritten, dass mehrere biologische Replikate von Nöten sind, um signifikante und aussagekräftige Ergebnisse zu erzielen. Aus diesem Grund wurde im Rahmen dieser Arbeit ein Vorgehen entwickelt, welches es erlaubt, mit aktuell existierenden Programmen trotzdem Replikate einzubeziehen. Dabei wurde zunächst das Signal aller Replikate in einem Meta-Replikat zusammengefasst und als Input für PureCLIP verwendet. Die initialen PureCLIP-Regionen wurden, wie oben beschrieben, in Bindestellen von neun Nukleotiden Breite zusammengefasst. Anschließend wurden alle so definierten Bindestellen bezüglich ihrer Replikat-Unterstützung untersucht. Dabei wurde zunächst für jedes Replikat ermittelt, wie viele reads in eine definierte Bindestelle fallen. Daraus konnte ein prozentualer Schwellenwert für jedes Replikat abgeleitet werden. Im Falle von U2AF65 wurde für das 10% Perzentil ausgewählt. Für jede Bindestelle wurde verlangt, dass dieser Schwellenwert von mindestens drei der vier Replikate erfüllt sein musste. Dieser Filter reduzierte die Anzahl um rund 17,5%, von den ursprünglichen knapp 300.000 auf nun 248.916 Bindestellen. Mithilfe dieses Verfahrens ist es möglich, mehrere Replikate einzubeziehen und damit sowohl sensitiver als auch spezifischer Bindestellen zu detektieren.

Teil einer jeden iCLIP-Analyse ist die Beschreibung des Bindespektrums des RBPs. Dazu müssen orthogonale Informationen, wie beispielsweise Gen-Annotationen, mit den definierten Bindestellen zusammengebracht werden. An dieser Stelle habe ich die entscheidenden Gesichtspunkte bei der Auswahl geeigneter Annotationsquellen beschrieben. In der vorliegenden Analyse wurden Annotationen von GENCODE verwendet. Es konnte gezeigt werden, dass U2AF65 hauptsächlich in Genen bindet, welche für ein Protein kodieren. Außerdem konnte beobachtet werden, dass 81% aller Bindestellen in Introns liegen; gefolgt von kodierenden Sequenzabschnitten (CDS, 10%) und den beiden 5' und 3' untranslatierten Regionen (3'UTR, 7%; 5'UTR, 1%). Dies bestätigt, dass U2AF65 als ein Spleißfaktor an der Erkennung der Grenze zwischen Exons und Introns beteiligt ist und somit potenzielle Spleißstellen definiert. Das Erkennen dieser Grenze ist in der Regel der Start der Spleißreaktion, für welche die Bindung von U2AF65 entscheidend ist.

Der in diesem Abschnitt beschriebene Teil der Arbeit legt einen Grundstein für eine einheitliche Prozessierung und Auswertung von iCLIP-Daten. Dadurch ist es beispielsweise möglich, das Bindeverhalten von RBPs über verschiedene Zelltypen oder Mutationen hinweg zu untersuchen, da der bindespezifische Fußabdruck des RBPs genauer charakterisiert werden kann. Man erhält zum

Beispiel Zugang zu Bindestellen auf Transkripten, welche im aktuellen Zelltyp nur niedrig abundant sind. Zusätzlich ist es auch möglich, Bindestellen mit geringerer Affinität zum untersuchten RBP zu finden, welche oftmals auf eine Sekundärfunktion des RBPs hindeuten. Daher ist es wichtig das Bindespektrum in diesem frühen Stadium der Analyse möglichst ganzheitlich zu definieren, ohne dabei jedoch Abstriche bei der Genauigkeit zu machen. Ein so definiertes Bindespektrum ist potenziell in der Lage die Genauigkeit von nachgeschalteten Vorhersagen zu verbessern. Programme, wie zum Beispiel GraphProt, versuchen genau die Bindestellen zu finden, welche aufgrund des aktuellen experimentellen Rahmens nicht gefunden werden können. Dazu werden zum Beispiel Sequenz- oder Strukturinformationen von beobachteten Bindestellen herangezogen. Es liegt also nahe, dass eine diverse initiale Beschreibung dieser beobachteten Bindestellen solche Vorhersagen positiv beeinflussen kann.

## Die Rolle von SRSF6 in pankreatischen Beta-Zellen

Wie eingangs schon erwähnt, führt die verminderte Expression von SRSF6 in pankreatischen Beta-Zellen zu einer erhöhten Apoptose der Zellen. Um den Einfluss der SRSF6-Bindung in Beta-Zellen zu beschreiben wurden iCLIP-Experimente in vier biologischen Replikaten in der Zelllinie EndoC-$\beta$H1 durchgeführt. Für die Charakterisierung des Bindeverhalten wurde zunächst die oben beschriebene Verfahrensweise angewendet. Damit konnten 160.320 Bindestellen identifiziert werden. Diese waren auf insgesamt 8.533 Gene verteilt, von denen 93% für ein Protein kodieren. Innerhalb dieser Gene konnten wir eine Präferenz für die Bindung von kodierenden Sequenzabschnitten, im Vergleich zu intronischen oder untranslatierten Regionen feststellen. Das genaue Sequenzmotiv, welches SRSF6 auf der mRNA erkennt und bindet, war allerdings unbekannt. Um dieses genauer zu charakterisieren wurden Tripletts sowie Pentamere in definierten Fenstern um die Bindestellen herum gezählt. Dabei konnte eine erhöhte Häufigkeit der Pentamere GAAGA, AGAAG, AAGAA und des Tripletts GAA festgestellt werden. Dies deutete auf ein Purinreiches Bindemotiv hin. Weiterhin konnte beobachtet werden, dass sich die Frequenz von GA-reichen Pentameren ab 50 Nukleotiden vor der Bindestelle bis ca. 25 Nukleotide nach der Bindestelle erhöhte. Das deutete darauf hin, dass SRSF6 GA-reiche Regionen erkennt und sich gegen Ende dieser positioniert. Die Triplett-Analysen ergaben weiterhin, dass sich ununterbrochene Wiederholungen von GAA positiv auf die Binderstärke auswirkten. Am deutlichsten war dies bei drei Wiederholungen des GAA-Tripletts zu sehen. Komplementiert wurden diese Beobachtungen durch eine *de novo* Motivsuche mit DREME, welche ebenfalls ein GA-reiches Sequenzmotiv identifizierte. Zudem ist das hier beschriebene Motiv von SRSF6 sehr ähnlich zu seinem orthologen Partner in Maus, sowie zu dem des „serine and arginine rich splicing factor 4“ (SRSF4), welches ebenfalls erst kürzlich in Mauszellen beschrieben wurde. Es ist bekannt, dass sich SR-Proteine gegenseitig regulieren, insbesondere die Kreuzregulation zwischen SRSF6 und SRSF4. Daher ist es plausibel, dass beide Proteine eine ähnliche Sequenzspezifität aufweisen. Mit der obigen Analyse konnten wir meines Wissens nach erstmals das Bindemotiv des menschlichen SRSF6-Proteins genau *in vivo* beschreiben, was den bisherigen Wissensstand deutlich verbessert.

Wie schon beschrieben, ist SRSF6 ein wichtiger Spleißfaktor, welcher in die Erkennung von Spleißstellen involviert ist. Dabei agiert das Protein meist als Verstärker, sodass insbesondere schwache Spleißstellen von einer SRSF6-Bindung profitieren. Es verstärkt diese und beeinflusst so die Definition von Exon-Intron-Grenzen, auch im Kontext von alternativem Spleißen. Bei diesem Prozess werden verschiedene Exons der „prä-messenger RNA“ (prä-mRNA) eines Transkripts zu unterschiedlichen reifen mRNA-Molekülen zusammengesetzt. Um diejenigen Gene zu ermitteln, welche durch SRSF6 in ihrem Spleißen betroffen sind, wurden RNA-Sequenzierungsdaten (RNA-Seq) aus EndoC-$\beta$H1 Zellen mit *SRSF6* im „knock-down“ (KD) und im „wildtype“ (WT) miteinander verglichen. Es konnten dabei 1.212 signifikant veränderte Kassetten-Exons (CE) beobachtet werden. Bei diesem klassischen alternativen Spleißereignis wird ein Exon entweder in das Transkript eingebaut oder übersprungen. Es existiert demnach im Kontext eines Spleißereignisses immer eine so genannte „inclusion“ und eine „skipping“-Isoform. Es konnte gezeigt werden, dass in ca. zwei Drittel aller Fälle die „skipping“-Isoform gegenüber der „inclusion“-Isoform vorherrschend war. Demzufolge scheint SRSF6 in diesen Fällen eine stabilisierende Wirkung auf das Spleißen von Exons zu haben. Das Fehlen von SRSF6 führt somit zu einem verminderten Einbau solcher Exons. Im Allgemeinen passen diese Beobachtungen zu der bekannten Funktion von SRSF6 als Spleißverstärker.

Im letzten Teil der Analyse habe ich die oben definierten Bindestellen mit den regulierten alternativen Exons kombiniert. Dazu wurden alle solche CE-Ereignisse betrachtet, welche entweder direkt auf dem regulierten Exon, oder auf einem der flankierenden konstitutiven Exons eine SRSF6-Bindestelle aufwiesen. Das Resultat war eine „RNA-Splicing-Map“, welche SRSF6-Bindung auf potenziell direkt regulierten Exons darstellt. Hierbei zeigte sich, dass alternative Exons, welche durch den *SRSF6* KD weniger häufig in die finale Isoform eingebaut wurden, eine verstärkte Bindung von SRSF6 direkt auf dem alternativen Exon aufwiesen. In dem umgekehrten Fall von hochregulierten alternativen Exons war die SRSF6-Bindung auf den flankierenden Exons im WT deutlich verstärkt. Dies deutet darauf hin, dass SRSF6 als Spleißverstärker spezifische „exonic splice enhancer“ (ESE) Sequenzelemente auf alternativen Exons erkennt. Das Spleißen dieser Exons wird demnach von der direkten Bindung durch SRSF6 reguliert. Dabei ist es durchaus denkbar, dass längere GA-reiche ESEs die Bindung von mehreren SRSF6-Proteinen begünstigen. Diese interagieren möglicherweise auch untereinander mittels der RS-Domäne, welche aber auch andere Spleißfaktoren rekrutieren kann.

Einige der so durch SRSF6 regulierten Gene sind für ihren Einfluss auf Typ-1- und Typ-2-Diabetes bekannt. Diese Gene weisen genetische Prädispositionen auf, sodass Mutationen meist in Zusammenhang mit der Krankheit stehen. Es konnte beobachtet werden, dass Prädispositionsgene vermehrt in Prozesse der intrazellulären Signalweiterleitung, aber auch in allgemeinere Funktionen zur Aufrechterhaltung des Zellzyklus eingebunden sind. Es liegt deshalb nahe, dass ein verändertes Spleißverhalten einer Vielzahl solcher Gene die Homöostase der Beta-Zellen nachhaltig beeinflusst und möglicherweise zu ihrem vermehrten Absterben führen kann. Dieses deutet auf einen Zusammenhang der Beta-Zellen-Apoptose mit der Regulation von SRSF6 durch GLIS3 hin. Dabei führt eine Verminderung des Expressionsniveaus von GLIS3 auch zu einer entsprechenden Reduktion der SRSF6-Transkriptmenge. Dadurch kommt es insbesondere bei Prädispositionsgenen für Diabetes zu einer Veränderung im Spleißen alternativer Isoformen, da diese durch direkte SRSF6-Bindung reguliert werden. Mit der vorliegenden Arbeit habe ich starke Hinweise liefern können, dass der Spleißfaktor SRSF6 maßgeblich an der Aufrechterhaltung der Funktion von Beta-Zellen beteiligt ist. Insbesondere die genaue Definition des Bindemotivs erlaubt möglicherweise eine bessere Vorhersage von ESE-Elementen, welche durch SRSF6 erkannt werden. Es ergeben sich dadurch auch potenzielle medizinische Behandlungsmöglichkeiten. So könnten beispielsweise Bindestellen auf alternativen Exons mit Hilfe von „antisense“-Oligonukleotiden (ASOs) blockiert werden, was sich positiv auf das Überleben von Beta-Zellen auswirken kann.

Zusammenfassend lässt sich sagen, dass ich mit der hier vorgelegten Arbeit eine umfassende Verfahrensweise zur Prozessierung von iCLIP-Daten beschrieben habe. Diese konnte anschließend auf die konkrete Fragestellung nach der spezifischen SRSF6-Bindingung in pankreatischen Beta-Zellen angewendet werden. Dadurch ergaben sich weitreichende neue Erkenntnisse über den Einfluss von SRSF6 auf die Beta-Zellen-Regulation im Zusammenhang mit Diabetes.

# Abstract

Most cellular processes are regulated by RNA-binding proteins (RBPs). These RBPs usually use defined binding sites to recognize and directly interact with their target RNA molecule. Individual-nucleotide resolution UV crosslinking and immunoprecipitation (iCLIP) experiments are an important tool to describe such interactions in cell cultures in-vivo. This experimental protocol yields millions of individual sequencing reads from which the binding spectrum of the RBP under study can be deduced. In this PhD thesis I studied how RNA processing is driven from RBP binding by analyzing iCLIP-derived sequencing datasets.

First, I described a complete data analysis pipeline to detect RBP binding sites from iCLIP sequencing reads. This workflow covers all essential processing steps, from the first quality control to the final annotation of binding sites. I described the accurate integration of biological iCLIP replicates to boost the initial peak calling step while ensuring high specificity through replicate reproducibility analysis. Further I proposed a routine to level binding site width to streamline downstream analysis processes. This was exemplified in the reanalysis of the binding spectrum of the U2 small nuclear RNA auxiliary factor 2 (U2AF2, U2AF65). I recaptured the known dominance of U2AF65 to bind to intronic sequences of protein-coding genes, where it likely recognizes the polypyrimidine tract as part of the core spliceosome machinery.

In the second part of my thesis, I analyzed the binding spectrum of the serine and arginine rich splicing factor 6 (SRSF6) in the context of diabetes. In pancreatic beta-cells, the expression of SRSF6 is regulated by the transcription factor GLIS3, which encodes for a diabetes susceptibility gene. It is known that

SRSF6 promotes beta-cell death through the splicing dysregulation of genes essential to beta-cell function and survival. However, the exact mechanism of how these RNAs are targeted by SRSF6 remains poorly understood. Here, I applied the defined iCLIP processing pipeline to describe the binding landscape of the splicing factor SRSF6 in the human pancreatic beta-cell line EndoC-$\beta$H1. The initial binding sites definition revealed a predominant binding to coding sequences (CDS) of protein-coding genes. This was followed up by extensive motif analysis which revealed a so far, in human, unknown purine-rich binding motif. SRSF6 seemed to specifically recognize repetitions of the triplet GAA. I also showed that the number of contiguous triplets correlated with increasing binding site strength. I further integrated RNA-sequencing data from the same cell type, with SRSF6 in KD and in basal conditions, to analyze SRSF6-related splicing changes. I showed that the exact positioning of SRSF6 on alternatively spliced exons regulates the produced transcript isoforms. This mechanism seemed to control exons in several known susceptibility genes for diabetes.

In summary, in my PhD thesis, I presented a comprehensive workflow for the processing of iCLIP-derived sequencing data. I applied this pipeline on a dataset from pancreatic beta-cells to unveil the impact of SRSF6-mediated splicing changes. Thus, my analysis provides novel insights into the regulation of diabetes susceptibility genes.

# Preface

The content of this PhD thesis is based on collaborative research projects between the group of Dr. Kathi Zarnack (BMLS, Goethe University FB15, Frankfurt am Main), the group of Dr. Julian König (IMB, Mainz) and the group of Dr. Décio L. Eizirik (ULB, Brussels). In my PhD project, I computationally investigated RNA processing mechanisms by analyzing high-throughput sequencing data. All bioinformatic analyses shown in this thesis were performed by myself, with the exceptions of the initial quality control steps, which were performed by Dr. Anke Busch (IMB, Mainz) and Dr. Stefanie Ebersberger (IMB, Mainz). Ines Alvelos (ULB, Brussels) and Reymond Sutandy (IMB, Mainz) performed iCLIP library preparation, optimization, sequencing and additional validations. For the sake of completeness some of the experimental results were included in this thesis. The projects described in this thesis were under the supervision of Dr. Kathi Zarnack and the described results have also been published in the following articles:

Busch, A., **Brüggemann, M.**, Ebersberger, S., and Zarnack, K. (2020). iCLIP data analysis: A complete pipeline from sequencing reads to RBP binding sites. *Methods*, 178:49-62.

Alvelos, M. I., **Brüggemann, M.**, Sutandy, F. R., Juan-Mateu, J., Colli, M. L., Buch, A., Lopes, M., Castela, Â., Aartsma-Rus, A., König, J., et al. (2020). The RNA binding profile of the splicing factor SRSF6 in immortalized human pancreatic $\beta$-cells. *Life science alliance*, 4(3).

# 1 Introduction

## 1.1 The complexity of the human transcriptome

### 1.1.1 RNA regulatory mechanisms

The flow of genetic information within a cell was first described in 1957 (Crick, 1958). This process, known as the 'central dogma of molecular biology', consists of two main steps. First information organized in genes and encoded in the deoxyribonucleic acid (DNA) is transcribed into ribonucleic acid (RNA). Next, RNA molecules are translated into proteins. RNA transcription and RNA processing are steps in between genes and proteins that serve as additional layers of regulation. Several forms of these RNA molecules exist, with the messenger RNA (mRNA) being the form that is explicitly translated into the final protein. Other RNA forms like transfer RNAs (tRNAs) or ribosomal RNAs (rRNAs), to name only the most prevalent, serve in regulatory processes to guide gene expression. The sum of all expressed types of RNA molecules comprise the transcriptome of a cell or tissue.

The mRNA life cycle consists of a complex interplay between RNA-binding proteins (RBPs) and further non-coding RNAs. These can interact at many different positions during mRNA transcription, maturation, transport, translation or degradation (Djebali et al., 2012). The initial steps of the mRNA life are shaped by transcription and maturation processes in the nucleus. In eukaryotes such as human, mRNA synthesis is catalyzed by the RNA Poly-

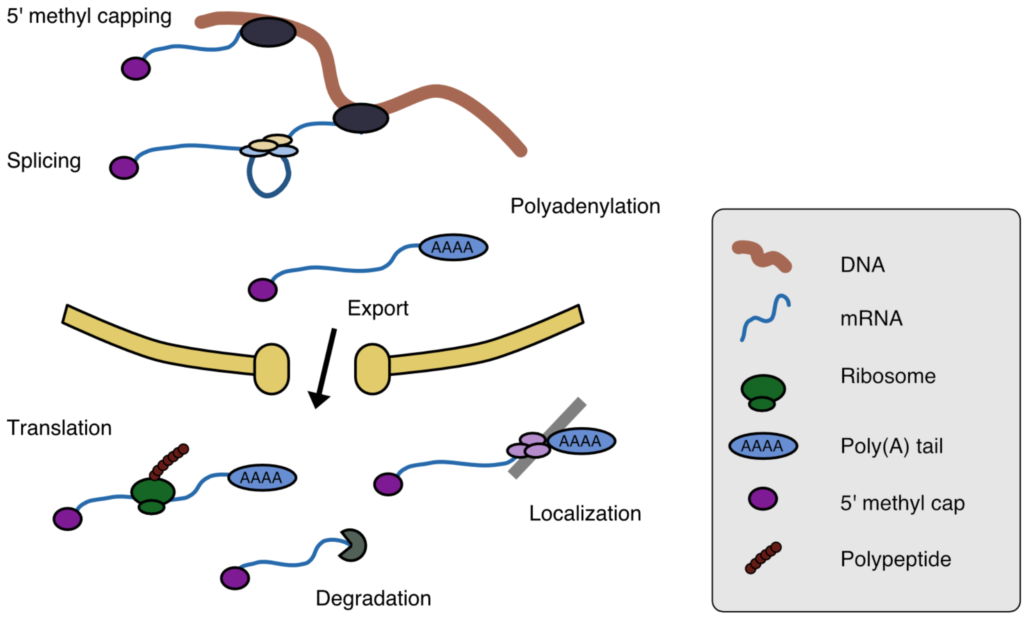

**Figure 1.1: Schematic overview of the mRNA life cycle.** Genetic information encoded in the DNA is transcribed into the mRNA in the nucleus. 5' capping and splicing are essential co-transcriptional modifications. Pre-mRNA processing ends with the polyadenylation at the 3' end. The mature mRNA is exported into the cytoplasm. Additional factors control the mRNA life in the cytoplasm, guiding translocation, transcription and degradation.

merase II (Cramer et al., 2001). The polymerase binds to the promotor region of the DNA and transcribes the DNA sequence. The resulting pre-mRNA is composed of exons and introns. Introns are non-coding sequence parts of the mRNA which are not translated into the final protein. This allows for alternative rearrangements during splicing and also provides anchors and binding points for RBPs needed in downstream processing steps. Exons represent coding parts and will be fused together in order to provide a single translatable sequence. The majority of the exonic sequence is translated into the protein, with exception to the 5' and 3' untranslated regions (UTRs). These regions serve as major anchor points for proteins to alter the fate of the mRNA, such as providing binding regions for nuclear export factors, directing the mRNA to its final location or to regulate translation. Already during transcription, a series of modifications is applied to the pre-mRNA. At first a cap structure is added to the 5' end that prevents the molecule from degradation via exonucleases (Proudfoot, 2000). This is followed by the splicing process in which exons are fused together while intronic sequence parts are removed. Lastly multiple adenines are added to the 3' end. This polyadenylation step completes the maturation process in the nucleus and leaves the mRNA ready for export, usually facilitated by nuclear export factors (Stewart, 2019). It is worth to mention that transcription in general is a heavily parallelized process and also many of the maturation steps happen in a co-transcriptional manner. Once exported to the cytoplasm an mRNA can be either directly translated into protein by the ribosome or stored for later usage. The mRNA can also be transported to a specific location in the cell prior to translation, such as to specific compartments or positions. The last step marks the degradation. One example is the nonsense-mediated decay (NMD) surveillance mechanism (Kurosaki and Maquat, 2016). It is a cellular pathway specifically for the degradation of aberrant mRNAs that harbor a premature stop codon. Such an error can be introduced in any of the maturation steps, for instance during transcription due to polymerase slippage or during splicing. NMD thus prevents the translation of truncated proteins and thereby keeps cell homeostasis.

### 1.1.2 Splicing and alternative splicing

The pre-mRNA splicing reactions are carried out by a large and highly dynamic RNA-protein complex, known as the spliceosome. The spliceosome catalyzes all reactions needed to remove the intron sequence between two exons in a first step, while fusing these exons together in a second step. In general, the spliceosome forms distinct complexes with well-known structures. These complexes are formed by the interaction of five small nuclear ribonucleoproteins (snRNPs), known as U1, U2, U4, U5 and U6, all of which consist of small proteins together with uridine-rich small nuclear RNAs (Shi, 2017; Lerner and Steitz, 1979). These snRNPs are considered the main building blocks of the spliceosome. However, a large amount of other proteins is required as well to catalyze each step of the splicing reaction.

Each intron consists of distinct sequence elements that are required for the initial spliceosome assembly. These are the 3' and 5' splice sites (3'SS and 5'SS, respectively), which mark the start and end position of the exon, as well as the branchpoint sequence (BPS) and the poly-pyrimidine tract (PPT). The 5'SS is also known as the donor site of the reaction, whereas the 3'SS is known as the acceptor site (Frendewey and Keller, 1985). The spliceosome assembly starts with the recognition of the 5'SS by the U1 snRNP, with direct base-pairing (Rogers and Wall, 1980). Of note, this interaction is typically influenced by members of the serine arginine-rich splicing factors (SR proteins) to modulate splicing efficiency (Jeong, 2017). 3'SS recognition starts with the binding of U2 small nuclear RNA auxiliary factor (U2AF), which is a heterodimer consisting

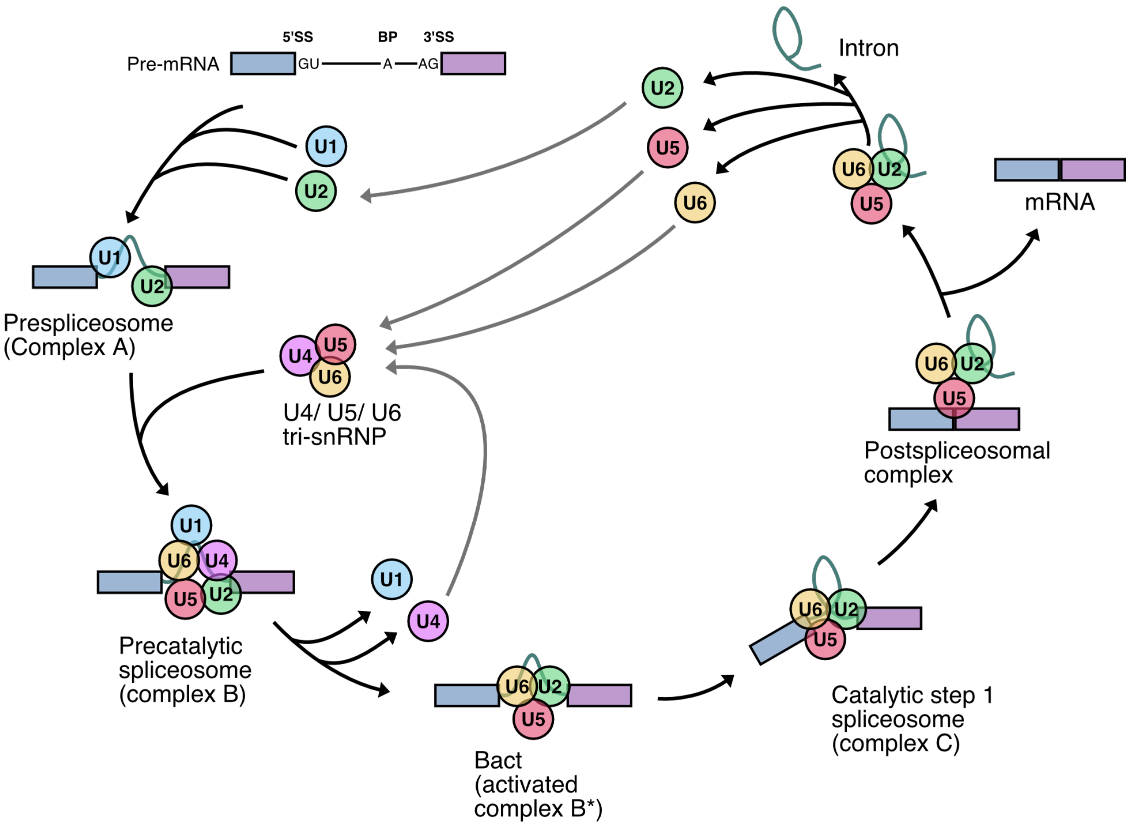

**Figure 1.2: Cycle of the spliceosome assembly and disassembly.** The spliceosome assembles on the pre-mRNA forming distinct complexes. The assembly is characterized by the stepwise interaction of small nuclear ribonucleoproteins (snRNPs). U1 and U2 facilitate the initial splice sites recognition and form complex A. The addition of the tri-snRNP, consisting of U4, U5, and U6 form complex B. U1 and U4 are released which results in the activated complex B*. Here the first catalytic reaction leads to complex C. The postspliceosomal complex contains the fused exons and the lariat structure. Lastly the complex disassembles and releases all produces, as well as U2, U5 and U6. Figure adapted from (Lee and Rio, 2015).

of the subunits U2AF1 and U2AF2 (also known as U2AF35 and U2AF65, respectively). Thereby U2AF1 recognizes the AG dinucleotide of the splice site itself, while U2AF2 binds to the PPT directly upstream of it (Berglund et al., 1998). This interaction is stabilized by the splicing factor 1 (SF1) which binds to the branchpoint sequence while interacting with U2AF2. Once fully assembled, U2AF recruits the U2 snRNPs which marks the beginning of the splicing reaction and forms the pre-spliceosome complex A. Next U4, U5 and U6 snRNPs join to form the precatalytic spliceosome (complex B). The complex undergoes further conformational changes, releasing U1 and U4, forming the activated spliceosome (B* complex). This complex carries out the branching reaction in which the phosphodiester bond at the 5'SS is broken and then fused to the branchpoint sequence. This is followed by the second part of the reaction in the post-spliceosomal complex where exon ligation occurs. Lastly, the spliceosome complex dissolves, freeing the ligated exons, as well as the intron sequence as a lariat structure (Padgett et al., 1984; Wahl et al., 2009; Shi, 2017).

A single human gene usually gives rise to at least two different transcript isoforms (Lee and Rio, 2015; Barbosa-Morais et al., 2012). The process of how and which exons are included in a final isoform is called alternative splicing. This is thought to provide many advantages by providing more flexibility in the genomic coding capacity (Kim et al., 2007). In the case of human, more than 95% of all genes are alternatively spliced (Kornblihtt, 2007). Setting up and maintaining alternative splicing patterns is performed as a combination of *cis*-acting sequence elements, as well as *trans*-acting binding factors (Lee and Rio, 2015). These *cis*-acting sites can be either located in the intronic and exonic part of the pre-mRNA. *Trans*-acting binding factors recognize such sites which can either have a positive effect on the splicing outcome in case of splicing enhancers, or a negative effect in the case of splicing silencers. Another level of regulation is set by the *cis*-acting sequence elements, which are usually short degenerate RNA motifs, that can be recognized by multiple different proteins. This gets even more complex since different splicing regulators might also depend on interacting co-factors as well as the cell type (Havlioglu et al., 2007). Finally, it was also observed that chromatin modification as well as the speed of the transcription influence the alternative splicing pattern (Dujardin et al., 2013). The most prevalent alternative splicing pattern in vertebrates is the exclusion or skipping of an entire exon, also known as cassette exons (CE). Exons can also be mutually exclusive (MXE), where either one or the other of two regulated exons is present in the final transcript. Additionally, exons can be modulated by alternative 3' and 5' splice sites, thus leading to slightly different versions of an exon. Another possible splicing event effects the intronic sequence, where the intron itself can be retained (IR) (Wang et al., 2015). In humans this happens primarily in the untranslated regions (Galante et al., 2004).

In summary, alternative splicing is a highly regulated mechanism that affects mRNA abundance. The final splicing outcome is controlled by specific sequence elements that allow the binding of specific splicing factors. Changes to the splicing pattern may influence and reprogram the entire transcriptome and thus change the fate of a cell or tissue. Unwanted changes to the splicing pattern due to for example mutations lead to a wide range of diseases, such as

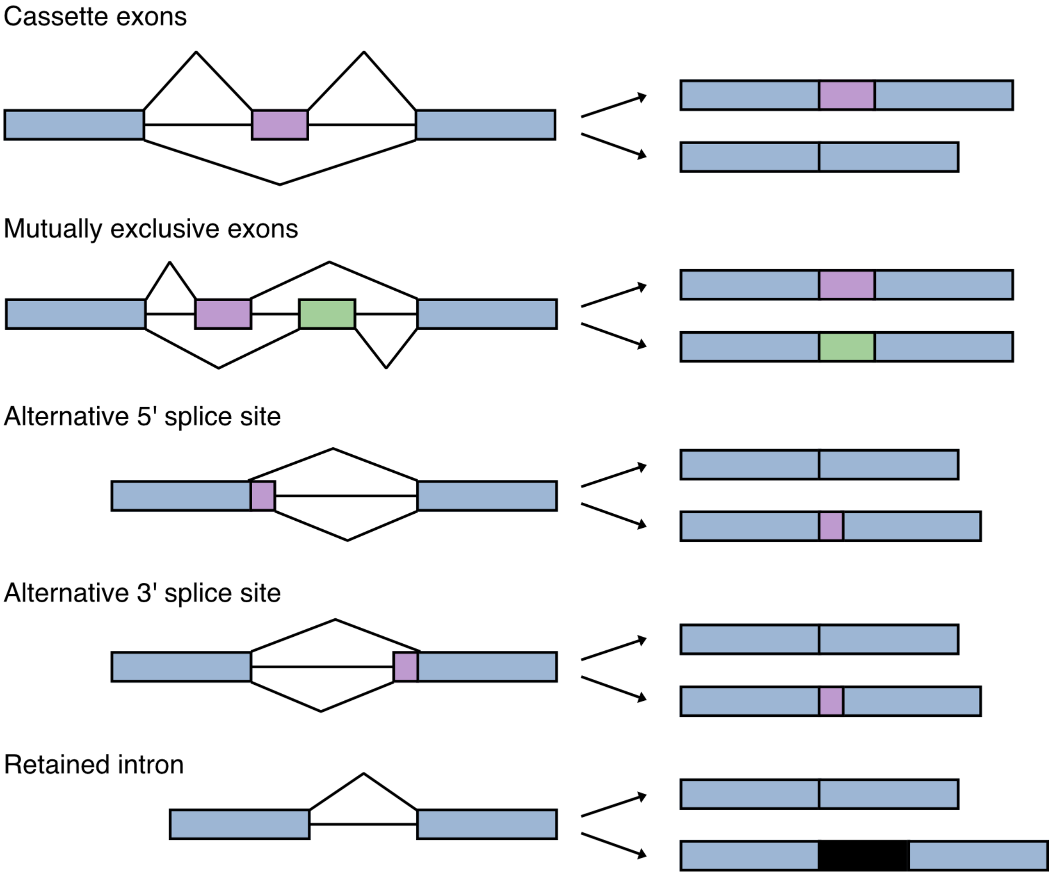

**Figure 1.3: Common mechanisms of alternative splicing.** A number of different alternative splicing processes can give rise to plenty of different transcript variants. Most commonly cassette exon skipping, the mutual inclusion of exons, the usage of alternative 5' and 3' splice sites as well as retaining an entire intron. Figure adapted from (Chen and Weiss, 2015).

neurodegenerative diseases and cancer (Cartegni and Krainer, 2002; Kashima and Manley, 2003; David et al., 2010; David and Manley, 2010).

### 1.1.3 Function and design of RNA-binding proteins

To function properly mRNAs must contain additional information besides the coding sequence. These elements are located in the UTR regions and are recognized by specific RBPs. RBPs are involved in nearly all regulatory processes within the cell, covering a wide range of functionalities. They can assemble in large complexes such as the spliceosome or ribosome but can also act alone directed by specific sequence motifs (Dreyfuss et al., 2002).

To allow for an interaction with the RNA RBPs usually contain specific structural domains (Lunde et al., 2007). Although a huge variety of different RBPs exists, most of them are built from few RNA-binding modules, such as the RNA recognition motif (RRM), the K homology (KH) domain, or zinc-finger type binding motifs (ZnF) (Burd and Dreyfuss, 1994). The variety of functional RBPs arises from the diverse structural arrangement and copy numbers of these domains. This allows for the formation of RNA-recognition units with the ability to express wide ranges of affinities and specifies to defined sequence motifs. These recognition units can be combined with a wide range of catalytic domains, to form enzymes that are specific for their target and the reaction they catalyze (Auweter et al., 2006). Such structural units interact with specific sequence motifs on the side of the RNA. Sequences range from highly specific binding sites to short and degenerate motifs. Depending on the tissue or regulation process the same motif might also be recognized by different RBPs. The affinity of an RBP to such a motif can also change in the context of other RBPs being present, leading to cooperative binding and competition. During alternative splicing for example, splicing enhancer and silencer proteins can compete for the same sequence element (Kyburz et al., 2006; Vagner et al., 2000).

More recently a new class of RBPs was discovered that lacked known RNA-binding domains, described as 'unconventional RBPs' (Hentze et al., 2018). Many of these RBPs were identified on a high-throughput scale using mass spectrometry anlaysis coupled with ultraviolet crosslinking. This suggests a direct form of interaction, some of which might be mediated by intrinsically disordered protein regions (Castello et al., 2016). Interestingly, CLIP-based studies already described many of such 'unconventional RBPs'. For example the alternative splicing regulator 3-hydroxyacyl-CoA dehydrogenase typ2 (HSD17B10) was identified as RBP yeast, but lacks known RNA-binding domains (Beckmann et al., 2015).

RBPs execute their diverse function in the cytoplasm as well as in the nucleus and some also shuttle between both compartments (Cáceres et al., 1998). Nucleic RBPs are typically involved in pre-mRNA maturation such as splicing regulators like SR proteins and heterogeneous nuclear ribonucleoproteins (hnRNPs). Cytoplasmic RBPs on the other hand predominantly bind to mature sequence elements, such as UTRs and CDS. The 3'UTR in particular can be seen as a hotspot for regulation, hosting several *cis*-acting sequence elements to guide RBP binding for translation and localization. Some 3'UTR-bound components are also deposited in the nucleus and travel to the cytoplasm with the mRNA, thus influencing the downstream metabolism (Müller-McNicoll et al., 2016; Singh et al., 2015).

## 1.2 SR proteins shape the transcriptome

### 1.2.1 The SR protein family

An essential protein family for cell survival that acts as core splicing factors are the serine arginine-rich splicing factors (SR proteins). They are involved in key steps of the splicing reaction, such as splicing initiation by interaction with U1 and U2, but also in later spliceosome formation by recruitment of U4, U5 and U6 (Wu and Maniatis, 1993; Blencowe et al., 1999). They assist in splice site recognition and their interplay leads to the stabilization or destabilization of exon boundaries, thus guiding the spliceosome. Usually SR proteins promote exon inclusion when bound to exonic or intronic splicing enhancer sequences (ESE and ISE, respectively) (Lin and Fu, 2007). hnRNP proteins on the other hand often act as splicing repressors binding to exonic and intronic splicing silencer sequence elements (ESS and ISS, respectively) (Zhu et al., 2001). Nevertheless, they also compete for certain sequence elements to either repress or enhance splicing.

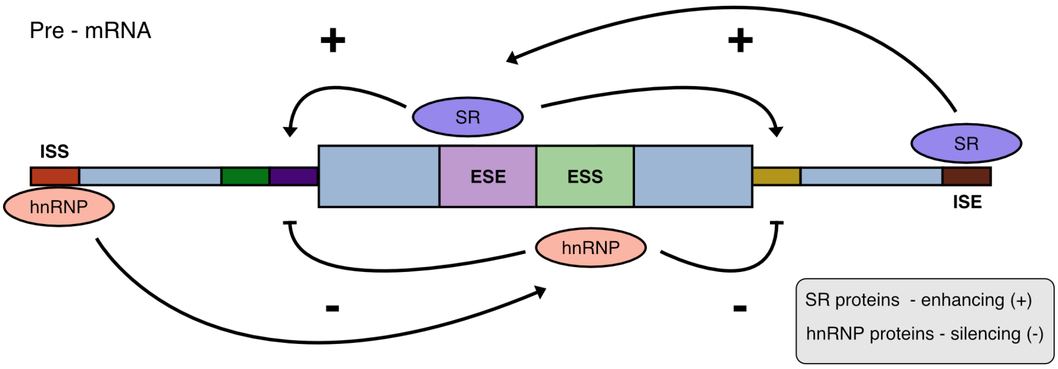

**Figure 1.4: Mechanisms of positive and negative control of pre-mRNA splicing.** The splicing process is influenced by *cis*-acting regions and *trans*-acting factors. Intronic and exonic sequence regions can both act as silencers (ESS, ISS) or enhancers (ESE, ISE). These sequence elements are bound by splicing regulator proteins, such as heterogeneous nuclear ribonucleoproteins (hnRNPs) or serine arginine rich proteins. SR proteins typically promote splicing of a nearby splice site and interact with enhancer regions. On the other side hnRNP proteins usually act to inhibit splicing of nearby splice sites, thus interacting with silencer regions. Figure adapted from (Lee and Rio, 2015).

The SR protein family consists of 12 evolutionary conserved proteins. They are defined by the presence of one or two N-terminal RRM domains, which are followed by an RS domain of at least 50 amino acids with a serine and arginine content of more than 40% (Manley and Krainer, 2010). The RRM domain facilitates the ability to directly bind RNA, whereas the RS domain is needed for protein-protein interactions (Bourgeois et al., 2004). Of particular importance are the serine residues, which regulate the protein activity by their phosphorylation status. Besides their well characterized function in pre-mRNA splicing, SR proteins are also involved in other gene regulatory processes (Sapra et al., 2009). With their ability to cycle between the nucleus and the cytoplasm they bridge the gap between transcriptional and translational regulation processes (Müller-McNicoll et al., 2016). The shuttling process is closely linked to the phosphorylation status of the RS domain. At the end of the splicing cycle SR proteins are dephosphorylated which allows the interaction with the nuclear export machinery. Once in the cytoplasm SR proteins can be re-phosphorylated in order to return to the nucleus. By this, the SR protein activity as well as its cellular localization are determined by the phosphorylation status (Zhou and Fu, 2013).

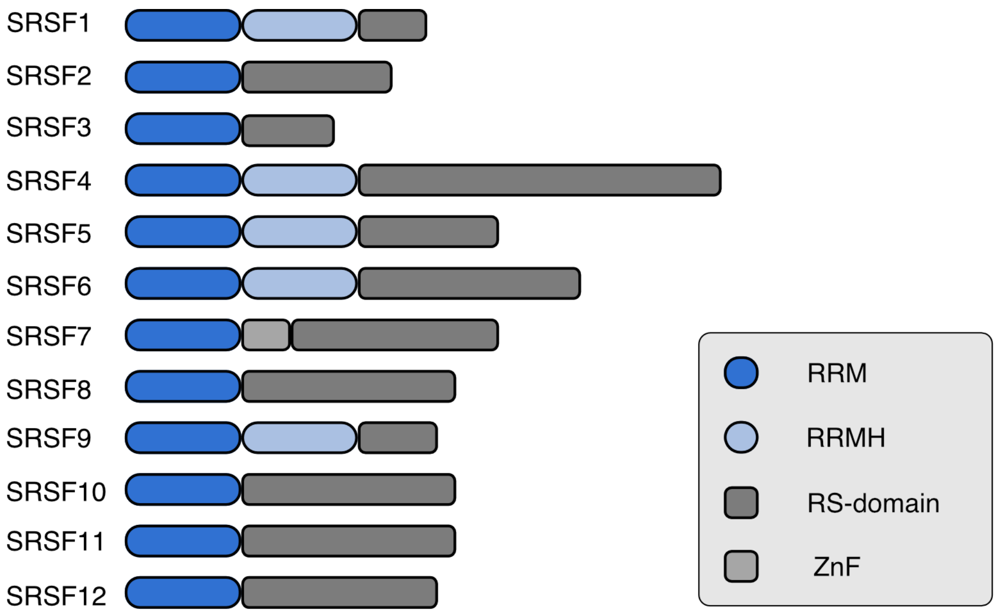

**Figure 1.5: Domain structures of the SR protein family.** SR proteins usually consist of a combination of four different protein domains (RRM, RRMH, RS domain, Znf). RRM, RRMH and Znf domains give rise to a wide range of RNA affinities. The RS domain is of variable size, but characterized by a serine-arginine content of above 40%. Figure adapted from (Änkö, 2014).

Although similar in their composition, individual SR proteins are not functionally equivalent (Zahler et al., 1993). In their RNA-binding capacity various different sequence preferences have been observed, suggesting unique specificities. For SRSF1 and SRSF7 binding to largely purine-rich sequence elements was observed *in vitro* via systematic evolution of ligands by exponential enrichment (SELEX) (Schaal and Maniatis, 1999). SRSF3 on the other hand binds predominantly pyrimidine-rich sequences. Furthermore, *in vivo* approaches such as crosslinking and immunoprecipitation (CLIP) were also applied to study the binding of SR proteins (Sanford et al., 2009). CLIP-Seq approaches are used to capture direct RNA-protein interactions *in vivo* (for details see chapter 1.3). Recently, CLIP-Seq studies revealed the splicing impact of SRSF1 and SRSF2 by cooperative and competitive binding interactions (Änkö, 2014; Müller-McNicoll et al., 2016).

### 1.2.2 SRSF6 a key splicing regulator

One important representative of the SR protein family is SRSF6. As almost all SR proteins it is involved in constitutive and alternative splicing, likewise it shuttles between the nucleus and the cytoplasm. It regulates splicing in a dose-dependent fashion, since overexpression of SRSF6 affects splice site recognition (Screaton et al., 1995). SRSF6 misregulation usually has severe effects on the cell. For example, it has been described as a proto-oncogene contributing to breast, lung, skin and colorectal cancer upon aberrant expression (Karni et al., 2007). Furthermore, it was identified as a master regulator of tenascin-C alternative splicing thus affecting wound healing and hyperplasia (Jensen et al., 2014). It was also shown that SRSF6 expression levels are coupled to those of SRSF4, indicating a compensatory cross-talk between the two proteins. Such co-regulatory effects are well known and described for other SR proteins, such as SRSF2 and SRSF5 (Müller-McNicoll et al., 2016). The human binding motif of SRSF4 has been described *in vivo* as GA-rich (Änkö et al., 2012). The binding motif of SRSF6 on the other hand is not as clearly described. *In vitro* SELEX based studies report differing motifs consisting of UCG and CAG enriched sequence elements (Änkö, 2014; Liu et al., 1998). Contrary to this a recent study described the SRSF6 binding motif in mouse as being GA-rich (Müller-McNicoll et al., 2016). With the known compensatory effects of SRSF4 and SRSF6 a similar binding motif seems reasonable. Yet the description of this motif for SRSF6 in human cells is still missing.

### 1.2.3 The role of SRSF6 in diabetes

Diabetes is a chronic disease characterized by patients with low insulin levels. Pancreatic beta-cells lose their ability to produce insulin, which leads to hyperglycemia and several other short- and long-term complications. Type

1 (T1D) and type 2 (T2D) are the two main forms of the disease. T1D is caused by an autoimmune response, where the loss of beta-cells leads to the failure of the pancreas to produce insulin. T2D is triggered by metabolic stressors, which lead to insulin resistance and lastly result in low insulin levels as well. In both cases however, genetic and environmental interactions trigger the loss or failure of the insulin-producing pancreatic beta-cells (Weir and Bonner-Weir, 2013). Physiologically, insulin-producing pancreatic beta-cells are located in the islet of Langerhans region (also known as pancreatic islets) (Zhou and Melton, 2018). In human, these islets comprise of about 2% of the total pancreatic tissue. Within the islets approximately 60% of the cells are insulin-producing beta-cells, followed by gluacon-producing $\alpha$-cells (30%). The remaining 10% is covered by $\delta$-cells, $\gamma$-cells and $\epsilon$-cells (Ionescu-Tirgoviste et al., 2015; Cabrera et al., 2006). Taken together all of these cells tightly regulate blood glucose levels by hormone secretion (Röder et al., 2016). In T1D these regions face increased inflammation promoted by cytokines. Cytokines are global regulators of the immune system typically responding to infection and inflammation. Different classes of cytokines exist, those that promote the disease (proinflammatory cytokines) and those that promote the healing process (anti-inflammatory cytokines). In T1D, proinflammatory cytokines are released by beta- and immune cells promoting pancreatic islet inflammation (Dinarello, 2000).

Early detection of diabetes remains challenging because of the complex nature of the disease. Genome-wide association studies (GWAS) however enable the usage of high-throughput data to detect single-nucleotide polymorphisms (SNPs) which might contribute to diabetes. Such SNPs can be combined with known annotated genes in order to derive potential diabetes susceptibility genes (Hakonarson and Grant, 2011). For T1D and T2D GWAS studies identified several such susceptibility genes (Barrett et al., 2009; Dupuis et al., 2010). One of these is the Kruppel-like zinc finger protein Gli-similar 3 (GLIS3), which is a known transcription factor (Barrett et al., 2009; Eizirik et al., 2020). The exact mechanism of GLIS3 influences beta-cell function was shown by our collaborators in 2013. They showed that the knockdown (KD) of GLIS3 lead to increased beta-cell apoptosis induced by proinflammatory cytokines. Further they showed that the triggered apoptosis pathway is activated by AS (Nogueira et al., 2013). That dysregulation of splicing can affect beta-cell function and impact their survival is not new. The impact of proinflammatory cytokines on AS changes in human beta-cells is well described (Eizirik et al., 2012). What was new however, was the observation that the activation of beta-cell apoptosis coincided with the inhibition of SRSF6. Explicitly, they found that reduced expression of GLIS3 also leads to a reduced expression of SRSF6 (Nogueira et al., 2013). Next, they described the impact of SRSF6 in a follow-up study and found that it is highly expressed in human pancreatic beta-cells (Juan-Mateu et al., 2018). Its direct downregulation via KD showed the same dramatic effect on beta-cell function and survival as if mediated by GLIS3. They showed that resulting splicing changes impacted diabetes-related biological processes such as insulin secretion and c-Jun N-terminal kinase (JNK) signaling (Juan-Mateu et al., 2018). Thereby they found a link between the expression of SRSF6 and beta-cell survival and death in the context of diabetes. Nevertheless, the exact mechanisms of SRSF6-mediated splicing changes remain an open question.

## 1.3 Approaches to study protein-RNA interactions

### 1.3.1 CLIP technologies

Within a cell, thousands of RBPs are involved in a wide range of cellular mechanisms, ranging from pre-mRNA splicing to translation, localization and degradation. In each of these processes they are presented to a variety of possible RNA targets, while competing or cooperating with other RBPs. It is thus important to study these RBP binding landscapes. Methods to study protein-RNA interactions typically make use of RNA immunoprecipitation (RIP). Formaldehyde can be used to retain protein-RNA interactions and allow the conservation of large RNP complexes (Niranjanakumari et al., 2002). These approaches were subsequently coupled with different read-out strategies, such as microarrays (RIP-Chip) or high-throughput sequencing (RIP-Seq) (Tenenbaum et al., 2000; Zhao et al., 2010). The conservation of large RNP complexes resulted in a rather broad identification of abundant RNAs bound to the RBP. The identification of direct protein-RNA interactions was not possible. To enhance the specificity, UV crosslinking and immunoprecipitation (CLIP) was developed. The idea was to preserve the endogenous protein-RNA contact while adopting stringent purification methods to avoid the detection of co-associations with other RBPs (Ule et al., 2003). Based on this idea a variety of different derivates were developed, each of which improved on different parts of the protocol, such as RNA fragmentation, RBP purification or cDNA library preparation. In the following I give a description of the individual steps, emphasizing the general concepts rather than the details of a specific protocol.

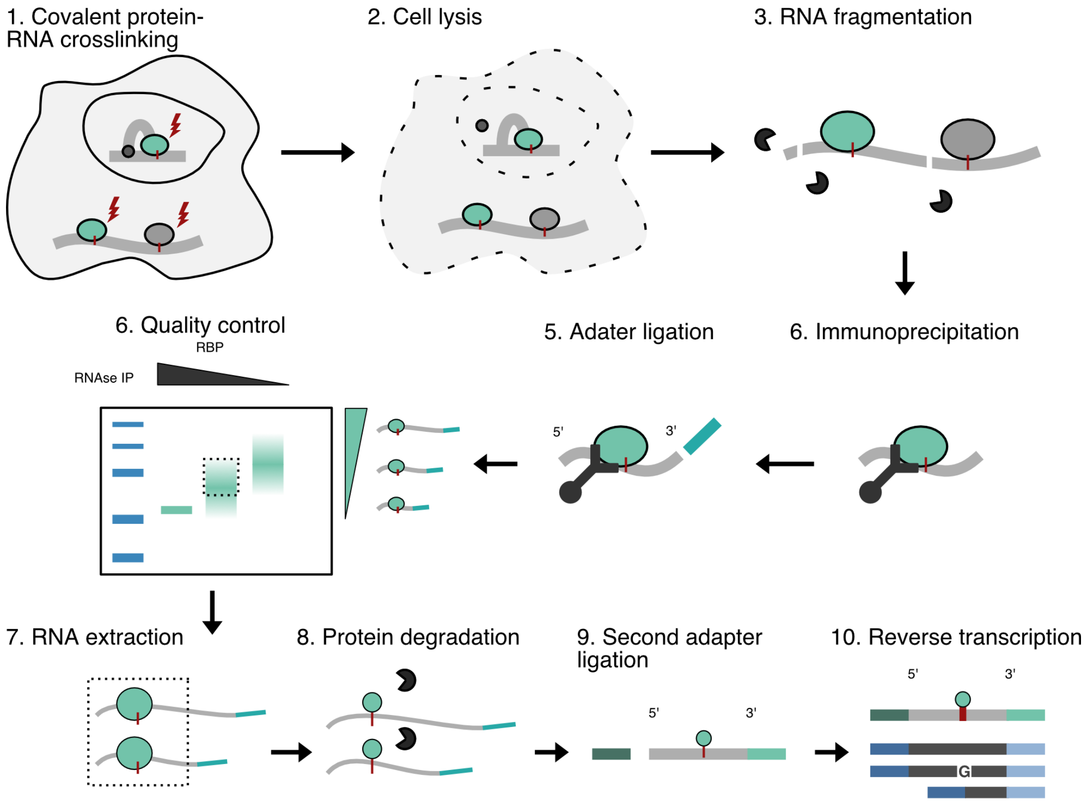

**Figure 1.6: The core steps of major variants of CLIP.** A covalent protein-RNA crosslink is introduced by exposure to UV light. Cellular structures are cracked open and the RNA is fragmented. The RBP of interest is captured by immunoprecipitation using antibodies. For reverse transcription the primer sequence is ligated to the 3' ends. All protocols usually include a quality control step to monitor the size of the fragmented RNA complexes. Complexes of the desired size are extracted and the protein particle is digested to prepare the RNA for reverse transcription. Depending of the protocol different approaches exist to ligate the second primer to the 5' end of the RNA molecule. cDNAs from all protocols are forwarded to high-throughput sequencing. Figure adapted from (Lee and Ule, 2018).

All CLIP-based approaches have an initial crosslinking step in common. Ultraviolet (UV) light introduces a covalent bond between the RBP and the bound RNA. This crosslink requires direct contact between the nucleobase of the RNA with an amino acid of the RBP. Thus, such covalent bonds preserve the original protein-RNA contact while being very stable, which allows for further stringent purification steps. Depending on the exact protocol, different wave-lengths and exposure times are used. In a special variant of the strategy cells are pre-incubated with photoactivatable ribonucleoside 4-thiouridine (4SU), which alters the crosslinking behavior (PAR-CLIP) (Hafner et al., 2010). The UV crosslinking step is followed by cell lysis. Here, all other nascent protein-RNA and protein-protein are disrupted, usually by buffering with ionic detergents. Thus, only the covalently crosslinked protein-RNA interactions remain intact in the lysate. Next, RNAs are split into fragments of cloneable size by treatment with RNase. This also removes the potential to co-purify other RBPs that crosslink to the same RNA. This step leaves the RBP of interest crosslinked to a small strand of RNA. The length of this strand has to be carefully controlled. Overdigestion for example can lead to the accumulation of short RNA fragments which narrows down the potential size distribution of cDNAs (Haberman et al., 2017). Insufficient RNA digestion on the other side might lead to co-purification of additional RBPs. Some protocols perform RNase treatment directly in the lysate, while others use a bead-based RNase treatment strategy (Hafner et al., 2010; Kargapolova et al., 2017). Purification of the RBP-RNA complex usually requires antibodies to be available for the immunoprecipitation step. If this is not the case, tagged RBPs can be expressed transiently from plasmids or genome editing can be used to express an epitope-tagged version of the RBP (Van Nostrand et al., 2017). In order for the resulting RNA fragments to be sequenced, common adapter sequences have to be ligated to the 5' and 3' end of the fragments. This include the forward and reverse primers for the sequencing as well as primer sequences used in the reverse transcription (RT). Again, to minimize RNA loss, most of the recent protocols perform this step on beads, compared to the initial ligation steps on purified RNA fragments. Depending on the exact protocol slightly different ligation strategies are used (Ule et al., 2005; Lee and Ule, 2018). A key quality control step is the visualization of the purified complexes via SDS-PAGE. This step is crucial to optimize any of the above steps, like the RNA fragmentation. Usually a high and low RNase condition are used for visualization. In the high RNase concentration, the specificity of the purified complex is visualized, whereas in the low RNase concentration RNAs are purified for further cDNA library preparation (Ule et al., 2003). Next, complexes are cut from the gel and prepared for reverse transcription (RT) into cDNAs. Proteinase K is used to digest the RBP, which leaves a short peptide at the crosslink site on the RNA. Depending on the protocol different RT enzymes and conditions are used. These have different impacts on the amount of cDNAs that truncate at the crosslink site, compared to those that are entirely revers transcribed (readthrough reads). In order to track PCR amplification duplicates, some protocols introduced random unique molecular identifier sequences (UMIs) to the barcode sequences that label each fragment (König et al., 2010). Free adapter sequences as well as RT primers are usually removed by gel purification step. These sequences could otherwise become templates that create contaminated cDNA libraries. Libraries dominated by PCR artifacts could obscure the CLIP signal and lead to read loss. While classical gel-based extraction provides the most precise method to select RNA fragments of specific sizes, it is also prone to read loss and inefficient for many replicates. To increase the convenience and speed some protocols use other size separation methods for example based on silica-like beads (Zarnegar et al., 2016). Lastly, the resulting library is forwarded to high-throughput sequencing. Here, usually single-end sequencing is sufficient if the cDNA inserts remained in original orientation. If this is the case, the read start will contain the barcode information which is needed for downstream computational analysis (Lee and Ule, 2018).

### 1.3.2 The iCLIP protocol

It has been observed that in standard CLIP conditions more than 80% of cDNA truncate at the peptide residue that is left from initial protein crosslinking. Individual-nucleotide resolution CLIP (iCLIP) was developed to exploit exactly that, thereby achieving single nucleotide resolution. As for other CLIP-based protocols, iCLIP starts with UV crosslinking to form covalent bonds between the RBP and the RNA. In particular cells are irradiated with 150 mJ/cm$^2$ at 254 nm in standard iCLIP. This is followed by cell lysis and partial RNA digestion to obtain RNA fragments. Optimizing the conditions for the RNase digestion is a critical step. For optimal read-out, it is desired to achieve a target size of purified RNA molecules between 50 and 300 nucleotides (Huppertz et al., 2014). Next are the immunoprecipitation and adapter ligation for the reverse transcription steps. Here iCLIP incorporates a special RT primer design. Two cleavable adapter regions, together with a defined barcode are ligated to the 3' end of the fragments. The barcode region contains two sequence parts. The first, sample-specific unique sequence tag allows for parallel sequencing, since it can be used for later de-multiplexing. The second barcode part represents a UMI that can be used to removed duplicates that might arise from the PCR amplification step (Huppertz et al., 2014; König et al., 2010). In iCLIP it is desired that the reverse transcription stops and the cDNA molecule truncates directly at the crosslinking site. After gel-based size selection those truncated cDNA molecules are re-captured via intramolecular circularization. This attaches the second adapter to the 3' end of the cDNAs. Before sequencing the construct is linearized again by the endonuclease restriction enzyme BamHI (Huppertz et al., 2014; König et al., 2012).

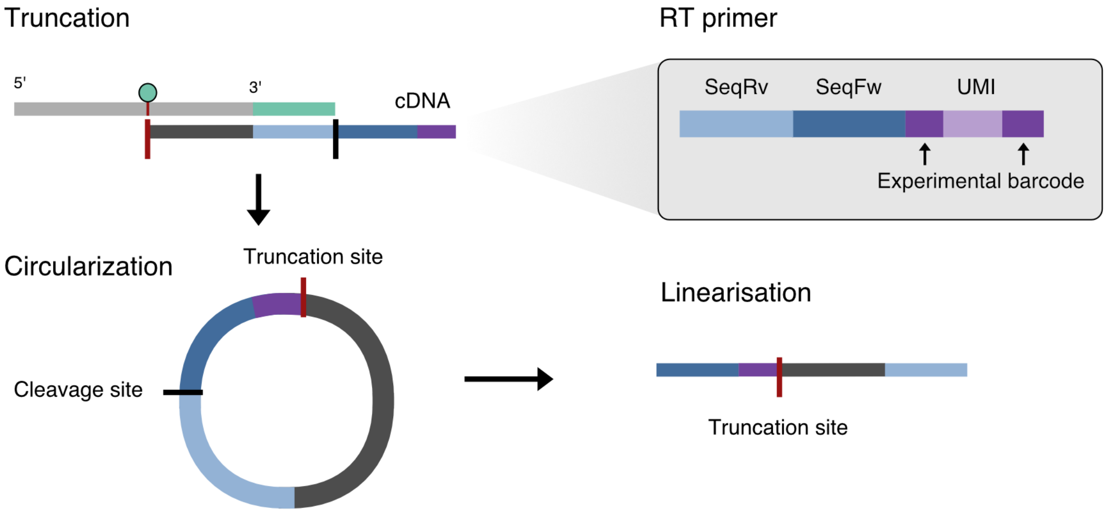

**Figure 1.7: Details of the adapter ligation and reverse transcription in iCLIP.** The iCLIP protocol achieves single nucleotide resolution by capturing the exact position of the truncation during reverse transcription. The protein particle covalently bound to the RNA causes the reverse transcriptase to truncate during cDNA generation. Truncated cDNAs are recaptured by a special RT primer design. This primer contains both, the reverse and the forward RT primer, with some additional barcode sequences. A circularization step is used to fuse the forward primer to the truncation site. This is followed by a linearization step and results in the final cDNA. Figure adapted from (König et al., 2012).

Of note, a recent update of the iCLIP protocol (iCLIP2) describes different optimization steps to improve the quality and complexity of the resulting libraries. In particular, the circularization step was replaced by two separate linker ligation steps. Further improvements are a PCR pre-amplification step as well as bead-based size selection for the cDNAs. In order to allow for larger library sizes also the composition of the barcode sequences was extended (Buchbender et al., 2020). In summary the iCLIP protocol allows to specifically identify protein-RNA interactions on a single-nucleotide level. With its transcriptome-wide and high-throughput scale detected interactions can be reinforced by making use of biological replicates. This ability to produce highly reproducible data makes it the perfect tool to characterize the binding of RBPs also in non-standard cell types, such as beta-cells.

## 1.4 The computational analysis of iCLIP data

### 1.4.1 Initial processing of the sequence data

The analysis of CLIP-Seq-derived data is usually comprised of three main steps. At first reads have to be mapped to a reference genome, followed by peak calling and postprocessing to gain functional insights. While in general being the same, most of these steps differ with regards to the specific CLIP protocol that was used. Depending on the protocol, different diagnostic events have to be used to identify the originally crosslinked position from the sequenced reads. HITS-CLIP and PAR-CLIP are based on readthrough reads, thus the whole read has to be used as read-out (Licatalosi et al., 2008). PAR-CLIP-derived reads for example carry a thymidine-to-cytidine transition directly at the crosslink site. This is because during cDNA generation the 4SU is translated into a guanine instead of an adenine (Hafner et al., 2010). Such an event can be exploited during the analysis to indirectly infer the exact position of the original crosslink event. In iCLIP however, the crosslink position is directly visible since about 80% of the cDNAs truncate directly at the crosslink site. Therefore, the read start itself can be used as indicative event to define the crosslink position (Sugimoto et al., 2012).

Common to all approaches is the initial quality control of the sequenced reads, which follows most next-generation sequencing pipelines. This includes the control of the sequencing performance by using per base read quality scores (Phred scores). Also, possible adapter sequences are removed from the read ends. Next reads can be quantified by aligning them to a selected reference genome (Chakrabarti et al., 2018). The choice of the alignment software for the initial read alignment however, depends on the RBP and the specific question. Reads can be aligned to the transcriptome, allowing for a potential transcript level resolution, while missing out on any binding to pre-mRNAs. Thus, alignment to the genome is preferred in cases without prior knowledge about the RBP. For the same reason, splice-aware aligners are preferred over non-splice-aware aligners (Baruzzo et al., 2017). After mapping PCR amplification biases can be removed in the case of iCLIP-derived reads. Reads mapping to the same location on the genome and sharing an identical UMI sequence are most likely derived by PCR duplication, rather than representing two independent crosslink events (Uhl et al., 2017).

### 1.4.2 Peak calling

RBP binding is a highly dynamic process with a wide range of affinities and kinetics between the RBP and its target RNA (Jankowsky and Harris, 2015). For that reason, no absolute threshold can be used to distinguish low-affinity sites from high-affinity sites. Furthermore, not every mapped read corresponds to a meaningful biological binding event. Peak calling analysis are used in order to differentiate between RBP binding and unspecific events (De and Gorospe, 2017). Peak calling usually consists of an initial round of peak identification, followed by stringent filtering steps. Similar to the wide range of CLIP-based protocols also a wide range of peak calling tools exists, each of which either specifically developed or adapted from already existing tools. Since all tools use different approaches for their specific use case, global comparisons and benchmarking are challenging (Bottini et al., 2017).

For instance, PARalyzer and PIPE-CLIP were specifically developed for mutation-based CLIP datasets, such as PAR-CLIP. The idea is to distinguish diagnostic mutations like the characteristic T-to-C transitions from sequencing errors, somatic mutations or single nucleotide polymorphisms. This is achieved by direct thresholding in the case of PARalyzer and slightly more complex by distributional modeling in the case of PIPE-CLIP (Corcoran et al., 2011; Chen et al., 2014). Truncation-based peak calling tools on the other hand use the nucleotide upstream of the mapped read as diagnostic event. Most tools model these crosslinks as variations of the negative binomial distribution. This distribution accounts for the over-dispersed nature of count data, where in general most counts are distributed over few positions. As a first tool ASPeak adapted this modeling approach (Kucukural et al., 2013). Next, it was shown that the zero-truncated variation of the negative binomial distribution is in general a better fit. Tools like Piranha and PIPE-CLIP made use of it (Uren et al., 2012; Chen et al., 2014). However, both tools depend on rather large bins which obscure the original single nucleotide resolution. In contrary to these distribution-based approaches, also permutation-based approaches were developed. In iCount for example, counts are randomly distributed over user-defined regions of interest taken from third-party annotation sources. These counts serve as background to which the observed count frequencies are compared in order to compute false discovery rates (König et al., 2010). CLIPer is the tool preferred by the ENCODE consortium (Van Nostrand et al., 2020). It consists of a two-step approach that uses distribution and permutation-based approaches. The first step yields a false discovery rate and in a second step a

Poisson distribution is used to remove peaks with lower counts than expected by chance (Lovci et al., 2013). It is worth to mention that an important requirement for truncation-based peak calling tools is to make use of the read start as diagnostic event. Using the whole reads would lead to misalignment and a loss of resolution (Van Nostrand et al., 2020). An inherent problem to all of the above tools is the resulting peak width. Ideally, the width of a peak would capture the binding footprint of the RBP. However, all of these tools are either based on internal binning or require a manually set window size for peak merging. This results in a wide spread of peak widths, in between and within distinct methods, which reduces the biological validity of the results and complicates downstream analysis. Additionally, the true binding affinity of an RBP might be obscured by the composition of the binding motif. It has been shown that uridine-rich motifs have the overall highest crosslinking efficiency. This uridine bias causes disproportional crosslinking to uridines. It is best noticeable in uridine-rich motifs that are shared between CLIP libraries of different RBPs (Haberman et al., 2017; Sugimoto et al., 2012).

A conceptually different peak calling algorithm, specifically developed for iCLIP and eCLIP data, comes in the form of PureCLIP (Krakau et al., 2017). It implements a non-homogeneous hidden Markov model (HMM). This allows not only to call peaks with single nucleotide precision, but also allows the integration of certain crosslink motifs to adjust for the uridine bias. The algorithm splits the initial information from mapped reads into read start sites and in a fragment density. The read starts relate to the single nucleotide crosslink events, whereas the fragment density serves as a measure of enrichment within the current region. The described model consists of four hidden states, which are based on the two input types. A position can be either categorized as crosslinked or non-crosslinked, as well as enriched or non-enriched. All positions of the state crosslinked and enriched are reported as peaks. To correct for high crosslinking biases, uridine-rich motifs can directly be incorporated into the model as covariates. This influences the emission probabilities of the two crosslink-based states. Further control experiments can also be included into the framework. Here the emission probability distribution of the enriched states is modulated by the additional covariates. In general, PureCLIP outperforms other peak calling methods on high-throughput iCLIP datasets. Its ability to correct for known biases allows to reliably detect peaks, leading to an increased precision in the definition of an RBP binding landscape.

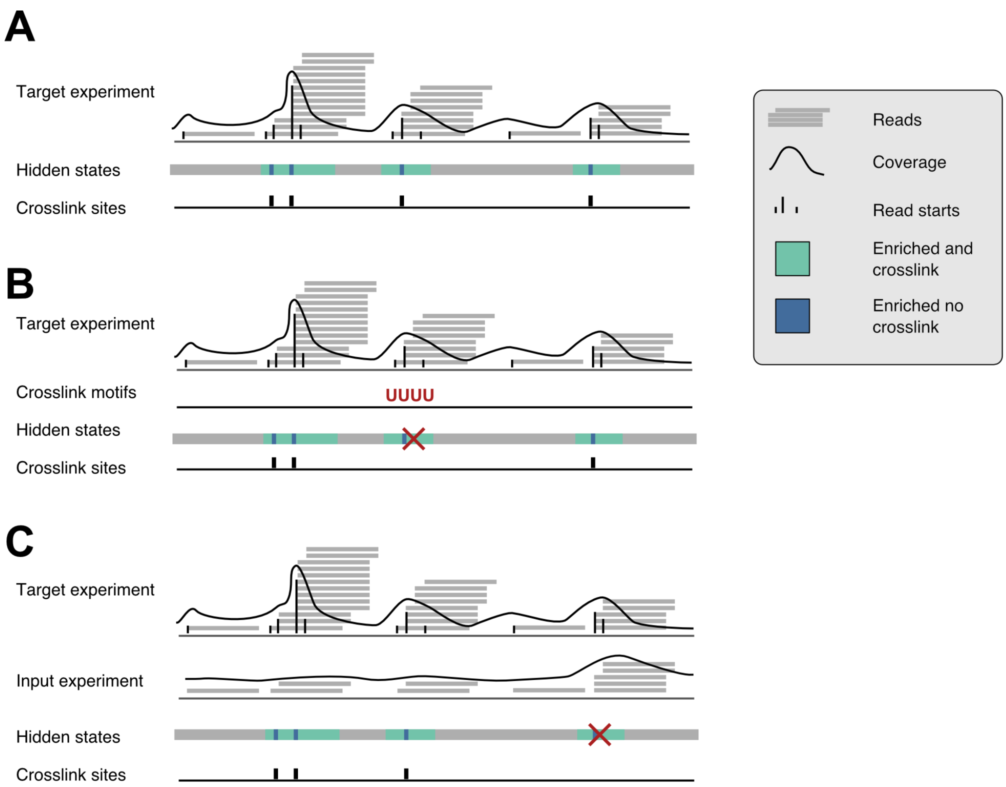

**Figure 1.8: Overview of the PureCLIP peak calling approach.** (A) Mapped reads from the iCLIP/ eCLIP experiments are used to define two input signals, the individual read start sites and a smoothed coverage of fragment densities. These signals are used to assign each nucleotide position to its most likely hidden state. Positions with the enriched and crosslinked state are returned as crosslink sites. PureCLIP allows the incorporation of crosslink motifs to adjust for known biases (B), as well as adjusting for experimental biases by incorporation of a sequenced input experiment. Figure adapted from (Krakau et al., 2017).

### 1.4.3 Post-processing

While the above steps follow a precisely defined question, various approaches and ideas are summarized under the term ’post-processing’. Usually all of these ideas have the extraction of biological insight in common. This can be based on the single nucleotide coverage derived by the initial data processing, or based on binding sites defined via peak calling. Most commonly the single nucleotide coverage is integrated with additional orthogonal data to generate a form of RNA-map (Ule et al., 2006). These maps are conceptually simple, yet very powerful. Usually the distribution of crosslink events is shown around regulated landmarks in the transcript. For example, regulated genes or exons can be identified by RNA-Seq and the crosslink distribution can be overlaid. Such approaches are used regularly and are also easily accessible for example directly by webservers (Rot et al., 2017; Park et al., 2016). Post-processing based on called peaks however is less clear. For example, no clear pipeline exits to adequately account for reproducibility between replicates. Some ideas exit, suggesting either replicate merging before the actual peak calling step, or merging final binding sites from replicates (Li et al., 2011; Chakrabarti et al., 2018). In both cases a clear trade-off between specificity and sensitivity is made, but the how and when to use either approach is not discussed so far. Also, methods to integrate binding sites with known gene and transcript annotations are not described. However, this is a common task since the usually general binding spectrum of an RBP is of high interest. In summary this suggest that the computational processing of iCLIP data covers most of the essential steps. Especially for peak calling a wide range of options exist. Sadly, no standardized workflows exist to derive easily reproducible biological insights from the processed data. This suggest that additional efforts have to be made in that direction.

## 1.5 Aim of this thesis

While it is common that computational analysis approaches for iCLIP datasets follow the three basic steps of mapping, peak calling and post-processing, a standardized workflow is still missing. Most workflows and pipelines are either custom solutions, or represent a patchwork of outdated tools applied to modern experimental protocols. Thus, iCLIP-based computational analysis lacks comparability between RBPs or research groups.

For this reason, the first part of this thesis sets out to define a standardized processing pipeline equal to other next-generation sequencing-based approaches. I describe a complete analysis workflow that allows for the reliable detection of RBP binding sites from iCLIP data. All steps from initial quality control via peak calling to binding site post-processing are covered. In contrary to available pipelines, a particular focus is held on the post-processing step. Here I give exact insights into the transcriptome-wide profiling of RBP footprints. A special effort is made to describe how peak calling can be managed with multiple replicates, while accounting for reproducibility. Additionally, the accurate integration with gene annotation data is described to streamline down stream analysis processes.

Next, I apply the pipeline defined in the first step to an iCLIP dataset of the splicing factor SRSF6 in pancreatic beta-cells. An in-depth profile of the SRSF6 binding spectrum in the context of beta-cells is computed. These findings are further integrated with SRSF6-mediated alternative splicing changes based on the analysis of RNA-Seq data. The combination of both datasets unveils how SRSF6 might recognize exons to alter their fate. These findings help to understand AS-mediated transcriptome changes in diabetes.

# 2 | Methods

## 2.1 Establishment of the iCLIP processing workflow

The described workflow resulted from a collaboration with our partners Dr. Anke Busch and Dr. Stefanie Ebersberger from the IMB in Mainz. In section 2.1.1 I summarized processing steps performed by Dr. Anke Busch. In section 2.1.2 the PureCLIP-based peak calling was performed by Dr. Stefanie Ebersberger, whereas the downstream binding site merging was performed by myself. All described processing steps are based on a published iCLIP dataset of U2AF65 (Zarnack et al., 2013).

### 2.1.1 Initial iCLIP data processing

At first, reads were checked for their sequencing quality using FastQC. Thereby a minimum Phred score of 10 was required for all nucleotides of the barcode region of the reads using the FastX-Toolkit (`http://hannonlab.cshl.edu/fastx_toolkit/`). Adapter demultiplexing and barcode trimming was performed with Flexbar (Dodt et al., 2012). A minimum overlap of 1 nt was required between read and adapter sequence for the trimming and reads were split into replicates by their barcode sequence. The following settings were applied:

```
Flexbar -r <data.filtered.fastq.gz>
        --zip-output GZ
        --barcode barcodes.fasta
        --barcode-unassigned
        --barcode-trim-end LTAIL
        --barcode-error-rate 0
        --adapter-seq adapter.seq
        --adapter-trim-end RIGHT
        --adapter-error-rate  0.1
        --adapter-min-overlap   1
        --min-read-length minReadLength
        --umi-tags
```

Processed reads were mapped to the human reference genome (GRCh38.p7) using STAR (Dobin et al., 2013). Soft-clipping was turned off to retain the crosslink position and up to two mismatches were allowed in the alignment. The detailed program call was as follows:

```
STAR --runMode alignReads
        --genomeDir genomeMappingIndex
        --outFilterMismatchNoverReadLmax 0.04
        --outFilterMismatchNmax
        --outFilterMultimapNmax 1
        --alignEndsType Extend5pOfRead1
        --sjdbGTFfile annotation.gtf
        --sjdbOverhang maxReadLength-1
        --outReadsUnmapped Fastx
        --outSJfilterReads Unique
        --readFilesCommand zcat
        --outSAMtype BAM SortedByCoordinate
        --readFilesIn <sampleX.fastq.gz>
```

Reads mapping to the same position in the genome while sharing an identical random barcode sequence are likely to be PCR duplicates. These technical artifacts were removed using UMI-tools (Smith et al., 2017). The detailed program call was as follows:

```
umi_tools_dedup -I <sampleX.bam>
        -L <sampleX.duprm.log>
        -S <sampleX.duprm.bam>
        --extract-umi-method read_id
        --method unique
```

Lastly, reads were reduced to their starting position and shifted by one nucleotide to transform mapped reads into crosslink events. For this step the BEDTools suite was used (Quinlan and Hall, 2010).

### 2.1.2 Binding site identification

PureCLIP was used as peak calling software to identify enriched peak regions (Krakau et al., 2017). The tool takes the mapped and de-duplicated reads as input and computes a list of crosslink sites. We used the tool with default parameters, besides using the $-ld$ flag for enhanced precision. The detailed program call was as follows:

```
pureclip -i <merged.bam>
        -bai <merged.bam.bai>
        -g <genome.fasta>
        -ld
        -nt 8
        -o <PureCLIP.crosslink_sites.bed>
        -or <PureCLIP.crosslink_regions.bed>
```

PureCLIP crosslink sites are of single nucleotide width and can be combined into binding sites by merging adjacent positions. This can be done by merging sites that are closer to each other than the desired binding site width -1 nt. To achieve 5-nt binding sites for example, all crosslink sites closer than 4-nt would be merged. This ensures that the final binding sites are not overlapping each other. To exclude spurious crosslink sites merged regions shorter than a specific threshold should be removed. For example, merged sites that remained single nucleotide width are probably caused by mapping artifacts since no other position in close proximity was detected. I developed an in-house merge and splitting routine for this purpose (see figure 3.4). The detailed scripts are shown in the supplementary material section (supplementary code 5). In the case of U2AF65 9-nt wide binding sites were computed. This means all crosslink sites closer than 8-nt were merged. Resulting merged regions were filtered to remove all regions shorter than 3-nt. To compute equally sized binding sites merged regions shorter than 9-nt were extended and regions longer than 9-nt were split up. Splitting was done iteratively always selecting the position with the largest crosslink signal as the binding site center. Each merged binding site was also required to harbor at least two of the original crosslink sites. The detailed function call for the in-house script was as follows:

```
peaksToBindingSites(
        peaks = peaksFiltered,
        peaksSize = 9,
        peaksSizeRemove = 2,
        minXlinksPerPeak = 2,
        minPureClipSites = 2,
        clipSignalPlus = clipSignalPlus,
        clipSignalMinus = clipSignalMinus
)
```

### 2.1.3 Binding site reproducibility and downstream processing

We investigated if a binding site is reproducible in each of the individual replicates. The number of crosslinks, that directly overlapped the range of the binding sites was counted for all replicates. This resulted in a crosslink event distribution per replicate. As filtering threshold, we selected the 10% quantile of each distribution, with a lower boundary of two crosslink events. This resulted in the following thresholds: Replicate 1 → 2 crosslinks, replicate 2 → 3 crosslinks, replicate 3 → 4 crosslinks and replicate 4 → 2 crosslinks. All binding sites where the threshold was met by at least three replicates were deemed reproducible.

To annotate direct gene targets, we overlapped the range of annotated genes and transcripts with those of the reproducible binding sites. Here a partial overlap was deemed sufficient. All binding sites overlapping multiple different gene annotation ranges were discarded. To assess the overlap with transcripts, binding site ranges were overlapped with those of annotated transcripts. Again, a partial overlap was deemed sufficient. We considered the following transcript regions for our annotation: Introns, coding sequence, 3'UTR and 5'UTR. Binding sites overlapping multiple different transcript type regions were resolved by applying the following scheme: (1) Majority vote; Assign the binding site to that transcript region that is the most overlapping. (2) Hierarchical rule; In case of ties, prefer intron > 3'UTR > CDS > 5'UTR. Binding site numbers after each processing steps are given in table 3.3.

## 2.2 Definition of the SRSF6 binding spectrum

We analyzed four biological iCLIP replicates of SRSF6 in the human pancreatic cell line EndoC-$\beta$H1. The library preparation and wet-lab work was performed by Ines Alvelos of the Eizirik group. Details about the exact steps can be found in section 2.4. The initial processing of the iCLIP data described in section 2.2.1 was performed by Dr. Anke Busch using the pipeline described in section 2.1.1. Here, I provide a detailed overview of the different steps that were performed to characterize the SRSF6 binding behavior.

### 2.2.1 iCLIP data processing

The initial sequencing yielded a total of 175,386,785 individual reads which were monitored for data quality using FastQC. Sequencing reads were filtered by the Phred score of the barcode region. Only reads with a sequencing quality of above 20 in the sample barcode, or with a minimum of one position below 17 in the random barcode were retained for further analysis. Next, Flexbar was used to demultiplex reads based on the sample barcode (Dodt et al., 2012). Potentially remaining adapter sequences were trimmed off from the right end of the reads only, again using Flexbar. Thereby a minimum overlap of 1 nt was required, while allowing up to one mismatch in 10 nt in the alignment. In a next step, the 9 nt barcode regions were completely removed from the read sequence, but stored as meta information in the read name. Reads shorter than 15 nt were removed and all remaining reads were mapped to the human genome using STAR (Dobin et al., 2013). With soft-clipping turned off and allowing for up to two mismatches, an average mappability of 78.4% was achieved and only uniquely mapped reads were retained. Reads with identical random barcode sequences mapped to the exact same position were removed by the dedup function of the bamUtils tool suit, using the random barcode sequence that was attached to the read name. Finally, reads were reduced to their starting position and shifted by one nucleotide to capture the truncation site and generate a crosslink coverage. A detailed overview of all read numbers after each processing step is given in table 3.5.

### 2.2.2 Peak calling and binding site definition

As a first step, the crosslink events from all four replicates were merged into a single file for peak calling with PureCLIP. Resulting crosslink sites were filtered by their PureCLIP-score and the lowest 5% scoring sites were removed. PureCLIP was used with default parameters, besides using the $-ld$ flag for enhanced precision. The detailed program call was as follows:

```
pureclip -i <merged.bam>
        -bai <merged.bam.bai>
        -g <genome.fasta>
        -ld
        -nt 8
        -o <PureCLIP.crosslink_sites.bed>
        -or <PureCLIP.crosslink_regions.bed>
```

All remaining sites were merged into 9-nt wide non-overlapping binding sites using an in-house R script (for details, see 5). These were filtered to contain at least two positions that were initially called by PureCLIP. The detailed function call for the in-house script was as follows:

```
peaksToBindingSites(
        peaks = peaksFiltered,
        peaksSize = 9,
        peaksSizeRemove = 2,
        minXlinksPerPeak = 2,
        minPureClipSites = 2,
        clipSignalPlus = clipSignalPlus,
        clipSignalMinus = clipSignalMinus
)
```

These sites were overlaid with gene annotations. Binding sites overlapping multiple different genes were excluded (3.1%). For protein-coding genes a majority rule was applied to assign binding sites to a specific transcript region, followed by a hierarchical rule to handle ties (intron > CDS > 3UTR > 5UTR). In-house scripts for the binding site definition can be found in the supplementary material section 5.

### 2.2.3 Description of the SRSF6 binding motif

The definition of the binding motif was based on 5,000 randomly sampled binding sites. Triplet and pentamer frequencies were counted using the biostrings R package allowing for overlaps (Pagès et al., 2017). For the counting three windows were spanned around each binding site, the 9-nt center, the 20-nt upstream as well as the 20-nt downstream of the binding site. Pentamer frequencies were computed for all four transcript regions, such as introns, CDS, 3'UTR and 5'UTR. The top 300 pentamers of the CDS region were chosen and analyzed in positional heatmaps. These were clustered by k-means clustering with three centroids using the complexHeatmaps R package (Gu et al., 2016). The de novo motif search was performed using DREME (Bailey, 2011) over all binding sites. 201-nt long sequences centered at the binding site were used as input. The position weight matrix of the second best hit (the GAA-rich motif) was taken as input for FIMO (Grant et al., 2011) to search against all input sequences. In-house scripts based on which these analyses were performed can be found in the supplementary material section 5.

## 2.3 RNA-Seq data analysis

For this analysis an RNA-Seq dataset previously published by our colleagues of the Eizirik group was re-analyzed (available at GEO under the accession GSE98485) (Juan-Mateu et al., 2018). The group performed a total RNA-Seq experiment in EndoC-$\beta$H1 cells with five replicates exposed to control (siCTL) and five to SRSF6 KD (siSRSF6). In the following I will give a detailed description of the methods used for its analysis.

### 2.3.1 Analysis of alternative splicing changes

Initial read quality was monitored using FastQC. Adapter sequences were removed from all 3' ends using Flexbar (Dodt et al., 2012). Resulting reads were trimmed to 98-nt in order to achieve a uniform read length. The exact program call was as follows:

```
for i in `less $ <files>`; do
        echo "---Processing Sample $i"
        s1=$<dir>/$i*R1*;
        s2=$<dir>/$i*R2*;
        flexbar --reads $s1
            --reads2 $s2
            --target $outdir/$i
            --threads $threads
            --post-trim-length 98
            --min-read-length 98
done
```

Genomic mapping was performed with STAR without allowing soft-clipping to maintain the defined read length (Dobin et al., 2013). The exact program call was as follows:

```
for i in `less $ <files>`; do
        echo "---Processing Sample $i"
        s1=$<dir>/$i*R1*;
        s2=$<dir>/$i*R2*;
        STAR --runMode alignReads
            --runThreadN $threads
            --genomeDir $index
            --readFilesIn $s1 $s2
            --outFilterMismatchNmax 2
            --outFilterMultimapNmax 1
            --outSAMtype BAM SortedByCoordinate
            --outFileNamePrefix $outdir/$i
            --alignEndsType EndToEnd
```

```
done
```

Alternative splicing changes were analyzed using rMATS-turbo specifying read type, read length and strand specificity (Shen et al., 2014). The exact program call was as follows:

```
rmats --b1 <replicates_condition_1>
        --b2 <replicates_condition_2>
        --gtf <annotations.gtf>
        --od <output_dir>
        --t paired
        --readLength 98
        --nthread 4
        --tstat 4
        --cstat 0.0001
        --libType fr
```

We determined the proportion of reported AS event types and focused the further analysis on cassette exon (CE) events only. Initial CE events were filtered for overlaps. If two or more CE events overlapped each other only the one with the lowest FDR was retained. We considered an alternative splice event to be significant if it passed a false discovery rate (FDR) $< 0.05$. Potential hits were also filtered by their absolute splicing change ($|\Delta PSI| > 0.05$). CE events were further filtered by their $log_2$-transformed sum of junction-spanning reads supporting the event $> 5$ (mean between replicates). To integrate SRSF6-regulated cassette exons with SRSF6 binding, we overlapped defined binding sites with the entire alternatively spliced region. This region was defined as the window from -100 nt of the 5' splice site of the upstream constitutive exon until +100 nt of the 3' splice site of the downstream constitutive exon. All scripts used to compute the described results are given in the supplementary materials section 5.

### 2.3.2 Differential expression analysis

Initial sequencing quality of the reads was monitored with FastQC. Adapter sequences were removed from 3' ends using Felxbar (Dodt et al., 2012). Reads were filtered for sequencing quality adapting window-based quality trimming again using Flexbar. The exact program call was as follows:

```
for i in 'less $ <files>'; do
        echo "---Processing Sample $i"
        s1=$<dir>/$i*R1*;
        s2=$<dir>/$i*R2*;
        flexbar --reads $s1
```

```
            --reads2 $s2
            --target $outdir/$i\\
            --threads $threads
            --qtrim WIN
            --qtrim-format i1.8
            --qtrim-threshold 20
            --qtrim-win-size 5
            --adapters $adapter
            --adapter-trim-end RIGHT
            --adapter-min-overlap 3
            --adapter-error-rate 0.1
done
```

Next reads were mapped with STAR to the human genome (Dobin et al., 2013). Soft-clipping was enabled, but a maximum of two mismatches were allowed and only uniquely-mapped reads were retained. The exact program call was as follows:

```
for i in `less $ <files>`; do
        echo "---Processing Sample $i"
        s1=$<dir>/$i*R1*;
        s2=$<dir>/$i*R2*;
        STAR --runMode alignReads
            --runThreadN $threads
            --genomeDir $index
            --readFilesIn $s1 $s2
            --outFilterMismatchNmax 2
            --outFilterMultimapNmax 1
            --outSAMtype BAM SortedByCoordinate
            --outFileNamePrefix $outdir/$i
            --alignEndsType EndToEnd
done
```

Expression levels were quantified by counting reads within annotated exons using genomicAlignments (Lawrence et al., 2013). The exact function call was as follows:

```
genomicAlignments::summarizeOverlaps(
        features = ebg,
        reads = <reads.bam>,
        mode = "union",
        singleEnd = FALSE,
        ignore.strand=TRUE
)
```

The resulting count matrix was filtered for genes covered by at least 10 reads over all replicates. Differential testing was performed using DESeq2

(Love et al., 2014). We accounted for the paired nature of the dataset by adding this information into the design formula. The exact function call was as follows:

```
DESeq2::DESeqDataSet(
        <summarizedObject>,
        design = ˜ samplePair + genotype
)
```

Resulting $P$ values were corrected for multiple testing using the Benjamini-Hochberg correction and genes were considered significant if they passed an adjusted P value threshold of $< 0.001$. We additionally filtered genes based on their absolute $log_2$-transformed fold-change ($|LFC| > 1$). Splicing regulators were identified by their association with the Gene Ontology term "RNA splicing" (GO:0008380).

### 2.3.3 RNA splicing maps

We grouped alternatively spliced exons based on their $\Delta PSI$ values into decreased or increased inclusion sets. For each regulated cassette exon, we defined a regulated region by including the flanking up- and downstream exons. The respective 5' and 3' splice sites of these exons serve as anchor point to span a symmetric window of 200-nt. This results in four distinct windows, each of which centered at a splice site. Intronic window parts are always 100-nt wide, whereas exon regions were defined based on the available exon length to avoid overlaps. For each position all exons with at least one crosslink event at a distinct position were counted and divided by the total number of exons. This results in a relative crosslink frequency. Frequencies on the PSI-matched background set were determined in the same way, but repeated 50 times to calculate a mean and standard deviation for each nucleotide position. These mean and standard deviation values were compared to the observed value of the respective increased or decreased inclusion exon, to compute a positional Z-score as well as a $P$ value for the identification of significant positions. Scripts for the computed maps can be found in the supplementary materials section 5.

## 2.4 Biological methods

For completeness biological validations as well as a description of the initial iCLIP are included in this thesis. All wet-lab experiments were performed by Ines Alvelos. Here I give a short summary of the methods that were used.

### 2.4.1 iCLIP library preparation

The SRSF6 iCLIP libraries were prepared using a previously described protocol (Haberman et al., 2017; Sutandy et al., 2016). Initially, crosslink conditions were optimized in HeLa cells and used as positive control. The final iCLIP libraries were prepared from EndoC-$\beta$H1 cells in four biological replicates. The cells were irradiated using 254 nm UV with 300 mJ/cm$^2$ to induce crosslinking between SRSF6 and the RNAs (König et al., 2010). Partial RNase digestion was performed by adding 2 U of RNase I (Ambion) to the sample lysates and immunoprecipitation was performed using a specific anti-SRSF6 antibody (Anti-SRSF6/SRP55 [aa250-300] LS-C290327, LifeSpan Bioscience, The Netherlands). For sequencing, the prepared libraries were run on an Illumina HiSeq 2500 sequencing system, sequenced as 75-nt single-end reads. Sample barcodes used in this experiment are given in table 3.5.

### 2.4.2 mRNA extraction, quantitative PCR and RT-PCR

Poly(A)+ mRNAs were isolated using the Dynabeads mRNA DIRECT Kit (Invitrogen, Carlsbad, CA). Reverse transcription was done using the the Reverse Transcriptase Core kit (Eurogentec, Belgium), after recovering mRNA molecules in Tris-HCl elution solution. Quantitative PCR amplification was performed using IQ SYBR Green Supermix (Bio Rad, Hercules, CA, USA) and Rotor-Gene Q (Qiagen, Venlo, Netherlands). The standard curve method (Overbergh et al., 1999) was used to calculate the PCR product concentration as copies per l, while correcting for gene expression based on beta-actin levels. A list of primers is given in table 2.1. For the validation of observed alternative splicing events by RT-PCR designed primers were annealed to the flanking constitutive exons and RT-PCR was performed using the Red-Taq DNA polymerase (Bioline, UK). PCR products were analyzed using the LabChip electrophoretic Agilent 2100 Bioanalyzer system and the DNA 1000 LabChip kit (Agilent Technologies, Diegem, Belgium). Quantification of the PCR bands corresponding to a specific splice variant was done using the 2100

Expert Software (Agilent Technologies, Diegem, Belgium), which was also used to calculate percent inclusion ratios.

**Table 2.1: Sequences of primers used for splicing analyses and quantitative RT-PCR** Abbreviations used are as follows: SPL, primers used to analyze splicing variants; qRT, primers used for quantitative RT-PCR.

| Gene | Application | Forward (5'-3') | Reverse (5'-3') |
|---|---|---|---|
| LMO7 | SPL | GGATAACAGAAGAAGTTGGGC | CCATTTTGCAAGGTCATCCTGC |
| RBM6 | SPL | GGTACCTGAAGATGCCACAAAAG | CCACCAATGTTTGCCTTACATCG |
| ITGB3BP | SPL | GCCTGTTAAAAGATCACTGAAG | CTACTGCCCTCCAAAGCCTGTAT |
| STARD10 | SPL | CCCTGAAGAACCGTGATGTC | CTTCTTCATGGCCTTGGGAGC |
| CDK2 | SPL | CATCAAGAGCTATCTGTTCCAGC | GCATAGAAGTAACTCCTGGCC |
| CENPO | SPL | GGGAATTCTCGCTTCTGGCCTG | GTTCCAAGAGCACCTTCCTGGG |
| BCAR1 | SPL | CAAAGGTGGTGGTGCCCACC | CACGTCGTAGAGGTCAGGAGCC |
| ACTB | qRT | CTGTACGCCAACACAGTGCT | GCTCAGGAGGAGCAATGATC |
| SRSF6 | qRT | CATAGGACGCCTGAGCTACA | TGCCGTTCAGCTCGTAAAC |

## 2.5 Programs

In the following part an overview of all programs and software tools used in this thesis is given. Table 2.2 provides a summary about all programs, table 2.3 provides a summary of all R packages.

**Table 2.2: List of tools used in this thesis.** For each tool the version and the respective reference is provided.

| Program | Version | Reference |
|---|---|---|
| FastQC | 0.11.4 | *online* |
| Flexbar | 3.0.3 | (Dodt et al., 2012) |
| STAR | 2.5.2b | (Dobin et al., 2013) |
| Samtools | 1.5.0 | (Li et al., 2009) |
| IGV | 2.3.97 | (Robinson et al., 2011) |
| PureCLIP | 1.0.0 | (Krakau et al., 2017) |
| DREME | 5.1.1 | (Bailey, 2011) |
| FIMO | 5.2.0 | (Grant et al., 2011) |
| rMATS | 4.0.1 | (Shen et al., 2014) |

**FastQC** FastQC was used to monitor the quality of sequencing reads (`https://www.bioinformatics.babraham.ac.uk/projects/fastqc/`). It accepts unprocessed reads in terms of fastq files, as well as mapped reads in the form of BAM/SAM files. It provides an overview about the reads phred quality scores per position to visualize the sequencing quality. It further shows useful statistics about the GC content, read duplication rates and potential

**Table 2.3: List of R packages used in this thesis.** For each package the version and the respective reference is provided.

| Program | Version | Reference |
|---|---|---|
| GenomicAlignments | 1.24.0 | (Lawrence et al., 2013) |
| GenomicFeatures | 1.40.1 | (Lawrence et al., 2013) |
| rTracklayer | 1.48.0 | (Lawrence et al., 2009) |
| DESeq2 | 1.28.1 | (Love et al., 2014) |
| Biostrings | 2.56.0 | (Pagès et al., 2017) |
| clusterProfiler | 3.10.1 | (Yu et al., 2012) |
| ggplot2 | 3.30.2 | (Wickham, 2016) |
| complexHeatmaps | 2.40.3 | (Gu et al., 2016) |

adapter contaminations. FastQC was used for the quality control of iCLIP and RNA-Seq data.

**Flexbar** Flexbar is a preprocessing tool for any kind of high-throughput sequencing data (Dodt et al., 2012). It provides options to efficiently handle reads in the form fastq and fasta data files. It for example provides options to remove adapter sequences from read ends, filter reads for phread quality scores as well as demultiplexing of libraries based on barcode sequences. Flexbar was used in the processing of iCLIP and RNA-Seq data.

**STAR** STAR (Spliced Transcripts Alignment to a Reference) is a reference based read aligner, that is splice-aware (Dobin et al., 2013). It was specifically designed with the challenges of RNA-seq data in mind, outperforming most other alignment tools in terms of mapping speed. It allows for several filter options on the resulting read alignments. The number of mismatches per alignment as well as the number of multimapping reads can be controlled. It also reports detailed statistics about uniquely mapped reads, such as average length, number of splices sites, mismatch or deletion rate per base. It also allows to remove non-matching parts of the read in the alignment to enhance mappability with a feature called soft-clipping. STAR was used for the mapping iCLIP and RNA-Seq reads to the human genome.

**Samtools** Samtools provides a collection of methods to manipulate alignment files (Li et al., 2009). Its tools can be used with any kind of alignment software that produces BAM or SAM output files. Samtools was used for sorting and indexing of BAM

**IGV** The Integrative Genomics Viewer allows used to visualize genome wide data (Robinson et al., 2011). It takes a wide range of inputs, ranging from very condensed data formats such as BigWig files, to total alignments in terms of BAM files. It was used to visualize all various dataset, such as RNA-Seq alignments, iCLIP BigWig files and GTF gene annotations.

**PureCLIP** PureCLIP is a peak calling tool that allows the detection of protein-RNA interaction footprint (Krakau et al., 2017). It was specifically developed for protocols that yield single-nucleotide data, such as iCLIP or eCLIP. It directly takes aligned reads as BAM files as input and returns crosslinked positions associated with a significance score. Conceptually it trains a hidden Markov model based on read coverage and read start sites (for details see xxx). PureCLIP was used to call crosslink sites for the U2AF65 as well as the SRSF6 iCLIP datasets. We always used the merge of all replicates as input.

**DREME** DREME (Discriminative Regular Expression Motif Elicitation) is part of the MEME-Suite and was designed to find short enriched sequence motifs (Bailey, 2011). It takes direct sequences as input in the format of fasta files. The input can be a single such file or it can be two sequences sets one to search for enrichment and one background set as reference. In the case where no background set is explicitly provided such a background is generated internally by permutation. DREME was used to calculated the de-novo SRSF6 binding motif.

**FIMO** FIMO (Find Individual Motif Occurrences) is also part of the MEME-Suit and was designed to can sequences for motif matches (Grant et al., 2011). It takes the position-weight matrix from a motif, as well as the sequences to search in as fasta file as input. It returns a set of sequences that contain the input motif. FIMO was used to identify binding sites that show the de-novo SRSF6 binding motif in their surroundings.

**rMATS** rMATS is the successor of the Multivariate Analysis of Transcript Splicing (MATS) tool (Shen et al., 2014). It is able to handle replicate experiments for the quantification of alternative splicing events (replicate-MATS). It takes bam files of mapped reads as input and quantifies splicing events which are deduced from an annotation file that has to be provided as input as well. The quantification is based on splice junction-spanning reads; thus, it can handle only reads that were aligned with a splice-aware alignment software.

rMATS was used to quantify the AS changes in EndoC-BH1 cells upon SRSF6 KD.

**R packages and scripts** All computational scripts and statistical analysis of this thesis were written in R (version 4.0.2). This ranges from the basic visualization of for example mapped RNA-Seq reads, to complex orthogonal data integration processes when creating RNA-splicing maps. It was also used to implement all steps of the iCLIP binding site processing pipeline. These custom scripts were created with functions from base R, as well as some additional open source packages that are either hosted by CRAN or Bioconductor. Any genome encoded positional data, such as read and crosslink coverages, but also gene and transcript annotations were processed with GenomicAlignments and GenomicFeatures (Lawrence et al., 2013). These packages provide a wide range of functionalities to structure and manipulated data that is encoded on genomic coordinates. For input, export and conversion of such genomic coordinate encoded files the rTracklayer package was used (Lawrence et al., 2009). Differential expression analysis was performed with DESeq2 package, that was specifically developed for count-based RNA-Seq data (Love et al., 2014). It takes a matrix as input that represents read counts per sample over genes, as well as meta information about the input samples. Based on a provided design formula the tool trains a generalized linear model (GLM) to compute fold-changes and significance statistics for each gene. DESeq2 was used to calculate fold-changes between the RNA-Seq data in SRSF6 KD vs. control condition. K-mer based analysis, such as counting over various different ranges was done with Biostrings (version 2.56.0) (Pagès et al., 2017). This package provides utilities for fast and efficient string matching and counting of biological sequence sets. Biostrings was used for the pentamer and triplet-based profiling of the SRSF6 binding motif. To analyze the functional enrichment of certain gene sets clusterProfiler was used (Yu et al., 2012). It provides an interface for the statistical analysis of GO and KEGG profiles. It takes a test set of GeneIDs and a set of background GeneIDs as input and performs overrepresentation analysis of the associated terms. clusterProfiler was used to calculate GO profiles of genes directly targeted by SRSF6 binding and genes that were differentially expressed in the RNA-Seq between the SRSF6 KD and control conditions. All graphics, plots and visualizations were created using the ggplot2 package, with the exception to matrices and heatmap visualizations, which were created using the complexHeatmaps package (Wickham, 2016; Gu et al., 2016).

## 2.6 Databases

**PubMed NCBI** PubMed (`https://pubmed.ncbi.nlm.nih.gov/`) is a free database from the National Center for Biotechnology Information (NCBI) (Wheeler et al., 2007). The debase comprises of millions of biomedical records and offers free full-text versions articles. They also provide related tools to search and query their literature database. In this thesis PubMed was used for literature search and review.

**GENCODE** The GENCODE database (`https://www.gencodegenes.org/`) provides genome wide gene and transcript annotations curated with biological evidence (Frankish et al., 2019). They identify and classify gene features in human and mouse while providing a periodically updated release. For this thesis gene and transcript annotations of release 29 for the human genome version xxx was used. These annotations were further filtered for their gene and transcript level support. Only genes with a support $\leq 2$ and transcripts with as support $\leq 3$ were used.

**GWAS Catalog** The GWAS database (`https://www.ebi.ac.uk/gwas/home`) provides a catalog of human genome wide association studies (Buniello et al., 2019). It is provided by the National Human Genome Research Institute (NHGRI) and the European Bioinformatics Institutes (EBI). They provide a comprehensive resource that archive thousands of studies investigating the association of specific phenotypes with single-nucleotide polymorphisms. In this thesis the database was used to retrieve a list of T1D and T2D diabetes susceptibility genes.

**Immunobase** The Immunobase database (`https://genetics.opentargets.org/immunobase`) also provides a database for genetic association with certain phenotypes. They specifically focus on immune-mediated diseases. T1D and T2D diabetes susceptibility genes were download from this resource as well. The database was recently integrated into the GWAS catalogue and can now be accessed via OpenTartes (Ochoa et al., 2020).

# 3 | Results

## 3.1 Establishing an iCLIP processing workflow

Most cellular processes are regulated by RBPs with a wide range of affinities, for example to specific mRNA targets. RBPs use defined binding sites on the mRNA to recognize their targets. One state-of-the-art method to study these interactions on a transcriptome-wide scale is individual-nucleotide resolution UV crosslinking and immunoprecipitation (iCLIP) coupled with high-throughput sequencing. Similar to other high-throughput methods, like RNA-Seq for example, iCLIP yields millions of sequencing reads. To generate biological insight from the sequenced data, computational analysis usually covers three essential steps (see introduction 1.4). Initial read quality is checked and pre-processing of the iCLIP reads is done to retrieve single crosslink events. This is followed by peak calling and a variety of different post-processing steps. In this chapter I set out to give a detailed overview of these steps, specifically highlighting details of the post-processing part.

### 3.1.1 Initial data processing

In the presented workflow a published iCLIP dataset for the RNA-binding protein U2AF65 in HeLa cells was chosen to demonstrate the individual processing steps (Zarnack et al., 2013). The experiment consists of four biological replicates. Library preparation as well as the sequencing was performed by the König group (IMB Mainz). Further, the presented workflow was published in 2019 and represents the shared work and efforts of all authors (Busch et al., 2020). Initial sequencing data processing up to the point of obtaining a single-nucleotide resolution crosslink coverage was performed by Dr. Anke Busch (IMB Mainz core facility).

The sequencing yielded a total of 134,386,066 individual reads which were subsequently processed to obtain crosslink events. The term 'crosslink event' is used throughout this thesis to describe the detection of a protein-RNA contact, measured by the sequenced reads. Major processing steps of particular importance are the initial quality filtering specifically on the barcode part of the read (figure 3.1). This was followed by sample de-multiplexing, where reads got assigned to the individual replicate based on their experimental barcode. Replicates were next mapped to the genome and PCR duplicates were removed based on the random barcode sequence. A summary of the processing steps is given in supplementary table 3.1 and described in the methods section (2.1.1).

**Table 3.1: Read counts before and after de-duplication.** Read counts are given after mapping to the reference genome. Technical duplicates were removed for downstream analysis.

| Sample | Uniquely mapped reads | Crosslink events after de-duplication |
|---|---|---|
| Sample 1 | 8,027,421 | 6,666,142 (83.04%) |
| Sample 2 | 28,978,914 | 22,619,516 (78.06%) |
| Sample 3 | 40,053,868 | 30,510,305 (76.17%) |
| Sample 4 | 16,713,002 | 12,490,440 (74.73%) |

Following up on these results I started to tackle the question of how to best identify individual binding sites from the characteristic transcriptome-wide iCLIP coverage. The term 'binding site' is used throughout this thesis to describe a small genomic region of around 10 nucleotides where crosslink events are highly enriched. Thus, binding sites are always deduced from detected crosslink events. Next, I also showed how binding site reproducibility can be ensured over several biological replicates, followed by the demonstration of useful downstream analysis such as the identification of target genes and transcripts.

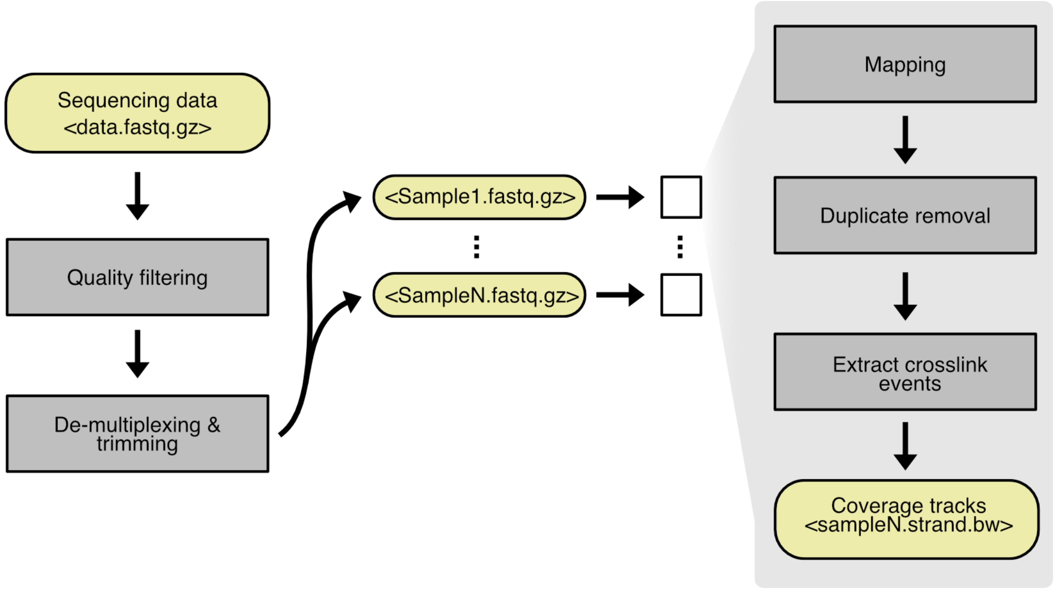

**Figure 3.1: Overview of the preprocessing workflow.** First quality filtering, trimming and sample de-multiplexing is applied on the entire library. Next, individual samples are mapped and coverage tracks are generated after the removal of technical duplicates. Processing steps are highlighted in grey, while resulting data files are shown in yellow. Figure adapted from (Busch et al., 2020).

### 3.1.2 The transcriptome-wide identification of RBP binding sites

RBPs typically bind to a variety of binding sites in coding and non-coding RNAs with a broad affinity range for the RNA sequence, RNA structure or a combination of both. Datasets are influenced by this binding behavior as well as by variable noise levels, such as technical aspects of the protocol or the efficiency of the immunoprecipitation. The analysis part that captures regions of increased crosslinking in short distinct clusters is known as ’peak calling’ (Chakrabarti et al., 2018). Up to this date a variety of different such tools exists, tailored to a specific CLIP protocol or adapted from DNA-based protocols such as ChIP-Seq analysis (Drewe-Boss et al., 2018). Here we used PureCLIP as a peak calling tool, since it is specifically designed to capture footprints of truncation-based protocols like iCLIP and eCLIP (Krakau et al., 2017).

To enhance the sensitivity of the peak detection reads from all four replicates were merged into a single file (Figure 3.2 A). The exemplary section of the *UBA2* gene showed two regions of dominant crosslink event pileups. Although visible in each of the individual replicates, these regions were most prominent in the merge of all replicates. This indicated that crosslink events caused by specific protein binding added up between replicates. More randomly distributed crosslink events, probably caused by unspecific experimental biases did not show the same effect. Thus, replicate merging was likely to enhance the signal from potential binding sites, while not amplifying protocol biases. PureCLIP-based peak calling resulted in a total of 1,929,191 crosslink sites. The term 'crosslink sites' is used throughout this thesis to describe a single nucleotide-wide position on the genome, that was called significant by PureCLIP. Significant sites are detected by prediction based on a hidden Markov model (HMM). Further, PureCLIP offers the automated merging of neighboring crosslink sites into larger regions. These regions were meant to capture the width of the RNA-binding footprint of the studied RBP. However, the size of these regions varied considerably, which could impair the downstream analysis (Figure 3.2 B). For example, more than 25% of all crosslink sites remained unmerged and thus spanned only a single nucleotide, whereas also regions larger than 24-nt could be observed. This again showcased the need of defining a unique binding site width to make conclusive statements about the binding behavior of the present RBP.

In order to remove low quality binding sites, I applied a filter on the PureCLIP-derived binding site strength score. PureCLIP computes a strength score for each single nucleotide crosslink site. This score is deduced from the log posterior probability ratio between the first and second most likely state of the HMM. It thus represents a binding site strength measure that is independent of the underlying transcript abundance (Krakau et al., 2017). We observed that the score distribution showed a near normal distribution with an added tail towards small values (Figure 3.3 A, B). These low-affinity positions were removed based on a threshold at the 5% quantile. This lead to an approximately normal shaped distribution among the 1,795,322 resulting crosslink sites. Depending on the binding mode the detected RBP footprint might differ. Thus, the computation of a unique binding site width has to be adapted whenever a new experiment is analyzed. To deduce an appropriate binding site width for U2AF65, we analyzed the spread of crosslink events around the center of the summarized binding sites (Figure 3.3 C, D). This can be done for a range of potential binding site widths. Here we decided to start with the smallest possible width of 3-nt, followed by an intermediate width of 9-nt and a large width of 29-nt. Thereby odd numbers were chosen to keep binding sites symmetrical. We observed that the majority of crosslink events piled up in a region of ten nucleotides around the binding site center. This clearly indicated that a binding site width of 29-nt would be too large. On the other hand, a binding site width of 3-nt appeared to be too small, since not all of the signal was captured. Thus, we decided to merge crosslink sites into 9-nt wide binding sites in the case of U2AF65.

Crosslink sites were merged by concatenating sites that were closer to each other than 8-nt. This ensured that the resulting 9-nt wide binding sites were not overlapping each other. The resulting regions varied in width and thus regions longer than 9-nt needed to be reduced, while shorter regions had to be elongated (Figure 3.4) (see methods 2.1.2). However, this procedure allowed also non-crosslink sites to arrive in binding sites. In order to control the number of these non-crosslink sites, we applied different filtering thresholds (Figure 3.5 A). Per definition each binding site harbored at least one crosslink site. A minimum of one crosslink site per binding site is thus the most inclusive threshold possible. With increasing stringency, the number of binding sites that passed the threshold declined constantly. For example, 91% of all binding sites were supported by two crosslink sites, whereas only 9.5% were supported by six crosslink sites. Based on the binding site definition, most binding sites showed the highest summed up crosslink signal directly at the center position (Figure 3.5 B, C). Binding sites with a higher proportion of crosslink sites on the other hand showed a broader crosslink profile. This hints towards a potential enrichment of neighboring binding sites on highly abundant transcripts. In order to achieve a balance between stringency and specificity we decided for a minimum threshold of two crosslink sites for each U2AF65 binding site. We further required each binding site to harbor a crosslink site at the center position. This center position is also required to show the maximum crosslink events within a binding site. In total these steps yielded 301,588 non-overlapping 9-nt wide binding sites (see table 3.2).

In summary, the described steps allow for an accurate definition of equally sized binding sites over the entire transcriptome. Potential pitfalls arise in the post-processing of crosslink sites, thus careful filtering steps needed to be applied to ensure the reliability of the final binding site.

**Table 3.2: Merging of crosslink sites into binding sites.** For the U2AF65 binding site processing in each filtering and resizing step the number of remaining regions is given. The initial crosslink sites were merged (resize routine) and filtered by the resulting width (position filter). Further the center of each binding site was required to be covered by a crosslink site, while showing the highest number of individual crosslink events (center crosslink and center maximum, respectively). Lastly a filter on the total number of crosslink sites is applied (crosslink site filter).

| Processing step | Number of crosslink/ binding sites |
|---|---|
| Crosslink sites | 1,795,322 |
| Resize routine | 332,949 |
| Position filter | 331,200 |
| Center crosslink | 330,820 |
| Center maximum | 323,086 |
| Crosslink sites filter | 301,588 |

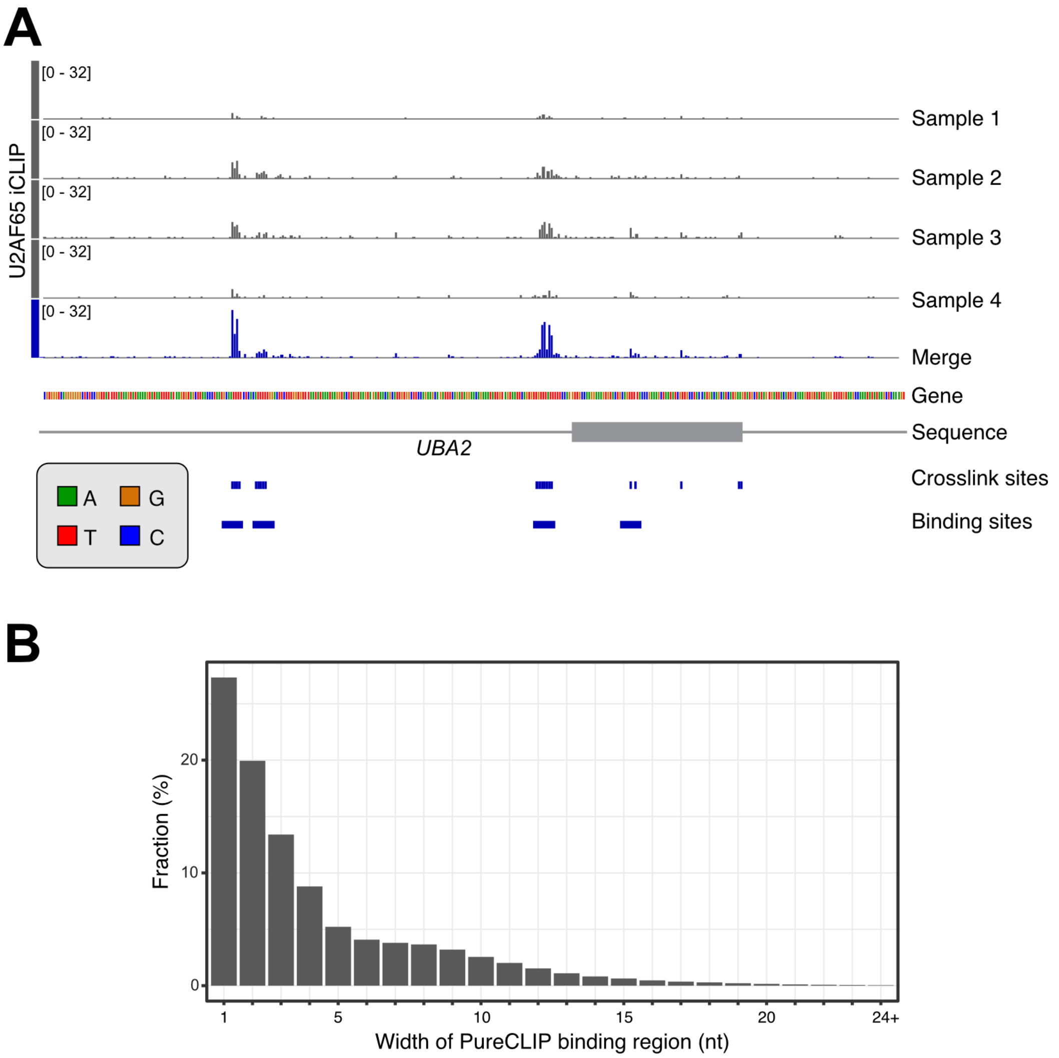

**Figure 3.2: Peak calling with PureCLIP.** (A) Overview of the *UBA2* gene locus showing the individual crosslink events for each sample as well as the merge of all four replicates. The PureCLIP-computed crosslink sites are shown underneath, together with the final U2AF65 binding sites after post-processing. (B) Distribution of the width of PureCLIP binding regions resulting from the concatenation of the predicted crosslink sites. Figure adapted from(Busch et al., 2020).

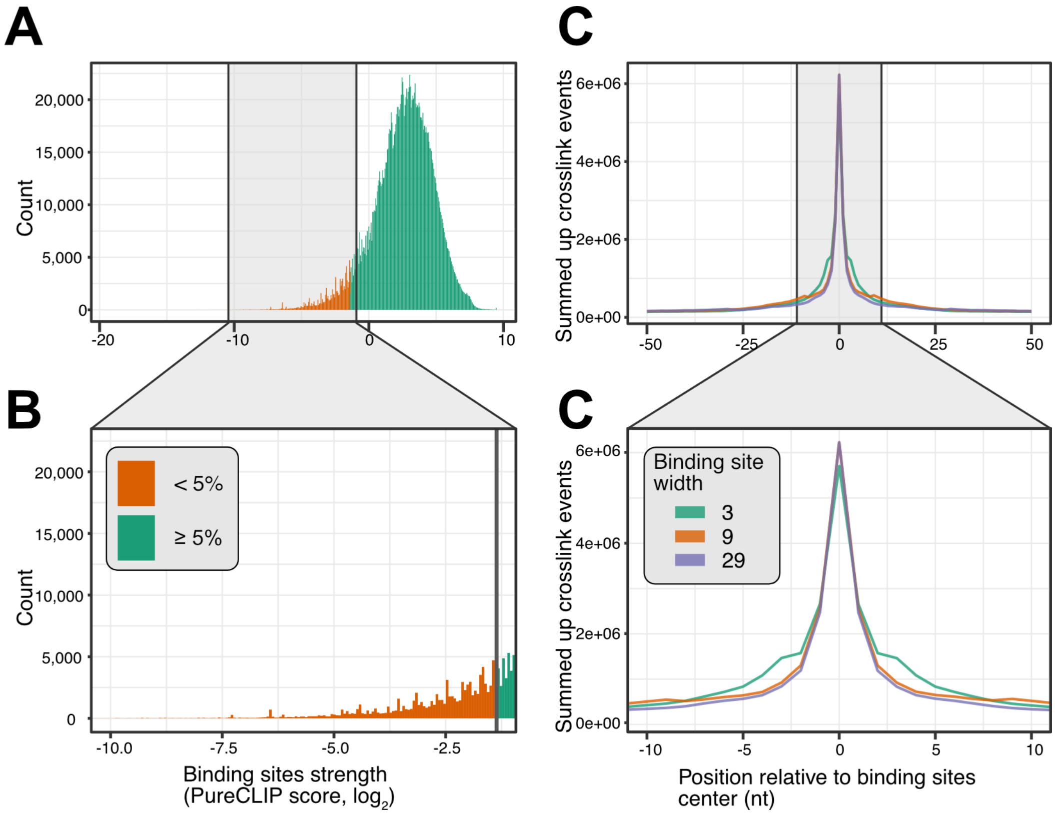

**Figure 3.3: Estimation of the binding site width.** (A, B) Distribution of the PureCLIP-score associated with each crosslink site. The 5% line indicates the score threshold that is used to remove poorly supported sites. (C, D) Meta profile of crosslink events around the center of merged binding sites with the indicated width. Binding sites were aligned at their central position and crosslink events were counted in a 100-nt window around that center.

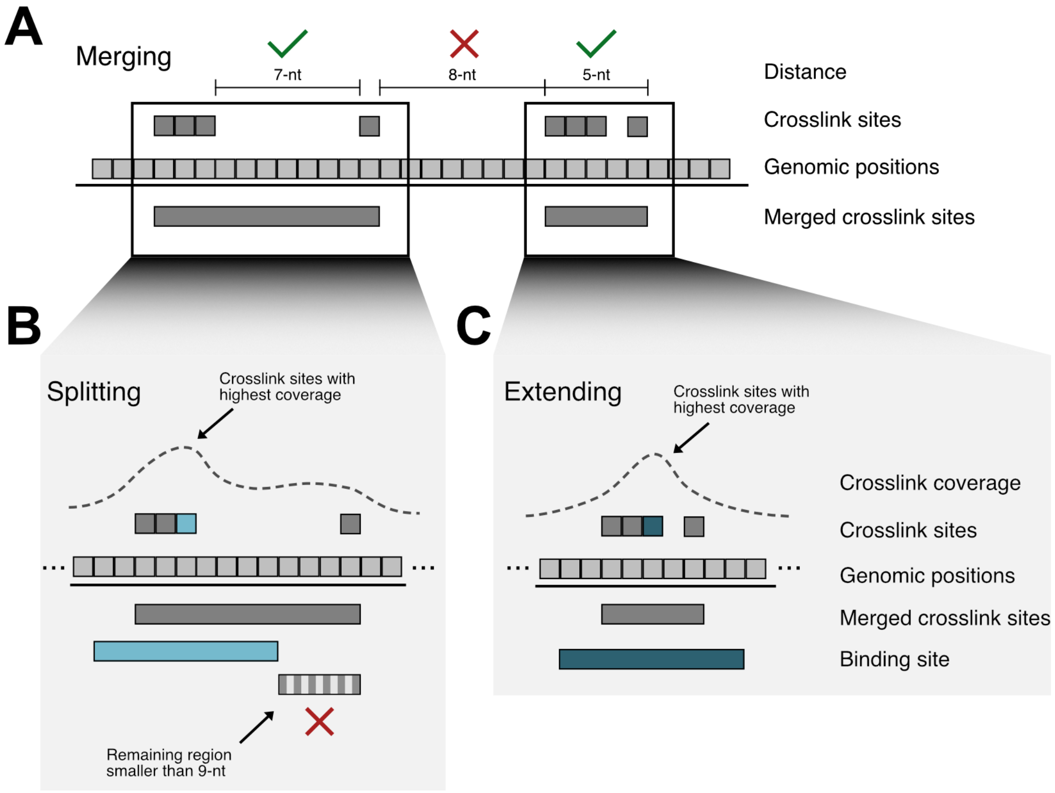

**Figure 3.4: Merging of the crosslink sites into binding sites.** Crosslink sites are subsequently merged into binding sites of 9-nt. (A) Crosslink sites closer than 8-nt are fused to merged sites. (B) Merged sites longer than 9-nt are iteratively splitted. (C) Merged sites shorter than 9-nt are extended symmetrically. In both cases the crosslink site with the highest signal is chosen as binding site center.

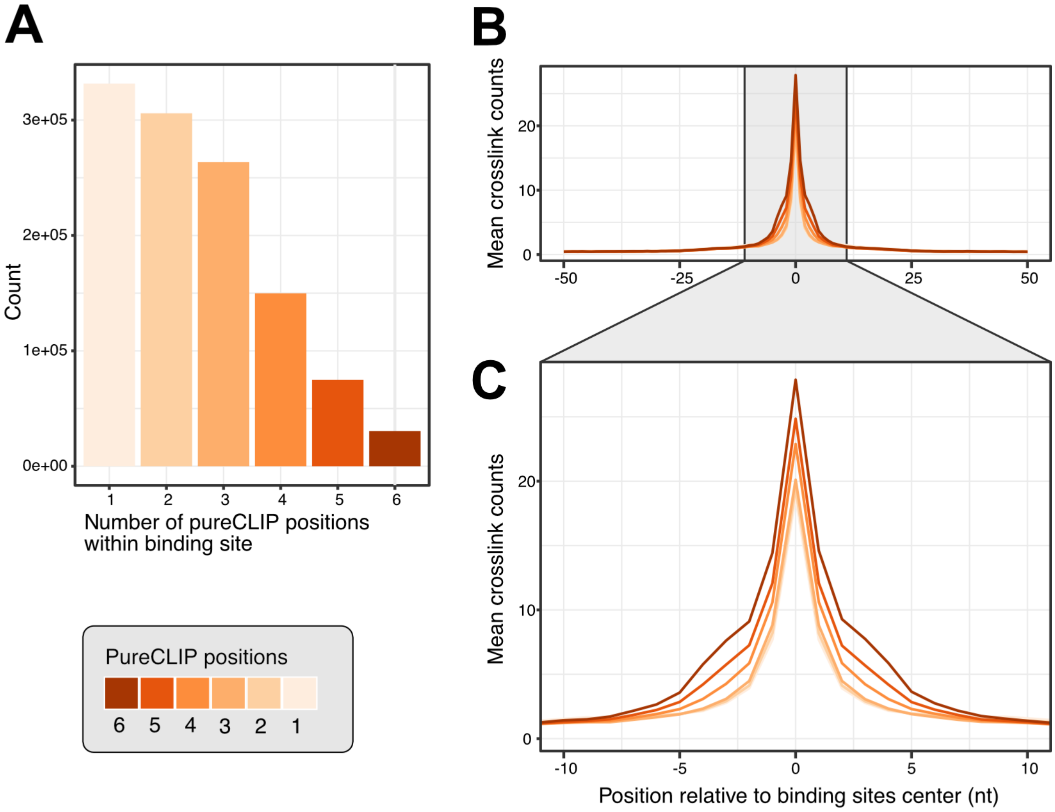

**Figure 3.5: Post-processing of computed binding sites.** Crosslink sites were merged into 9-nt wide binding sites, which are further filtered by additional constraints. (A) Bar chart indicating the number of binding sites retained given the indicated number of crosslink sites that overlapped a given binding site. (B, C) Meta profile of crosslink events around the center of the 9-nt wide binding sites. Binding sites were aligned at their center position and the mean number of crosslink events were counted in a 100-nt window around the center.

### 3.1.3 How to assess binding site reproducibility among replicates

Initial peak calling was based on the merge of all replicates to enhance the binding signal. This allowed the peak caller to also identify potential low-affinity binding sites, or binding sites from lowly abundant transcripts. On the other hand, this procedure might create artificial binding sites, caused by erroneous amplifications or sequencing biases in one of the replicates. To control for these types of errors, we implemented a replicate reproducibility filter. Since replicates might vary in their sequencing depth applying a single count-based threshold over all replicates would oversimplify the problem. We therefore developed a quantile-based approach to account for each replicate individually (Figure 3.6 A). Here we counted the number of crosslink events per binding site over the number of binding sites. The resulting distribution showed for each replicate the degree of contribution to all binding sites. Sample 1 for instance showed a left-tailed distribution with many binding sites not covered at all and only few binding sites covered by 40 or more crosslink events. Sample 3 exhibited the opposite behavior, such that almost all sites were covered with at least one crosslink and a substantial number of binding sites were covered by 40 or more crosslinks. To account for this huge variability, we proposed a quantile-based approach to define a replicate-specific crosslink threshold. Here we used the 10% quantile, deeming all binding sites above this threshold to be supported by the respective replicate. In addition, we adopted a lower boundary of two crosslink events, to ensure a certain minimum signal level. A further level of control was integrated by the number of replicates that need to support a specific binding site (Figure 3.6 B). Depending on the desired stringency any threshold from all to a single replicate support can be applied. In the present example, a binding site was assumed to be reproducible if the respective threshold was met by at least three of the four replicates. This filter removed 17.5% of the initially computed binding sites, resulting in 248,916 reproducible binding sites. This showcased how the peak calling step can be boosted by replicate merging while also ensuring sufficient individual replicate support. The described method can be adapted to any number of replicates or conditions.

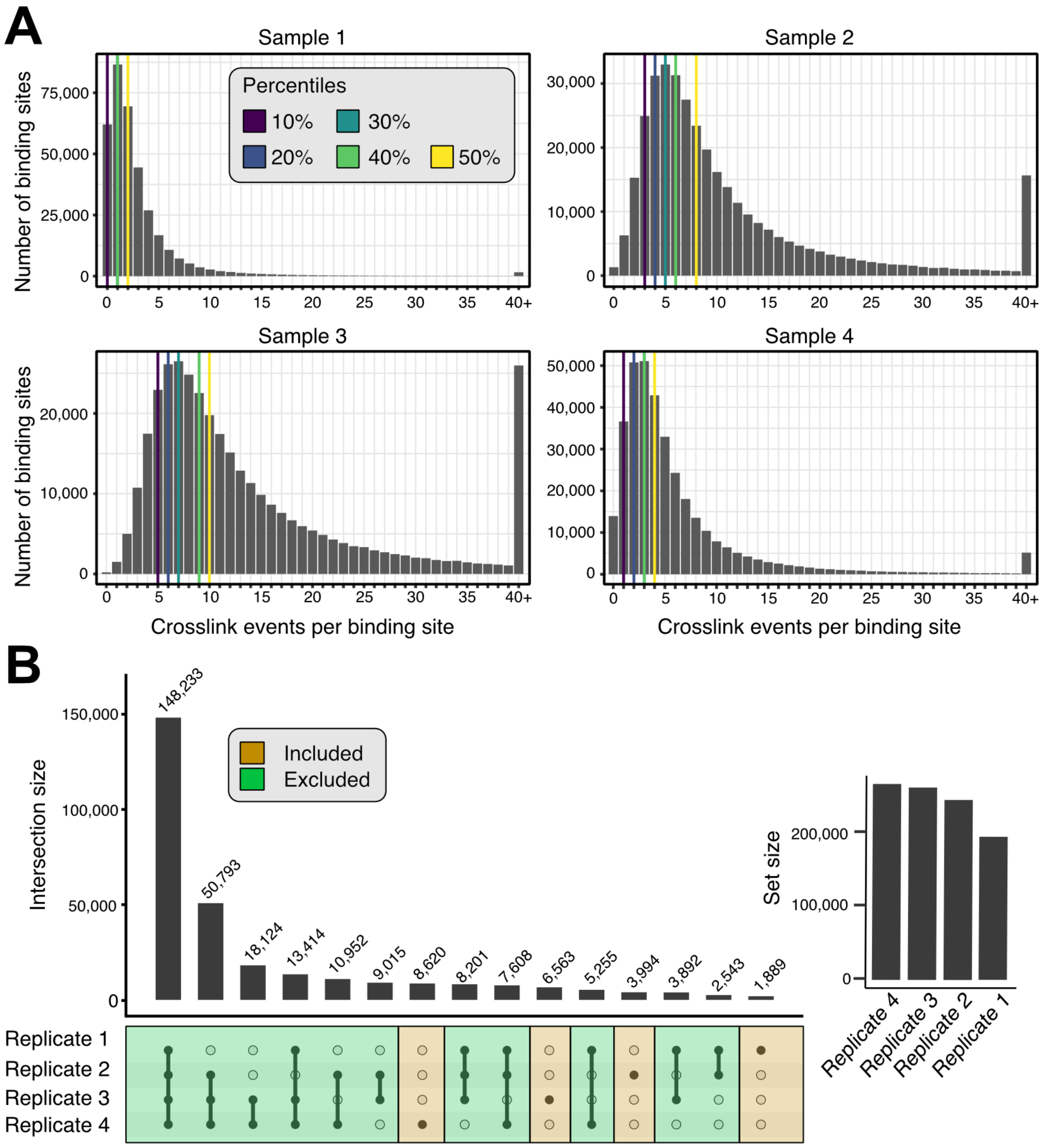

**Figure 3.6: Reproducibility assessment between replicate experiments.** (A) The distribution of crosslink events per binding site is shown for all four U2AF65 replicate experiments. For each replicate five possible thresholds are indicated by the 10% - 50% quantiles which represents different stringency thresholds. Here the 10% quantile is used as quality cutoff. (B) Summary of the binding site numbers that are shared between the different replicates. Reproducible binding sites must be supported by any three of the four replicates, indicated by the color code. The number of binding sites in each set is shown to the right.

### 3.1.4 Accurate annotation of gene and transcript regions

In order to describe the binding spectrum of a given RBP one is typically interested in the proportion of bound genes. This requires orthogonal data sources to be overlaid with the previously defined binding sites. Depending on the source, these annotations might differ in scope and reliability. Annotations provided by GENCODE/ ENSEMBL for example provide the full spectrum of putative isoforms, NCBI RefSeq in contrast provides a manually curated set (Wheeler et al., 2007; O'Leary et al., 2016). For the analysis of the U2AF65 data we decided to use annotations provided by GENCODE (version 29 of the human genome assembly version GRCh38). To remove redundancies and to filter for annotation reliability, we applied additional filtering steps on the gene and transcript support levels. We retained only genes with a support level $\leq 2$ and transcripts with a support level $\leq 3$. These filtering steps reduced the number of annotated genes by 13% and the number of annotated transcripts by 42%. This is particularly important since binding sites might overlap with several different annotations, which complicates the assignment process. These issues can potentially be resolved by the application of hierarchical rules, such as prioritizing protein-coding genes over non-coding RNA genes. However, one has to be aware that any of these rules might impact further analysis downstream. In the present case we decided for a conservative approach and completely removed ambiguously overlapping binding sites (2.1%). All remaining binding sites expressed a U2AF65-specific binding spectrum that was heavily dominated by protein-coding genes (Figure 3.7). In total 96.4% of all binding sites resided in protein-coding genes. These genes made up a total 90.4% of all genes to which a binding site could be unambiguously assigned. This also highlights that a binding spectrum of an RBP can be described from a gene or binding site driven perspective. For example a gene might be described as a target, based on a single or hundreds of binding site overlaps. It is thus important to describe the binding behavior in the context of bound genes as well as specific binding sites.

Transcript annotations for protein-coding genes discriminate between introns and exons, where the latter are further divided into coding sequence as well as the 3' and 5'UTRs. Integrating this type of information is particularly important when resolving the binding spectrum of an RBP, since its relative positioning in the transcript hints towards potential functions. For example, binding within intronic regions might point towards a role in pre-mRNA processing, while a positioning in the 3'UTR hints towards for example translational regulation. Mechanisms that give rise to multiple different transcript isoforms per gene such as for example alternative splicing or alternative polyadenylation increase the complexity of the assignment process (Lee and Rio, 2015; Barbosa-Morais et al., 2012). Annotations of these transcript isoforms usually overlap by a large degree, thus making it very difficult to trace a binding site back to a single specific isoform. With 9.7% of all binding sites overlapping different transcript regions, the problem was much more prevalent than in the gene type overlaps described above (Figure 3.8 A). For instance, 17,743 binding sites overlapped with at least two different transcript regions. Of note, we observed a total of 1,977 binding sites that did not overlap with a single transcript region. These binding sites resided within the annotated range of a protein-coding gene, but outside of any annotated transcript for the particular gene. In the present case we removed these suspicious cases resulting in 217,909 final binding sites (see table 3.3).

**Table 3.3: U2AF65 binding site processing overview.** The number of initial crosslink sites was filtered by their PureCLIP score (global filter) and merged into 9-nt wide binding sites (merged sites). These were accounted for replicate reproducibility (reproducible sites) and assigned to the final gene and transcript regions (assigned gene target and assigned transcript target, respectively). For each filtering step also the number of target genes and detected crosslinks over the indicated number of binding sites is given.

| | Number of crosslink/ binding sites | Number of target genes | Number of crosslinks |
|---|---|---|---|
| Crosslink sites | 1,889,813 | 10,804 | 14,453,030 |
| Global filter | 1,795,322 | 10,792 | 14,162,759 |
| Merged sites | 301,588 | 10,476 | 11,891,988 |
| Reproducible sites | 248,916 | 10,249 | 11,187,556 |
| Assigned gene target | 226,561 | 9,416 | 9,643,611 |
| Assigned transcript region | 217,909 | 8,766 | 9,281,805 |

To resolve the transcript annotation challenge, one again has the option between the application of specific rules and the conservative approach to discard all ambiguous binding sites. In the present case we decided to implement a majority vote system together with a hierarchical rule strategy, prioritizing the type of transcript region that was most often overlapping (Figure 3.8 B, C). In the case of U2AF65 we observed that 81% of all binding sites resided in introns, followed by 3'UTR (10.5%), CDS (7.6%) and 5'UTR (1%). It is worth mentioning that for most RBPs binding to introns does make up a high percentage. This is due to the different region size. With an average length of 6,529-nt introns are typically much longer than CDS (161-nt), 3'UTRs (562-nt) or 5'UTRs (132-nt). We resolved this issue by dividing the total binding site counts per transcript region by the respective length (Figure 3.8 D). However, such a relative enrichment is heavily in favor of shorter regions, thus underestimating the influence of intronic binding. This is especially the case when binding is focused on specific intronic regions such as the 3' and 5' splice site in the case of U2AF35 and U2AF65 (Jeong, 2017). We therefore recommend that both statistics, absolute binding as well as binding relative to the region size, have to be looked at simultaneously.

In summary, I presented an accurate workflow to define the specific binding spectrum of an RBP based on high-throughput iCLIP sequencing data. I described how the performance of the peak calling can be pushed while preserving replicate integrity. Additionally, I described a method to define binding sites of equal width based on PureCLIP-called crosslink sites. Lastly the integration of orthogonal annotation data from common sources was used to narrow down the specific binding profile of U2AF65. In this case we identified 217,909 binding sites which revealed a preference for intronic binding. This nicely matched the known function of U2AF65 as a splicing regulator and its role in early spliceosome formation by splice site recognition (Jeong, 2017).

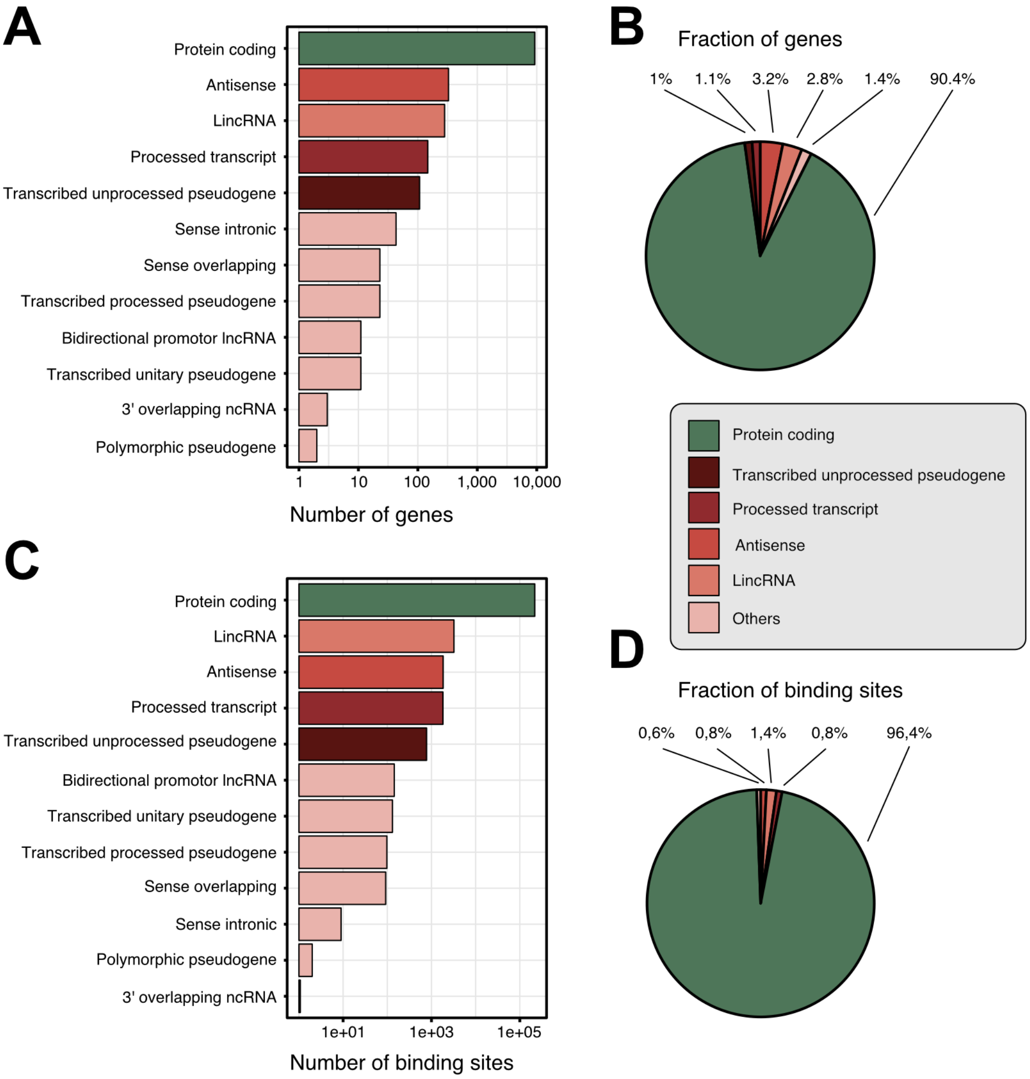

**Figure 3.7: Gene level assignment of binding sites.** Reproducible binding sites overlap with gene annotations. (A, B) The gene perspective shows the number of different gene target types that overlap with at least one binding site. Absolute numbers are given in (A), whereas relative fractions are shown in (B). (C, D) The binding site perspective that shows how many binding sites overlap with the indicated type of target gene. Absolute numbers are given in (C), whereas relative fractions are shown in (D).

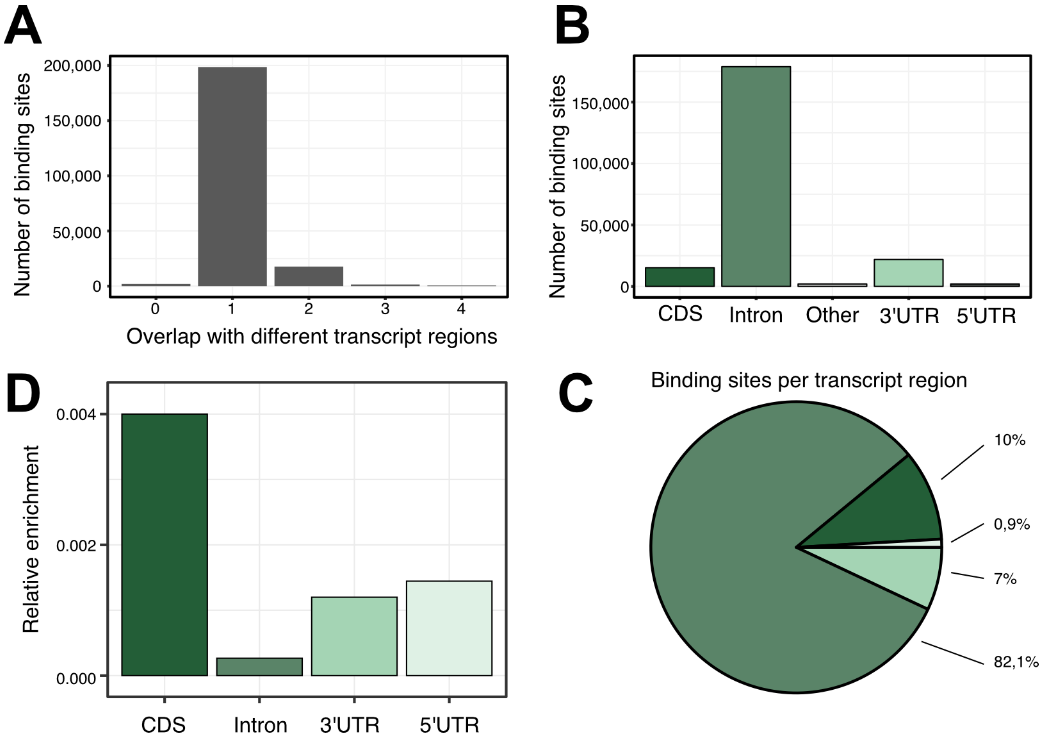

**Figure 3.8: Transcript level assignment of binding sites.** (A) Bar chart indicating the number of binding sites that overlap with zero, one or more conflicting annotations of different transcript regions. (B, C) Bar and pie chart giving the number of binding sites assigned to a specific transcript region after resolving overlaps. Absolute numbers are given in (B), whereas relative fractions are shown in (C). (D) Bar chart that shows the relative enrichment of each region. Absolut numbers of binding sites are normalized by the summed length of the respective transcript region.

## 3.2 The role of SRSF6 binding in human pancreatic beta-cells

In diabetes pancreatic beta-cells face increased apoptosis, in parts, based on the expression levels of GLIS3. It could be observed that SRSF6 is a downstream target of GLIS3. The mechanism of the GLIS3-mediated regulation was shown in 2013, where also the influence of SRSF6 was first noticed (Nogueira et al., 2013). This was followed up in 2018 where the influence of SRSF6 on beta-cell survival was shown directly via KD experiments (Juan-Mateu et al., 2018). These results originated from the group of Décio Eizirik (Brussels, Belgium), with which we collaborated for the current project. We followed up on their results, and the second aim of this thesis sets out to characterize the SRSF6-dependent transcriptome changes in detail.

Individual nucleotide resolution UV crosslinking and immunoprecipitation (iCLIP) experiments were carried out for the splicing factor SRSF6 in the human pancreatic beta-cell line EndoC-$\beta$H1. Four biological iCLIP replicates were used to describe the transcriptome-wide binding spectrum of SRSF6 and integrated with associated alternative splicing (AS) profiles to gain mechanistic insights of the SRSF6-mediated splicing regulation.

The project was carried out in cooperation with the group of Décio Eizirik and the group of Julian König (IMB, Mainz) and was recently published in the *Life Science Alliance* journal (Alvelos et al., 2020). The initial iCLIP experiment as well as protocol adaptation and optimization were performed by Ines Alvelos of the Eizirik group. Final iCLIP library preparation was done at the IMB in Mainz by FX Reymond Sutandy and sequencing was performed at the IMB core facility. Initial sequencing quality control as described above was performed by Dr. Anke Busch (IMB, Mainz). RNA-Seq data for the AS analysis was taken from a recently published research project of the Eizirik group (Juan-Mateu et al., 2018). All splicing event validations by RT-PCR were also performed by Ines Alvelos. The remaining bioinformatic analysis were performed by myself. This included the processing of the iCLIP data using the workflow described above, as well as the characterization of the SRSF6-specific binding motif. I also analyzed the RNA-Seq data to detect AS events, which I integrated with the iCLIP-derived binding sites. This resulted in the discovery of SRSF6-mediated AS regulatory events in the context of diabetes. A summary of all datasets used in this study is given in tables 3.4 and 3.5.

**Table 3.4: Summary of the SRSF6 RNA-Seq experiment in EndoC-$\beta$H1 cells.** The RNA-Seq experiment was performed in five paired replicates and yielded a total of 353,558,932 reads for the *SRSF6* KD and 334,335,772 reads for the WT condition.

| Sample | Condition | Sequenced reads | Mapped reads (%) |
|---|---|---|---|
| S1 | SRSF6 KD | 64,178,184 | 93.1 |
| S1 | WT | 61,408,355 | 93.3 |
| S2 | SRSF6 KD | 74,711,329 | 92.8 |
| S2 | WT | 42,236,869 | 92.7 |
| S3 | SRSF6 KD | 56,381,192 | 92.8 |
| S3 | WT | 74,475,186 | 92.9 |
| S4 | SRSF6 KD | 110,057,700 | 92.8 |
| S4 | WT | 91,835,192 | 92.7 |
| S5 | SRSF6 KD | 75,823,673 | 92.5 |
| S5 | WT | 89,816,570 | 93.1 |

**Table 3.5: Summary of the SRSF6 iCLIP experiments in EndoC-$\beta$H1 cells.** The iCLIP experiment was performed in four independent replicates, which yielded a total of 68,449,054 crosslink events.

| Sample | Sample ID | Barcode | Sequenced reads | Crosslink events |
|---|---|---|---|---|
| S1 | imb_ koenig_ 2017_ 10_ JKRS33_ SRSF6_ rep1 | NNNCGCCNN | 29,496,620 | 12,318,770 |
| S2 | imb_ koenig_ 2017_ 10_ JKRS33_ SRSF6_ rep2 | NNNTACGNN | 77,766,901 | 27,542,289 |
| S3 | imb_ koenig_ 2017_ 10_ JKRS33_ SRSF6_ rep3 | NNNATACNN | 38,186,626 | 16,161,814 |
| S4 | imb_ koenig_ 2017_ 10_ JKRS33_ SRSF6_ rep4 | NNNCGAGNN | 29,936,638 | 12,426,181 |

### 3.2.1 SRSF6 binds to thousands of protein-coding genes

An iCLIP experiment typically yields millions of individual single nucleotide crosslink events. Thus, extracting relevant information from the binding profile is not straightforward. Several summarizing and filtering steps have to be carefully performed to separate true binding signal from background noise. For this step we build upon the previously established iCLIP processing pipeline (see chapter 3.1) (Busch et al., 2020).

As a first step, the crosslink events from all four replicates were merged into a single file. The merged file was subjected to a peak calling step using PureCLIP. The resulting crosslink sites were filtered to retain only sites with the 95% highest PureCLIP scores. This first filter removed low-confidence sites probably resulting from very lowly abundant transcripts. Filtered crosslink sites were further merged into 9 nucleotide-wide binding sites, by employing the iterative merge and split routine, as described above (figure 3.4). Resulting binding sites were filtered to contain at least two crosslink sites. The center position of each binding site was also required to be a crosslink site as well as to exhibit the highest number of crosslink events (for details see methods 2.2.1). This process resulted in a set of 214,830 high-confidence binding sites. By applying stringent filtering, we ensured that any downstream result was based on binding sites with solid peak calling support. An overview of each peak-to-binding site processing step is given in table 3.6.

**Table 3.6: Merging of SRSF6 crosslink sites into binding sites.** For each filtering and resizing step the number of remaining peak regions is given. The initial crosslink sites were merged (resize routine) and filtered by the resulting width (position filter). Further the center of each binding site was required to be covered by a crosslink site, while showing the highest number of individual crosslink events (center crosslink and center maximum, respectively). Lastly a filter on the total number of crosslink sites is applied (crosslink site filter).

| Processing step | Number of crosslink/ binding sites |
|---|---|
| Crosslink sites | 1,600,050 |
| Resize routine | 269,353 |
| Position filter | 264,829 |
| Center crosslink | 264,418 |
| Center maximum | 259,544 |
| Crosslink sites filter | 214,830 |

Since peak calling was performed on the merge of all four replicates, the resulting binding sites had to be verified by the individual replicates. As described above the merging process boosted the power of the peak calling, while potential biases were removed by the reproducibility filtering. Erroneous amplifications for instance could lead to an artificially increased coverage in one of the replicates and by this introduce artificial binding sites, or mask actual binding sites supported by the other replicates. For each replicate the 20% quantile of the crosslink distribution per binding site was set as threshold, with a lower boundary of two crosslinks per binding site (figure 3.9 A). Replicate 2 for instance showed the overall highest number of crosslink events per binding site, which resulted in a threshold of at least six crosslink events. Replicate 1 in contrast showed a rather weak binding site coverage resulting in the minimum threshold to be used. In total, a binding site needed to be supported by at least three of the four replicates to be deemed reproducible and thus be retained for further analysis (figure 3.9 B). This procedure allowed to include also binding sites with weaker reproducibility support although the vast majority of 133,892 binding sites is supported by all four replicates. In total, these steps cumulated in 185,266 binding sites.

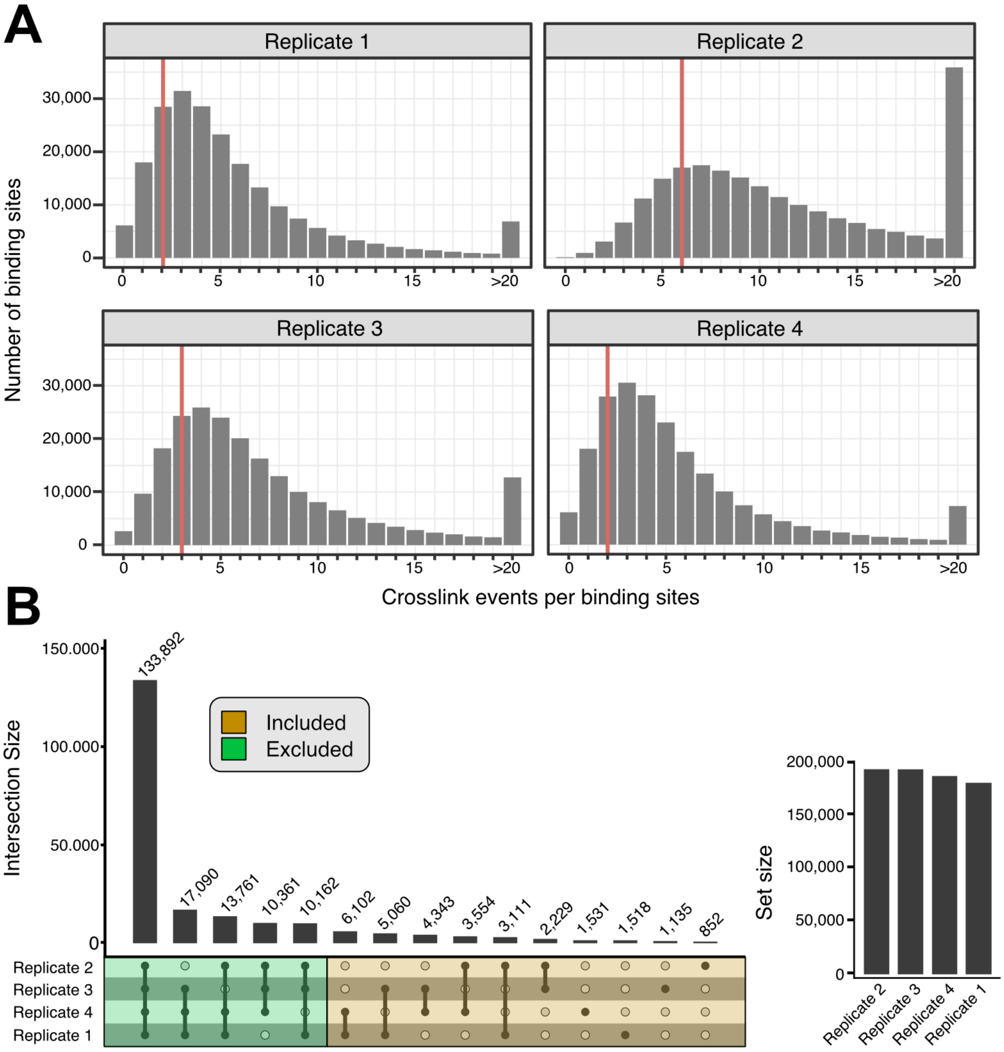

**Figure 3.9: Reproducibility filtering of the SRSF6 binding sites.** SRSF6 binding sites are reproducible among four replicates. (A) Bar charts showing the number of crosslink events per binding sites for each replicate. The red line indicates the 20% quantile, which was used as replicate-specific threshold. (B) Reproducible binding sites met the threshold of at least three different replicates (green). The bar chart to the right indicates the set size for each replicate.

In order to describe the targets that are bound by SRSF6, the final binding sites were overlaid with gene and transcript annotations, which resulted in 8,533 overlapping genes (for details see methods section 2.2.2). The vast majority of binding sites mapped to protein-coding genes (93%), followed by lincR-

NAs (3.2%) and antisense-RNAs (2.1%) (figure 3.10 A). This is expected, since SRSF6 as a splicing regulator assists in the early recruitment of core spliceosomal units and thus is involved to a large extend in pre-mRNA processing of protein-coding genes. To further narrow down the location of binding sites in transcripts, we focused on those in protein-coding genes (figure 3.10 B). Here, binding sites seemed to overlap intronic regions (48.3%) with a slight dominance over coding sequences (40.1%), followed by 3'UTRs (8%) and 5'UTRs (3.6%). Since the coverage over transcript regions is heavily influenced by the region length, we normalized binding site counts to the summed length of each transcript region (figure 3.10 C). This relative enrichment shows a strong dominance of CDS over introns. The strong preference of SRSF6 to bind in exons is further exemplified by the *CCDC50* gene (figure 3.11). A broad spread of crosslink coverage can be seen over the entire range of the annotated transcripts, but binding sites predominantly resided in exons. Introns on the other hand showed a more randomly distributed signal, probably pointing towards more unspecific background binding. An overall summary of all binding site processing steps is given in table 3.7.

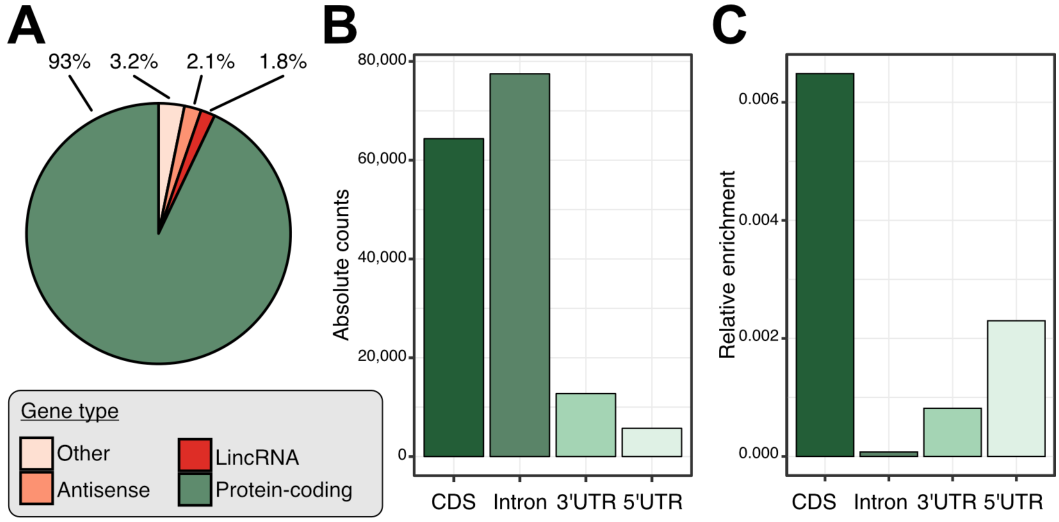

**Figure 3.10: SRSF6 predominantly binds to protein-coding genes.** (A) The pie chart shows the distribution of SRSF6 binding sites per annotated gene type. Gene types that made up less than 1% were summarized under the term 'Others'. (B) SRSF6 primarily binds to coding sequences (CDS). The bar chart shows the distribution of SRSF6 binding sites per transcript region on protein-coding genes. (C) Bar chart displaying the relative enrichment of binding sites per region. Binding site numbers per region were normalized by the summed length of the respective bound transcript regions.

The number of initial crosslink sites was filtered by their PureCLIP score (global filter) and merged into 9-nt wide binding sites (merged sites). These

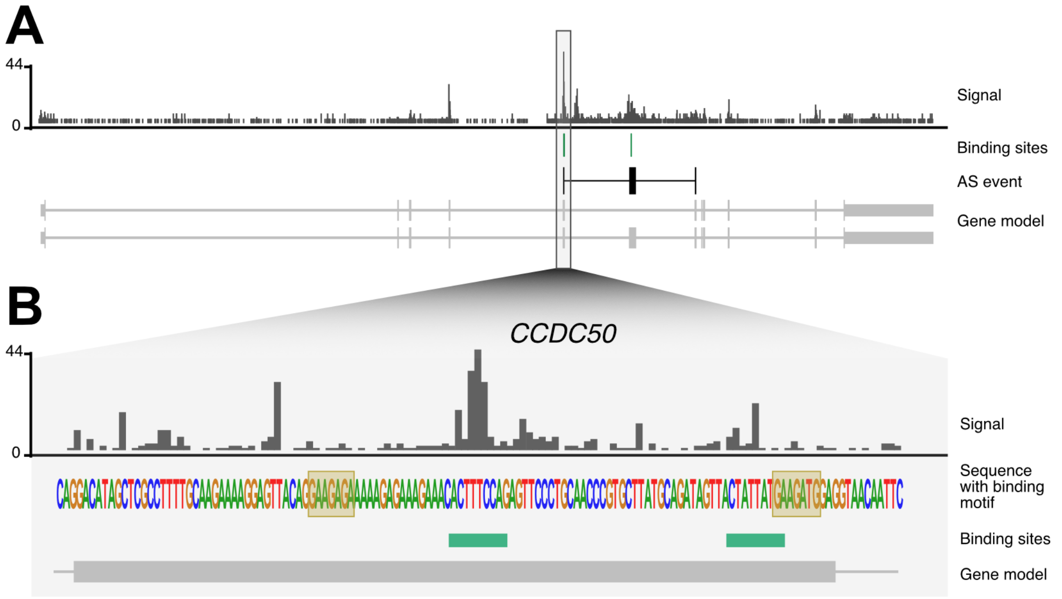

**Figure 3.11: SRSF6 preferentially binds on exons of the *CCDC50* gene.** Genome browser view of the *CCDC50* gene (A) of the SRSF6 iCLIP crosslink events (signal), with deduced binding sites (green) and the SRSF6 binding motif (yellow boxes). Binding is shown on selected annotated transcript isoforms (grey), with the SRSF6-regulated alternative exon alongside the flanking constitutive exons (black). (B) Zoom-in of the alternative exon region.

were accounted for replicate reproducibility (reproducible sites) and assigned to the final gene and transcript regions (assigned gene target and assigned transcript target, respectively). For each filtering step also the number of target genes and detected crosslinks over the indicated number of binding sites is given.

To gain functional insights into the binding spectrum of SRSF6, a Gene Ontology (GO) enrichment analysis was performed for all genes with bound transcripts (figure 3.12). A broad range of different GO terms were associated with the genes of SRSF6-bound transcripts, such as cell cycle progress, histone modification and DNA repair. This is expected for a splicing regulator that binds to the vast majority of all expressed transcripts. However, also terms pointing towards splicing and mRNA processing were found, which indicates a potential cross-regulation among the protein class of splicing regulators.

**Table 3.7: Summarized SRSF6 binding site processing overview.** The number of initial crosslink sites was filtered by their PureCLIP score (global filter) and merged into 9-nt wide binding sites (merged sites). These were accounted for replicate reproducibility (reproducible sites) and assigned to the final gene and transcript regions (assigned gene target and assigned transcript target, respectively). For each filtering step also the number of target genes and detected crosslinks over the indicated number of binding sites is given.

| | Number of crosslink/ binding sites | Number of target genes | Number of crosslinks |
|---|---|---|---|
| Crosslink sites | 1,684,248 | 11,857 | 13,053,409 |
| Global filter | 1,600,050 | 11,821 | 12,754,216 |
| Merged sites | 214,830 | 10,329 | 7,831,081 |
| Reproducible sites | 185,266 | 9,766 | 7,451,934 |
| Assigned gene target | 168,096 | 9,222 | 6,341,329 |
| Assigned transcript region | 160,320 | 8,533 | 6,036,281 |

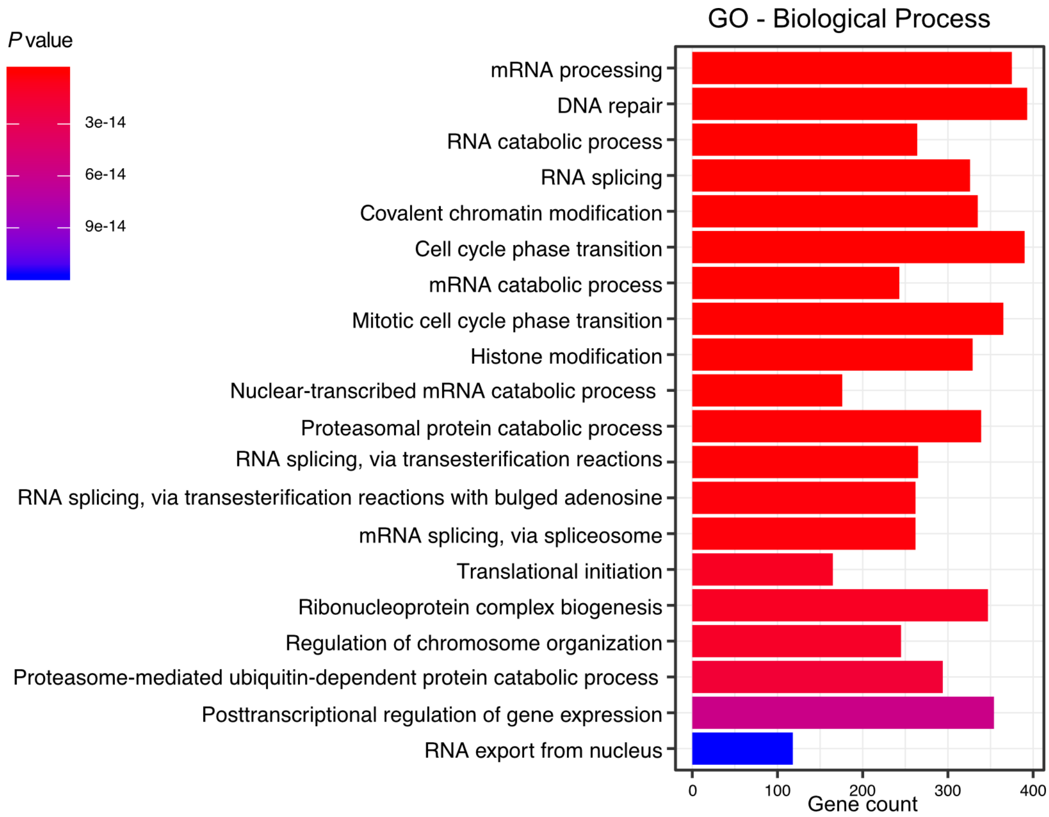

**Figure 3.12: SRSF6-bound genes are associated with different functions.** Gene Ontology (GO) enrichment in SRSF6-bound genes. Gene counts indicate the number of genes in the respective set. All expressed genes were selected as reference background and P values from the hypergeometric distribution are shown. A gene was deemed expressed if it exceeds a TPM threshold of 1.

### 3.2.2 Characterization of the SRSF6 binding motif

In order to interact and bind to their target RNA, RBPs typically consist of RNA-binding domains which specifically recognize short degenerate sequence motifs. The length, position and composition of such a motif varies largely between different RBPs. Thus, it is important to carefully define such a binding motif for every protein under study. To investigate the sequence footprint of SRSF6 we analyzed the local sequence content around the defined binding sites. For computational reasons binding sites were randomly down-sampled to 5,000 representative binding sites from introns and coding sequences (CDS). For a first quantitative description of the sequence content, all 1,024 possible pentamers were counted in three distinct windows. These windows consisted of the 9-nt binding site itself, as well as a 20-nt flanking region up- and downstream (figure 3.13 A). The mean count of each pentamer per window indicated an increased frequency of GA-rich pentamers in CDS and an increased frequency of U-rich pentamers in introns (figure 3.13 B). The pentamers GAAGA, AGAAG and AAGAA were most prominently enriched in CDS, especially in the window downstream of the binding site. Binding sites in introns on the other hand were dominated by uridine pentamers, most noticeably UUUUU, with the highest observable frequency directly at the binding site. Such an enrichment is well documented for CLIP-based experiments, since UV crosslinking is biased towards uridines. It is also typical that this bias is enriched in non-coding intronic regions compared to CDS (Sugimoto et al., 2012; Haberman et al., 2017; Chakrabarti et al., 2018). Therefore, the immediate position of binding sites follows to some extend this so-called uridine bias. For that reason, the real sequence motif that is potentially recognized by SRSF6 can be found outside, but in close proximity to the binding site. The GA-rich pentamers identified in the flanking regions indicated exactly that.

We further focused our attention on these flanking regions and excluded the bias driven 9-nt binding site part. Since uridine-rich pentamers were also more frequently seen in introns, we limited the subsequent motif description on the CDS. Most pentamers showed little to now clear enrichment, thus we filtered all 1,024 initial pentamers by their observed frequencies and retained only the top 300 with the highest frequencies in the CDS. In order to capture start and end point of the enrichment, a larger window of 200-nt was spanned for the up- and down-stream flanking regions (figure 3.14 A). Pentamer frequencies were counted again and summarized in a heatmap to gain further positional insights. Next k-means clustering was used to split pen-

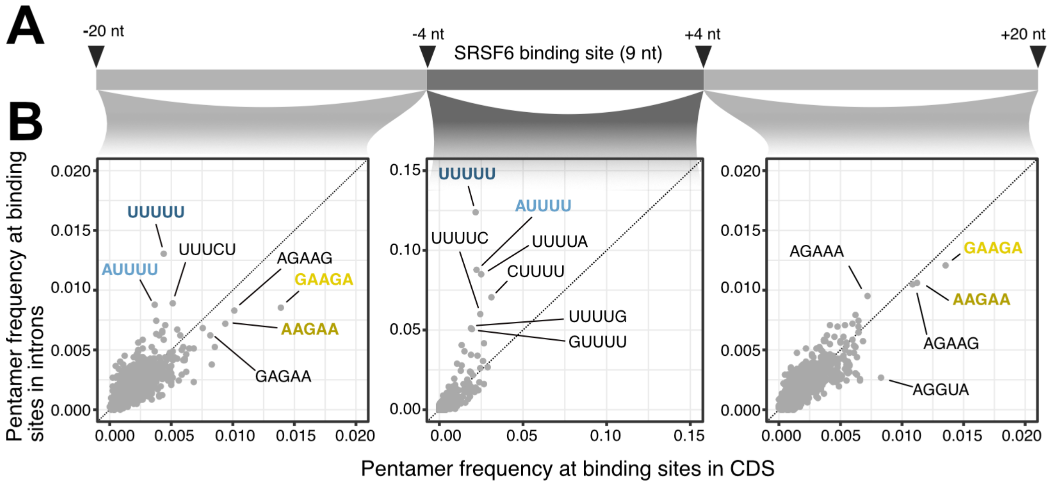

**Figure 3.13: GA-rich pentamers are enriched around SRSF6 binding sites in CDS.** (A) A schematic representation that indicates the three different windows in which pentamers were counted. The central 9-nt window of the binding site itself (dark grey) and the two flanking 20-nt windows (light grey). (B) Pentamers in CDS are enriched for G and A compared to introns, where UUUUU is most prominent.

tamers into three clusters based on positional preferences (figure 3.14 B-D). Cluster 1 almost exclusively consisted of AG-rich pentamers, that seemed to enrich towards the binding site. Cluster 2 on the other hand showed an inverse behavior, displaying the depletion of U-rich pentamers around the binding site. Cluster 3, as the largest cluster of all three, consisted of all other pentamers with no clear positional preferences. This indicated that the majority of the sequence signal was already summarized by cluster 1 and 2.

In fact, when computing meta-profiles based on the two most frequent pentamers in each cluster, the positional preferences became more clear-cut (figure 3.15). The GA-rich pentamers GAAGA and AAGAA displayed a rising frequency from up to 100 nucleotides before the binding site, that continued to around 25 nucleotides after the binding site and then dropped sharply. Mixed pentamers from cluster 3 exhibited no positional preferences and rather displayed the background level. U-rich pentamers spiked at the binding site center, strongly pointing towards the uridine bias. The GA-enrichment on the other side suggested a strong preference for SRSF6 to position precisely towards the end of such a GA-enriched region.

Binding motifs are usually short RNA sequences that can be recognized by specific RNA-binding domains (RBDs) of the RBP. These motifs can differ in their affinity based on their sequence composition. The strength of the protein-RNA interaction might depend on the number of proteins that assemble at a

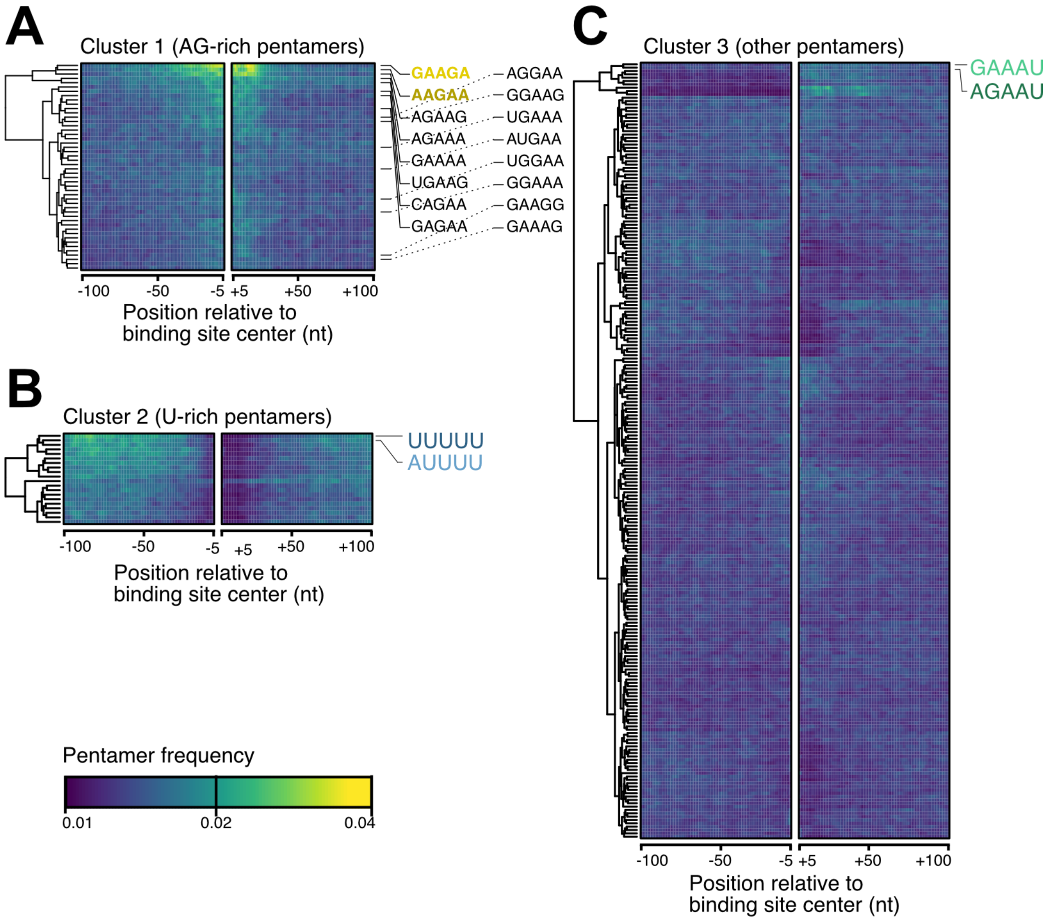

**Figure 3.14: GA-rich pentamers are enriched in windows flanking SRSF6 binding sites.** (A) Scheme of the extended 100-nt binding site flanking windows, where the binding site itself is excluded. (B) Cluster 1 from the k-means clustering of the pentamer heatmap shows mainly GA-containing pentamers that are enriched towards the center. (C) Cluster 2 is enriched for U-rich pentamers, which are depleted towards the center. (D) Cluster 3 displaying the remaining pentamers with no clear positional preference.

given motif, as well as the type or the number of RBDs that facilitate the contact (Lunde et al., 2007; Burd and Dreyfuss, 1994). Since the triplet GAA was common among most pentamers analyzed so far (figures 3.13, 3.14, 3.15), we asked whether such a triplet serves as a building block for a larger motif. Therefore, we counted the number of GAA triplets in a 49-nt wide window around the binding site center, to focus on the enriched region only. Next, binding sites were grouped based on the counted number of GAA occurrences and for each binding site the respective binding strength was assessed. Here binding site strength is accounted for in terms of the PureCLIP score which measures binding intensity independent of the underlying transcript abundance (Krakau et al., 2017). This score was computed for each crosslink site

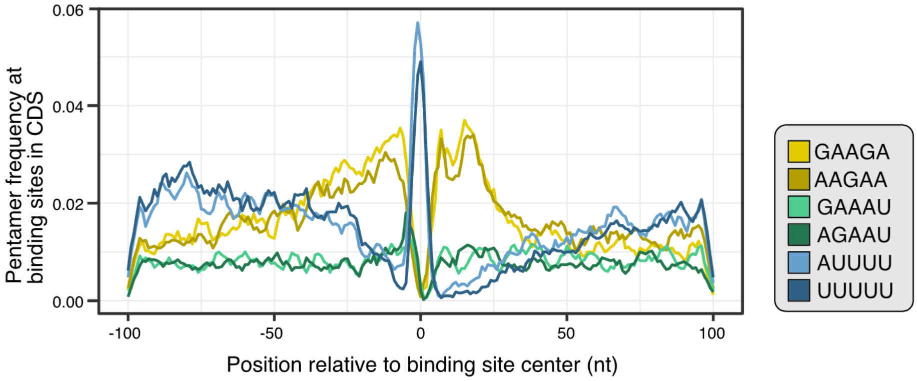

**Figure 3.15: SRSF6 binds predominantly towards the end of GA-rich pentamer stretches.** Metaprofile that shows the frequency of selected pentamers in a 201-nt window centered around SRSF6 binding sites. The two most enriched pentamers of each heatmap cluster are shown (figure 3.14). GAAGA and AAGAA frequencies increase towards the binding site and peak about 20-nt after the binding site. UUUUU and AUUUU are depleted in the GA-rich regions and peak directly at the binding site center.

and the respective mean over all crosslink sites was chosen as representative measure for each binding site. The same counting scheme was also applied to the reverse complement triple UUC, which serves as a control (figure 3.16 A). Interestingly, an increase of binding site strength coincided with an increase in GAA triplet numbers, starting from two up to seven or more triplet occurrences. For the reverse complement, such an effect could not be observed. This indeed pointed towards stronger SRSF6 binding in regions that exhibit a higher number of GAA triplets. This finding was further underpinned by the number of binding sites associated with the respective triplet occurrences (figure 3.16 B). With only 9 binding sites at seven or more triplets, UUC numbers declined much more rapidly compared to GAA, with 671 binding sites at seven or more triplets.

In a similar fashion we further analyzed whether gaps would be allowed in between repetitions of the GAA triplet. We compared the strength of binding site groups that harbored one (GAAxGAA) and two (GAAxxGAA) nucleotide gaps, as well as those with no gap (GAAGAA), in between two or more occurrences of the GAA triplet (figure 3.17). Strikingly already a single nucleotide gap disrupted the motif, causing the binding site strength to drop significantly. We observed 1,093 binding sites with a three-fold repetition of the GAA motif to exhibit the strongest binding, compared to 11,256 binding sites with a two-fold repetition. This indicated that SRSF6 binds strongest at binding sites with three repetitions of the GAA triplet with no gaps in between.

To complement the motif description analysis, we lastly conducted a *de*

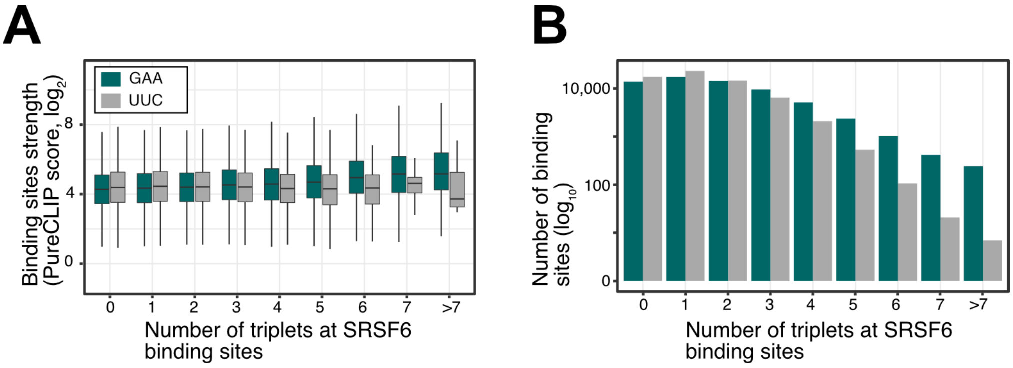

**Figure 3.16: SRSF6 specifically recognizes GAA triplets.** The binding site strength increases with the number of GAA triples. (A) The boxplot compares the binding site strength of the triplet GAA with the reverse complement UUC. The frequencies of these triplets are counted in a 30-nt window around the binding site center and each box shows the distribution of the associated binding site strength ($log_2$-transformed PureCLIP score). (B) Bar chart that shows the number of binding sites in each of the described categories.

*novo* motif search using DREME, this time including binding sites from all regions of protein-coding transcripts (Bailey, 2011). For all SRSF6 binding sites, a 201-nucleotide long sequence window was extracted and submitted to the search in order to capture the full sequence range. The resulting second-best hit motif nicely reflected the GAA enrichment detected by the k-mer analysis (figure 3.18 A). This motif was preceded by a U-rich motif, again reflecting the influence of the uridine bias (figure 3.18 B). Next the position weight matrix of the GAA-rich DREME motif was submitted to FIMO, in order to search for all binding sites that are supported by that motif (Grant et al., 2011). We found that a total of 25,148 (19%) of all binding sites showed the motif support, with binding sites overlapping the CDS showing the highest support, followed by introns, 3'UTRs and 5'UTRs (figure 3.18 C).

In summary we could show that SRSF6 positioning is determined by a GA-rich sequence motif. This motif is most likely constructed of GAA triplets as central building blocks and that higher, but uninterrupted, repetitions of these blocks lead to stronger SRSF6 binding on these sites.

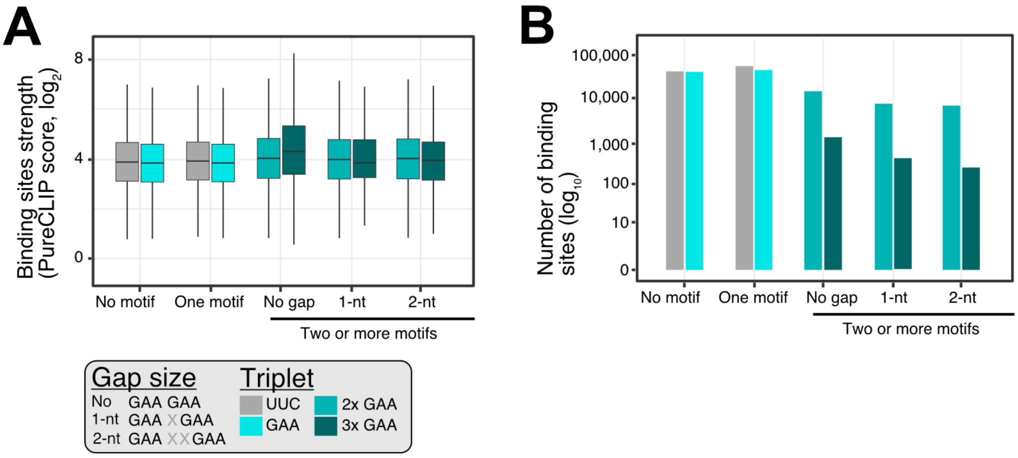

**Figure 3.17: GAA-triplets in direct sequence lead to stronger binding.** (A) Boxplot that shows the distribution of binding site strength ($log_2$-transformed PureCLIP score) for binding sites associated with two or more triplets with no, 1-nt or 2-nt gaps in between. These scores are compared to the strength of binding sites with no or one triplet (GAA, UUC) for comparison. B) Bar chart the shows the number of binding sites in each of the described categories.

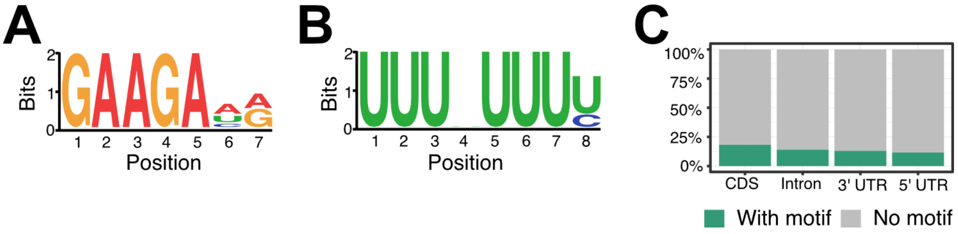

**Figure 3.18: *De novo* motif analysis indicates a GAA based binding motif for SRSF6** (A, B) Purine (second hit) and uridine (first hit) rich motifs computed from the motif enrichment analysis using DREME (Bailey, 2011). (C) The GAA-rich motif is present at 25,148 SRSF6 binding sites reinforcing the role of the GAA regions on SRSF6 binding.

### 3.2.3 *SRSF6* KD reshapes the beta-cell transcriptome

We followed up on the above analysis of the transcriptome-wide binding pattern of SRSF6 with the analysis of expression level changes upon *SRSF6* KD. As a major splicing regulator SRSF6 is in general involved in pre-mRNA processing. Thus, it might globally effect gene expression and transcript levels. In order to analyze differential expression in the context of pancreatic beta-cells and SRSF6 binding, we reanalyzed an RNA-Seq dataset published by our colleagues of the Eizirik group (Juan-Mateu et al., 2018). The dataset consisted of five individual replicates. Five of which captured a *SRSF6* KD condition, whereas the others represented the control state (table 3.4). We used the RNA-

Seq dataset for differential expression testing between the *SRSF6* KD and WT condition with DESeq2 (Love et al., 2014). Mapped reads were counted in annotated exons to quantify gene expression levels in both conditions

For an initially quality control principal component analysis (PCA) was performed. This type of analysis works best for homoscedastic data that shows the same range of variance at different ranges of mean values. For expression data, this is typically not the case since the absolute count differences grow larger, the higher a gene is expressed (Huber et al., 2002). To avoid this bias, we used the regularized-log (rlog) strategy developed by the authors of DESeq2. It essentially stabilizes the mean across different ranges of variance applying a $log_2$ derivate (figure 3.19 A). Resulting rlog transformed counts were forwarded to the PCA analysis which showed a clear separation of the samples based on their condition (figure 3.19 B). The scree plot further showed that PC1 comprised 68% of all between-sample differences (figure 3.19 C). PC2 and PC3 were more similar to each other with 14% and 7.5%, respectively. All other components showed only minor contribution and thus can be neglected. In total the PCA indicated high-quality data, ideal for a differential expression analysis, where no potential batch effects have to be accounted for.

We next went to test for gene-specific fold changes asking for expression changes between the WT and *SRSF6* KD conditions. In total we observed 6,893 genes to significantly change their expression levels (Benjamini-Hochberg adjusted $P$ value $< 0.001$) (figure 3.20 A). With 49.6% of these genes showing an up- and 50.4% showing a downregulation no general trend could be observed. Both groups of regulated genes showed an equally good support by all of the replicates (figure 3.20 B). The heatmap indicated that replicates reacted evenly in the indicated condition. This is further supported by k-means clustering performed on this heatmap, which nicely captured the two main groups of up- and downregulated genes. Functional gene ontology (GO) enrichment analysis did reveal a broad spectrum of affected functionalities. For example, genes were associated with terms like DNA replication, cell cycle and digestive systems (figure 3.20 C). However, most genes showed less than a 2-fold change. It is likely that these changes, although statistically significant, have little biological impact. In order to filter for larger changes, we applied a fold-change cutoff (absolute $log_2$ fold-change $\geq 1$). This greatly reduced the number of regulated genes to as little as 106, again not indicating a preference for either up- or downregulation. Among these were mainly housekeeping genes involved maintaining cell homeostasis, such as ribosomal genes or members of the solute carrier family (*SLC*) (Liu, 2019). Interestingly, with the *BCL2*

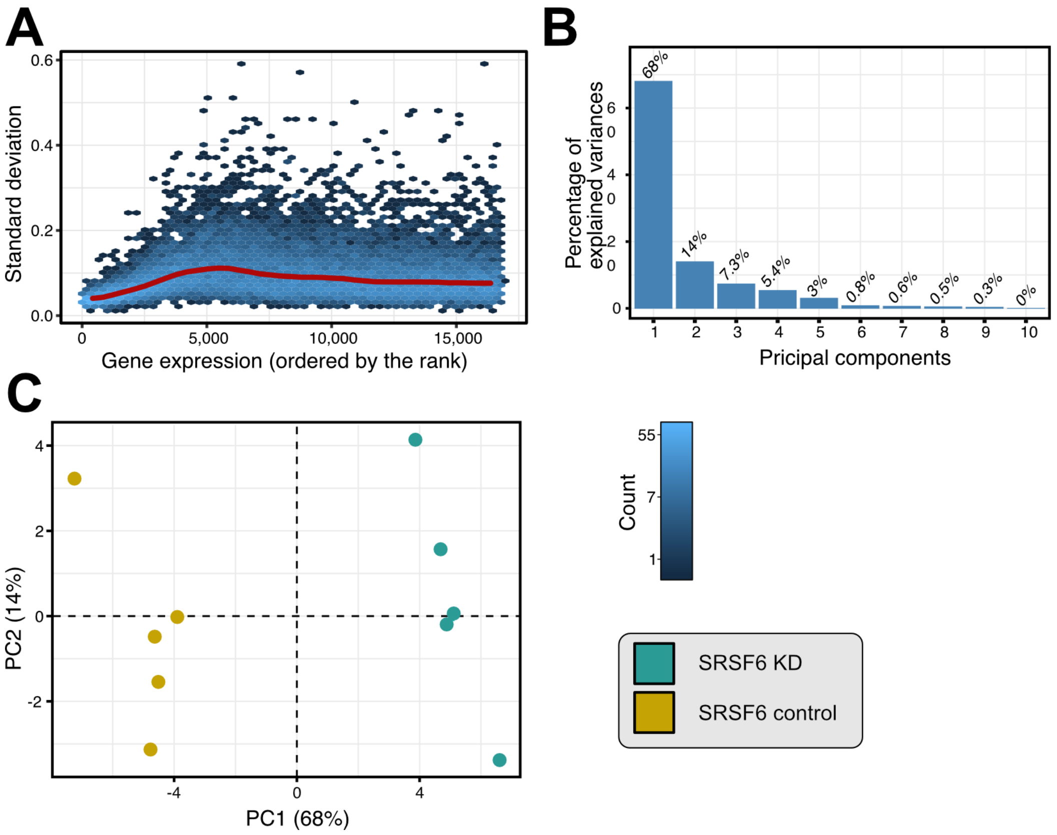

**Figure 3.19: RNA-Seq replicate expression is dominated by the *SRSF6* KD.** Principal component analysis (PCA) separates *SRSF6* KD replicates from controls. (A) Variance stabilizing transformation of the count data using the rlog transformation (Love et al., 2014). The standard deviation is shown over the mean expression, with a running median shown as a red line. (B) Scree plot indicating the percentage of the variance that each component contributes. (C) PCA plot of the first and second component (PC1, PC2) explaining over 80% of the observed variance.

Interacting Protein 3 Like (*BNIP3L*) a first hint towards a potential influence in cell apoptosis regulation is given (Edlich, 2018).

Since the differential expression analysis did not reveal many severe changes in transcript expression levels, we conducted alternative splicing analysis, which serves as an additional layer of transcriptional regulation (Lee and Rio, 2015). We quantified differences in splice junction-spanning reads using the replicate multivariate analysis of transcript splicing software (rMATS). Five different alternative splicing events were reported, namely: ”cassette exon” (CE), ”retained intron” (RI), ”mutually exclusive exon” (MXE), ”alternative 3’ splice site” (A3SS) and ”alternative 5’ splice site” (A5SS). With a total of 71.8% the majority of events corresponded to CE, followed by MXE, A3SS, A5SS and RI (figure 3.21 A). Since rMATS reported many splicing events overlapping between the different categories, we applied a stringent filtering approach. Thereby we focused on the most prevalent category of the CE events only. From the initially reported 19,308 CE events 73.5% were removed by a filter on the number of splice junction-spanning reads. Another 26.4% were excluded by removing overlapping events that could not be resolved. This resulted in a total of 1,380 statistically significant CE events with a false-discovery-rate (FDR) $\leq 0.01$. Depending on how SRSF6 regulates a cassette exon, increased or decreased exon inclusion might be the observable result. In order to remove exons that do not show a clear directional trend towards either skipping or inclusion, CE events with changes smaller than 5% were excluded and treated as unchanged (figure 3.21 B). The *SRSF6* KD mainly leads to an increase in exon skipping, with 975 cassette exons showing a significantly decreased inclusion level, compared to 237 cassette exons with an increased inclusion level. Thus, our results confirmed the findings of the alternative splicing analysis previously conducted by our colleagues, which indicated that the *SRSF6* KD mainly leads to cassette exon skipping (Juan-Mateu et al., 2018).

Next, we overlaid the computed AS changes with the differential expression results computed above. From all significantly differentially expressed genes with a $log_2$ fold-change $> 2$ only 31 also harbored a CE event and only five out of these were also significant (figure 3.22 A). When further directly comparing DGE fold-changes with AS PSI values, no relation could be observed (figure 3.22 B). PSI values appeared to be independent of the underlying transcript level change and vice versa. For instance, genes with exons displaying the highest absolute inclusion level changes exhibited a near-zero fold-change. With the exception of one gene, the majority of strongly regulated CE events did not reside in genes with equally strong affected expression levels. This suggests that *SRSF6* KD does not lead to an increased splicing of for example poison cassette exons, which is a common mechanism to regulate gene expression through alternative splicing via NMD (Kurosaki and Maquat, 2016). On the other side, SRSF4 showed an expression increase (LFC = 0.49), while having an alternative exon that is more consequently spliced out ($\Delta PSI =$ -0.01). This hints towards a possible compensatory effect, which is known among the SR proteins (Müller-McNicoll et al., 2016). Among all genes significantly affected by differential expression, many other splicing regulators were detected (figure 3.22 C). Some of these also showed effects in AS, predominantly increased cassette exon skipping. We again observed SRSF4 among these which further strengthen the idea of a stronger compensatory mechanisms within the SR protein family.

In summary, differential gene expression analysis revealed widespread global changes in transcript levels. These however were minor in magnitude, indicating possible compensatory effects, for example by other SR proteins such as SRSF4. AS analysis revealed a preference for CE skipping over inclusion upon the *SRSF6* KD. The overlap of both types of analysis pointed away from the idea that SRSF6 directly induced drastic transcript level changes via alternative splicing. However, many different splicing regulators were affected, which points towards the fact that SRSF6 might act as a major splicing regulator in human pancreatic beta-cells and influences the transcriptional response.

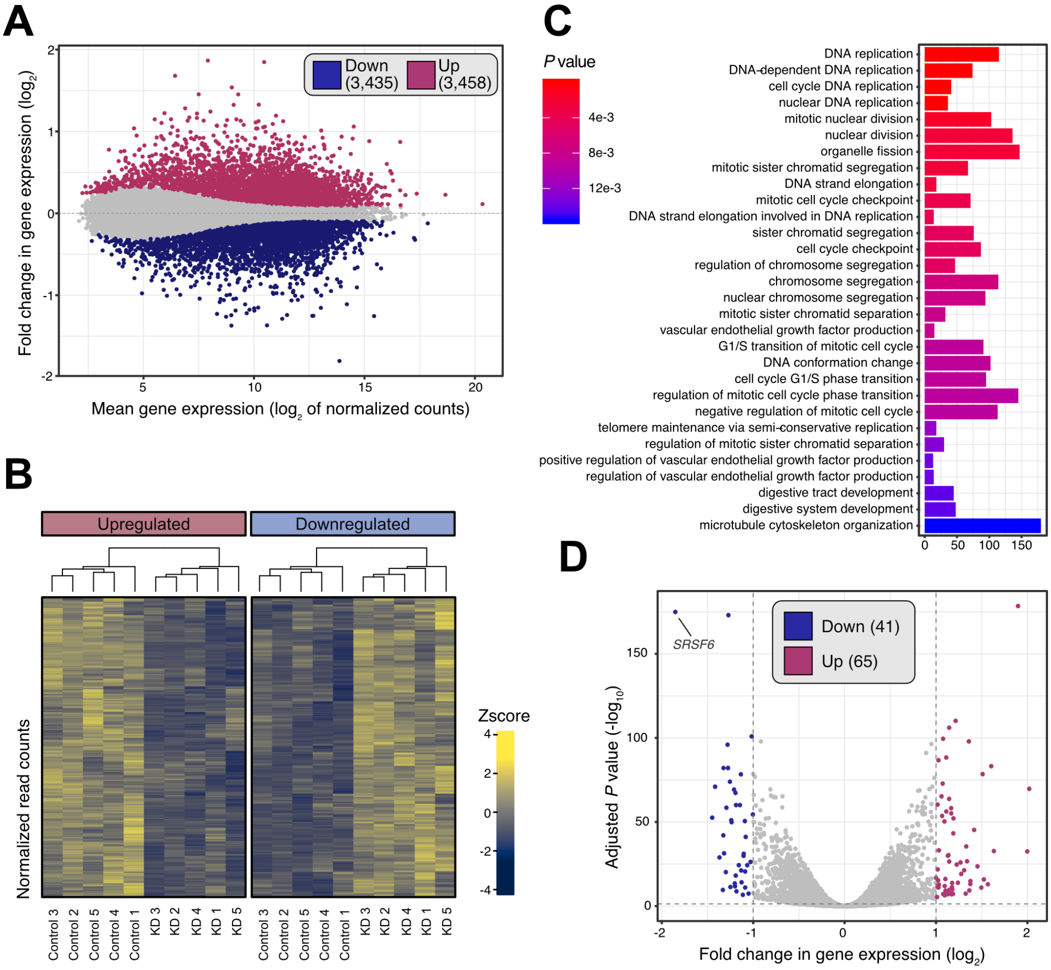

**Figure 3.20: SRSF6 modulates global transcript level expression.** Differential expression result comparing the effect of the *SRSF6* KD against the control using DESeq2 (Love et al., 2014). (A) MA plot showing the estimated $log_2$ fold-changes (LFC) against the expression level for each gene. Significant changes are colored (Benjamini-Hochberg adjusted $P$ value $< 0.05$) and LFCs are corrected for the influence of lowly expressed genes. (B) Heatmap of all significantly regulated genes. Counts for each replicate are rlog transformed and normalized by a Z-score for plotting. The heatmap is split in two groups using k-means clustering with two centroids. (C) GO enrichment analysis for genes that significantly changed their expression levels upon the *SRSF6* KD. Gene counts indicate the number of genes in the respective set and $P$ values from the hypergeometric distribution are shown. (D) Volcano plot that displays the shrunken LFCs over the adjusted $P$ values. An absolute threshold of 1 ($|LFC| > 1$) on the fold-change is indicated by the two vertical dashed lines. The adjusted $P$ value cutoff of 0.05 is indicated by the horizontal dashed line.

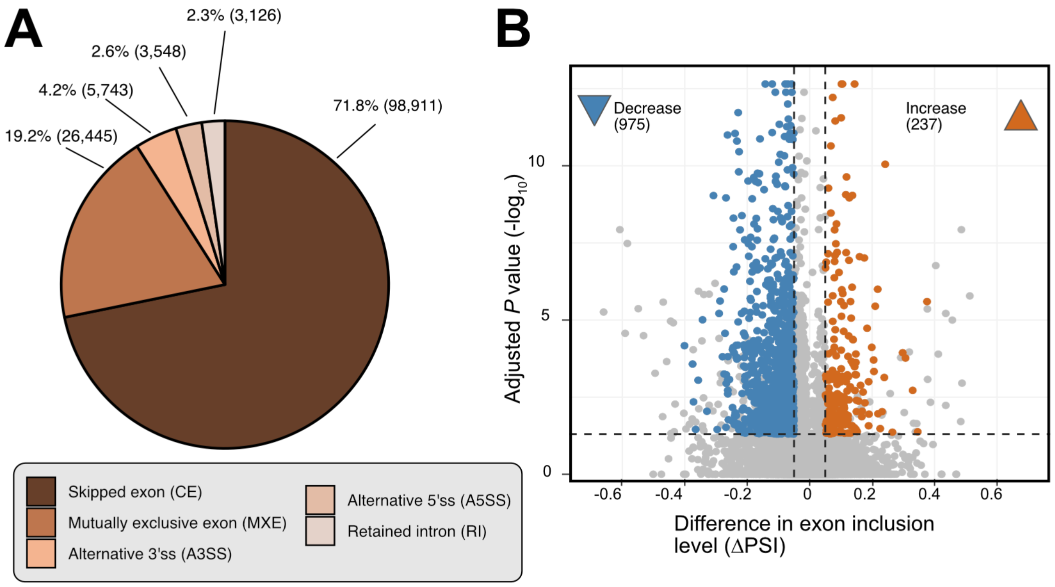

**Figure 3.21: *SRSF6* KD mainly affects cassette exons.** Alternative splicing analysis using rMATS (Shen et al., 2014) reveals an influence of the *SRSF6* KD in the splicing of alternative cassette exons. (A) Pie chart showing the distribution of the detected alternative splice events. (B) Volcano plot like illustration of the alternatively spliced cassette exon events. The difference in 'percent spliced-in' ($\Delta PSI$) is shown against the $P$ value (Bejamini Hochberg corrected). The dashed lines represent the thresholds for the $P$ value (0.05) and the $\Delta PSI$ (5%). The 1,212 significant splice events are colored according to the direction of the observed change.

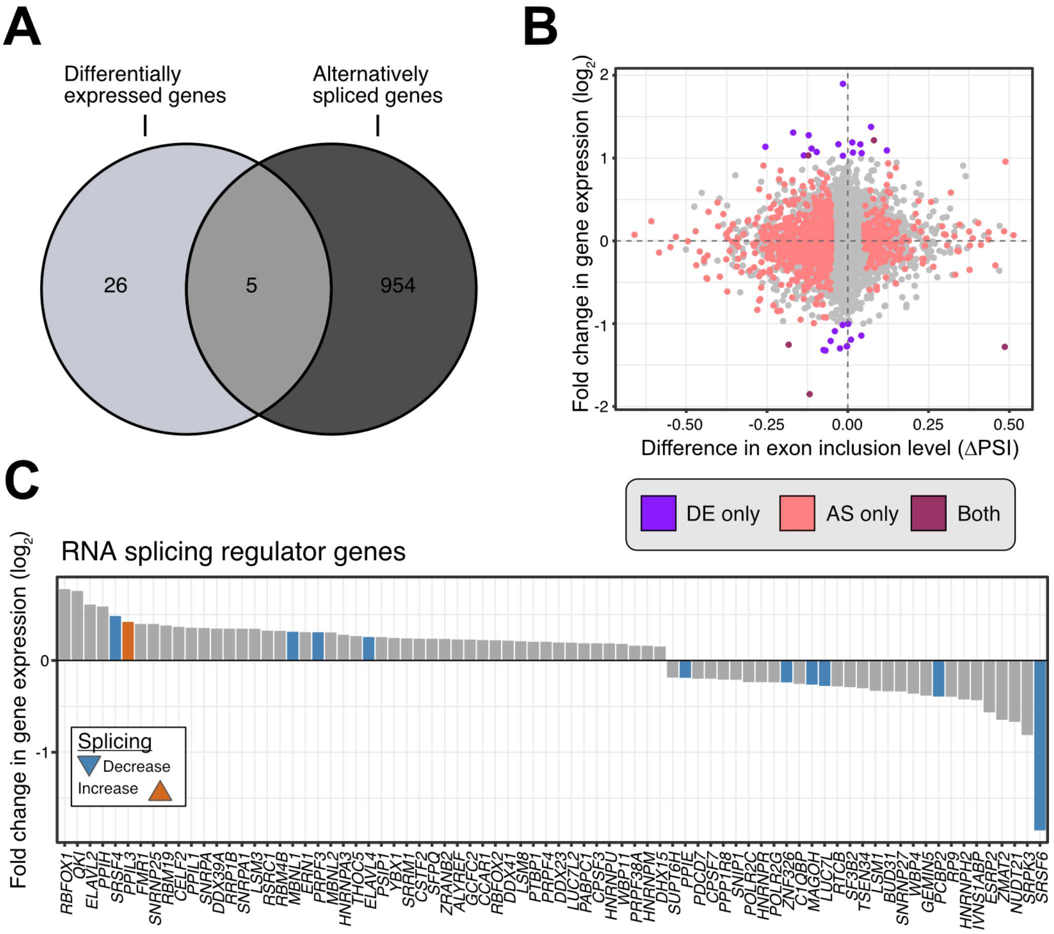

**Figure 3.22: SRSF6-regulated splicing events affect gene expression.** (A) Venn diagram showing the overlap between alternatively spliced cassette exons in differentially expressed genes. (B) Scatterplot that compares the differences in exon inclusion (percent spliced-in, PSI) to the gene expression change ($log_2$ fold-change, LFC). Genes are colored based on significance in the differential expression (DE) and/ or the alternative splicing (AS). In the case of multiple AS events per gene, the most significant one was selected as representative. (C) Many splicing regulator genes are affected by the *SRSF6* KD. The bar chart shows the LFC for selected splicing regulator genes, which are additionally color-coded by the AS change.

### 3.2.4 Alternative splicing is coupled to SRSF6 positioning on cassette exons

After analyzing the general splicing response of the *SRSF6* KD, we further asked how the position of SRSF6 impacts the observed splicing decisions. In order to understand how exactly SRSF6 influences CE changes we integrated our previously defined binding sites with the AS results. To detect CE events that are direct targets of SRSF6 binding, we overlapped the binding sites with specific CE models. We defined these models to range from the beginning of the upstream exon, to the end of the downstream exon (figure 3.23 A). We found that 765 (63%) cassette exon evens were overlapping with a least one SRSF6 binding site (figure 3.23 B). Here we again observed that with 593 CE events, the majority belonged to the decreased inclusion category, whereas in contrast 172 CE events belonged to the increased inclusion category. This overlapping approached allowed us to define a set of CE events that are direct targets to SRSF6.

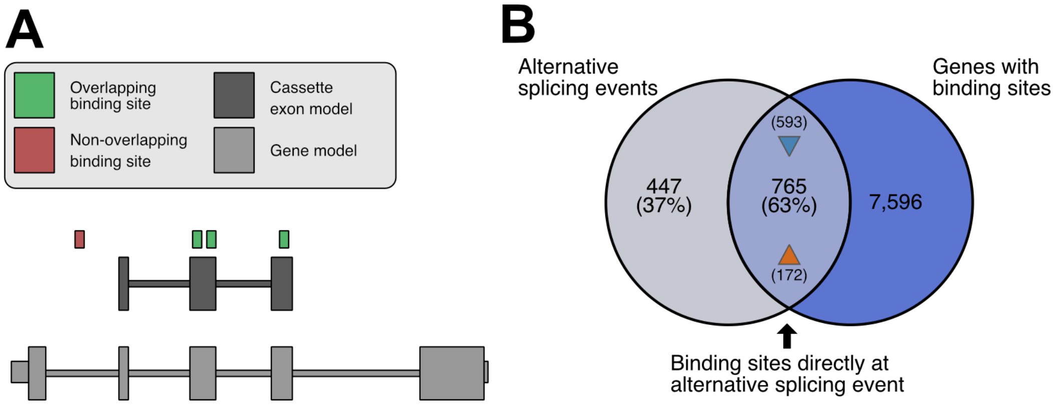

**Figure 3.23: SRSF6-regulated cassette exons associate directly with SRSF6 binding sites.** (A) Scheme depicting the overlap of a gene and cassette exon (CE) model with SRSF6 binding sites. Gene models are based on GENCODE-derived transcript annotations, whereas a CE model consists of the cassette exon, together with the flanking constitutive exons. (B) Venn diagram showing the overlap of CE events with genes that harbor a SRSF6 binding site. The numbers in brackets at arrowheads specify exons with significantly decreased or increased inclusion.

One gene falling into this set encodes for the SR protein SRSF4. It serves as a good candidate to exemplarily highlight the combined regulatory events we observed so far. *SRSF4* showed a large number of SRSF6 binding sites, with a preference of CDS over introns (figure 3.24). It also is well expressed, since high coverage was observed in both RNA-Seq conditions. The annotated SRSF4 transcripts further showed two predominate isoforms that seemed to be regulated by the *SRSF6* KD, one being the regular protein whereas the other includes a poison cassette exon. It is well known that SR proteins regulate their transcript levels by including exons harboring premature stop codons, thus triggering degradation via NMD (Kurosaki and Maquat, 2016). In the present case the *SRSF6* KD led to increased CE skipping, potentially indicating compensatory effects. The cell might react to low SRSF6 levels by an increased expression of the functional isoform of SRSF4. This example perfectly showcased the complex interplay between direct binding and transcriptional changes in the context of SR proteins.

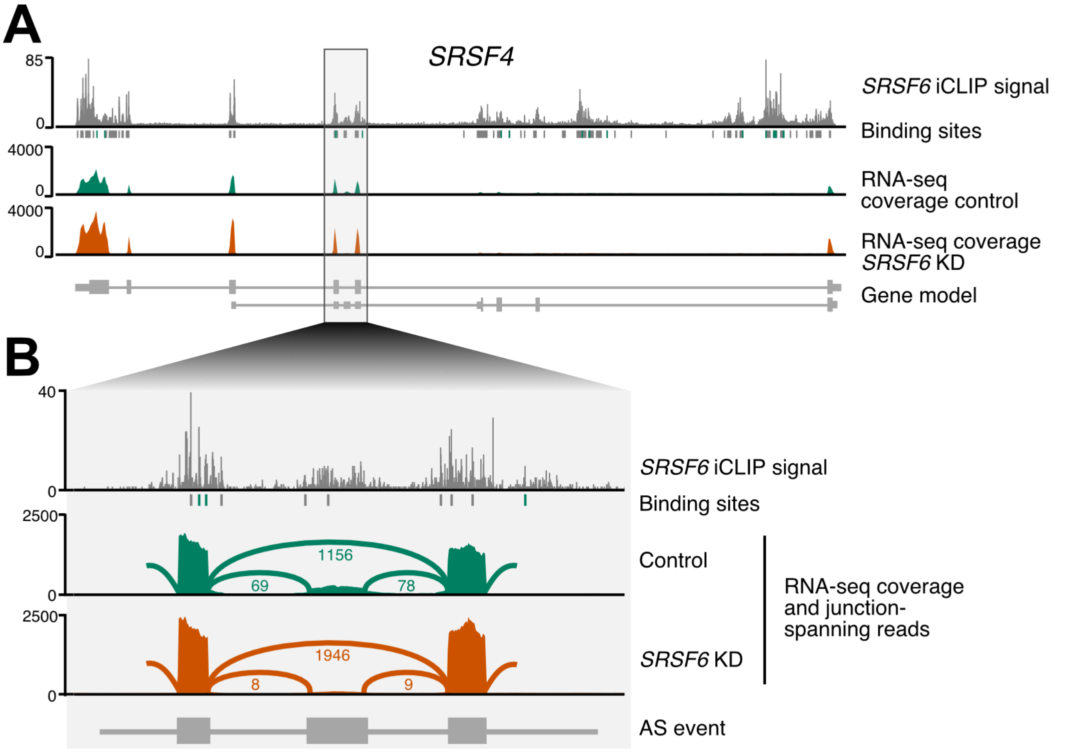

**Figure 3.24: SRSF4 alternative splicing is regulated by the *SRSF6* KD.** The SR protein-encoding gene *SRSF4* is significantly upregulated upon the *SRSF6* KD. (A) Genome browser shot that shows the SRSF6 iCLIP crosslink events, with computed binding sites and the coverage of the RNA-Seq over selected transcript annotations of *SRSF4*. (B) Zoom-in on the SRSF6-regulated cassette exon event with sashimi-plot representation of the splice junction spanning reads of the RNA-Seq coverage. Lines show exon-exon junctions and numbers indicate the supporting reads.

To further address the exact positioning of SRSF6 on all identified targets, metaprofiles such as RNA-maps are a state-of-the-art method. Typically, regular eukaryotic exons are either predominantly included or excluded given a certain cell status. This is showcased by the bimodal distribution of estimated PSI values over all analyzed exons (figure 3.25 A). The distribution also showed that most exons are included whereas fewer exons are excluded. This immediately points towards a common problem when integrating orthogonal data with alternatively spliced exons. Differences in the base inclusion level between exons must lead to differences in the observed iCLIP coverage. For instance, an exon with increased inclusion upon *SRSF6* KD must have shown less inclusion in the control and thus shows fewer iCLIP crosslink counts (figure 3.25 B). This was no minor effect in our data and could not be neglected. On average exons with an inclusion rate above 80% showed more than 150 crosslink events per exon. On the other side, less than 50 crosslink evens could be observed for exons with an inclusion rate below 20%. When remained unresolved, this bias makes it impossible to compare crosslink coverages between the increased and decreased inclusion set, as well as between any regulated set and a potential background of non-regulated exons.

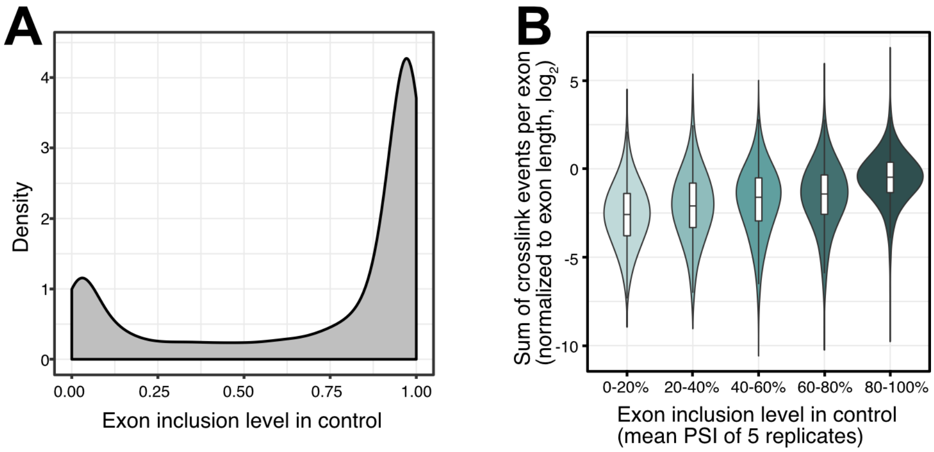

**Figure 3.25: Exons with high inclusion harbor more crosslink events.** (A) Alternative exons have different inclusion levels in control ('percent spliced-in', PSI). Density plot that shows the distribution of PSI values in control condition. (B) More iCLIP crosslinks are detected on exons with high inclusion levels, compared to exons with low inclusion in control. Box and violin plots show the distribution of crosslink events per exon ($log_2$ transformed, normalized to exon length). All 19,308 CE exon evens were stratified into 20% bins (color shading) by their PSI

We resolved this issue by compiling PSI-matched background sets for the exons with decreased inclusion, as well as for those with increased inclusion. It was important to compute two separate background sets, since the decreased inclusion exon set showed a completely different baseline exon inclusion than the increased inclusion set (figure 3.26 A). The universe of all non-regulated exons served as the total background set. We started adjusting the total background to a PSI-matched background, by splitting the inclusion level distribution of a regulated exon set into 5% quantiles. These quantiles were then transferred over to the total background set and an adjusted background set was picked randomly. The random picking step ensured that the number of exons included in the PSI-matched distribution was similar to the number of exons of the regulated exon set. We then repeated the matching and subsampling process 50 times, to avoid any local pitfalls (figure 3.26 B). The strong corrective effect of the PSI-matching becomes visible when compared to a uniform background selection strategy (figure 3.26 C, D). Regardless of the type of regulation set, increased or decrease, the uniform background set showed the same baseline inclusion distribution. The PSI-matched distribution on the other hand followed the distribution of the respective regulated set it got adjusted for.

Next, we counted single nucleotide positions covered with crosslink events on regulated and PSI-matched background exons by spanning a 200-nucleotide window around each cassette exons splice sites. The flanking up- and downstream exons were used as reference points, so the same window was opened (figure 3.27 A). This resulted in four windows, each of which centered around a splice site. For exons that were shorter than 200-nt the largest possible range was displayed. In these cases, the exons were split in half to ensure that the windows remained non-overlapping. For each window we calculated the relative crosslink frequency by counting the number of crosslink events per nucleotide and dividing them by the number of exons that covered the particular nucleotide. For the PSI-matched background set we applied the same counting scheme, but the mean and standard deviation over the 50 iterations were displayed as reference (figure 3.27 B). In general, we observed increased binding on the regulated exons compared to the flanking ones. This is in line with SRSF6 being a global splicing regulator. However, the observed patterns differed between decreased and increased inclusion exon sets. Exons that got decreased in their inclusion upon *SRSF6* KD showed increased binding to the cassette exon itself. This hints towards a direct enhancer function of SRSF6 on these exons. The opposing trend could be observed for increased inclusion exons. Here SRSF6 bound more dominantly to the flanking up- and downstream exons, compared to the cassette exons. This suggested that SRSF6 in these cases reinforced the flanking exons in order to facilitate exon skipping.

Meta-profiles are useful visualization methods, which in the present case were used to summarize crosslink signal over multiple exons. However, one has to be careful not to overestimate signal from for example exons of highly expressed transcripts, or exons with high-affinity binding sites. For instance, a single exon from a highly abundant transcript might distort the profile, leading to false discoveries. For that reason, we computed heatmaps to complement our meta-profile (figures 3.28, 3.29). Both trends observed in the meta-profile could be reproduced by the RNA-splicing heatmaps. Here, rather than summed up, each exon is displayed as an individual row and positions with crosslinks were highlighted. For increased and decreased inclusion level exon sets these maps indicated that the observed trends were supported by the signal from all exons, rather than being a bias forged by a minority.

In summary, these results suggest that SRSF6 modulates alternative splicing in a position-dependent manner. Increased binding to the up- and downstream flanking exons seemed to result in increased skipping of the regulated exon. Increased binding directly at the cassette exon seemed to stabilize the inclusion of that exon into the mature transcript. Thus, our results help to understand how direct SRSF6 binding impacts the splicing outcome. Further, this type of regulation was also described for other SR proteins, however the exact mechanism is still debatable (Han et al., 2011; Sanford et al., 2009).

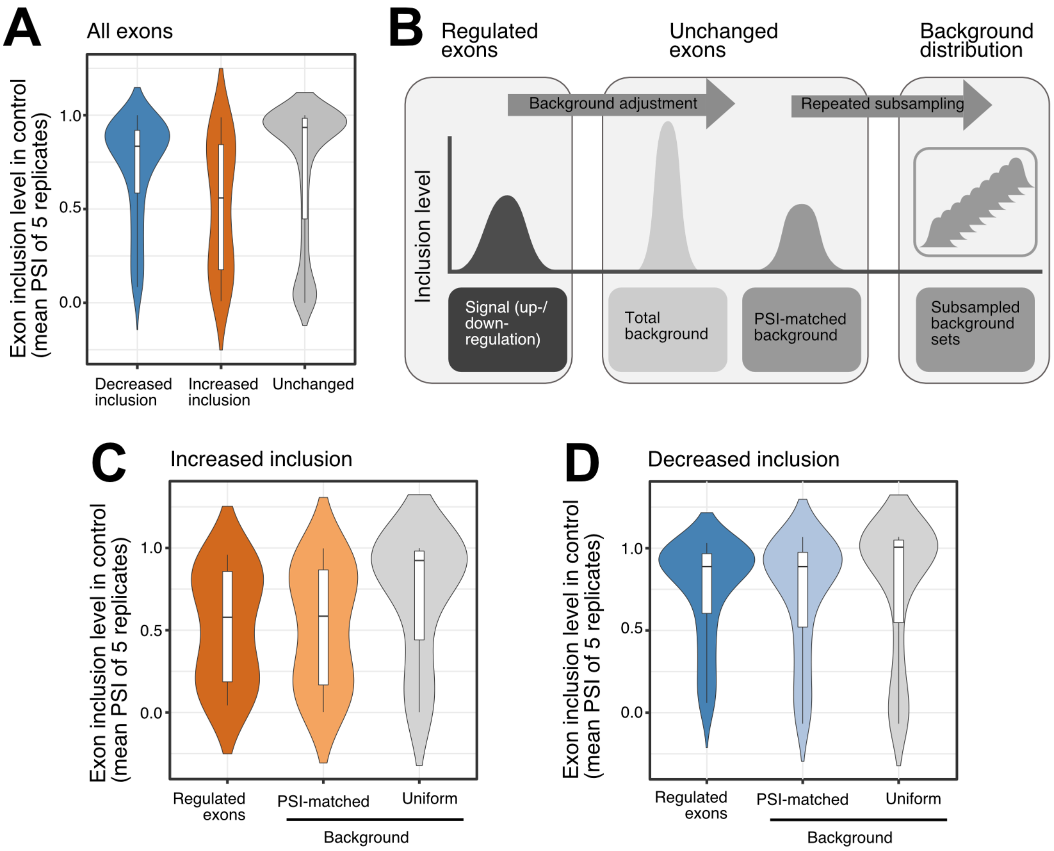

**Figure 3.26: Selection of PSI-matched background sets.** PSI-matched background sets show a similar mean inclusion level as up- and downregulated exons. (A) Inclusion levels (PSI) differ between exons that decrease, increase or remain unchanged. Box and violin plots show the distribution of PSI values in control for the three groups. (B) Schematic overview of the computation of the PSI-matched background from all unchanged exons. A random subset is picked and corrected to exhibit a similar mean inclusion level to the upregulated set (N=237) or to the downregulated set (N=975). This process was repeated 50 times to retrieve mean and standard deviations. (C) Exemplary distribution of a selected PSI-matched distribution compared to a uniform background set for exons with increased inclusion levels. (D) Exemplary distribution of a selected PSI-matched distribution compared to a uniform background set for exons with decreased inclusion levels.

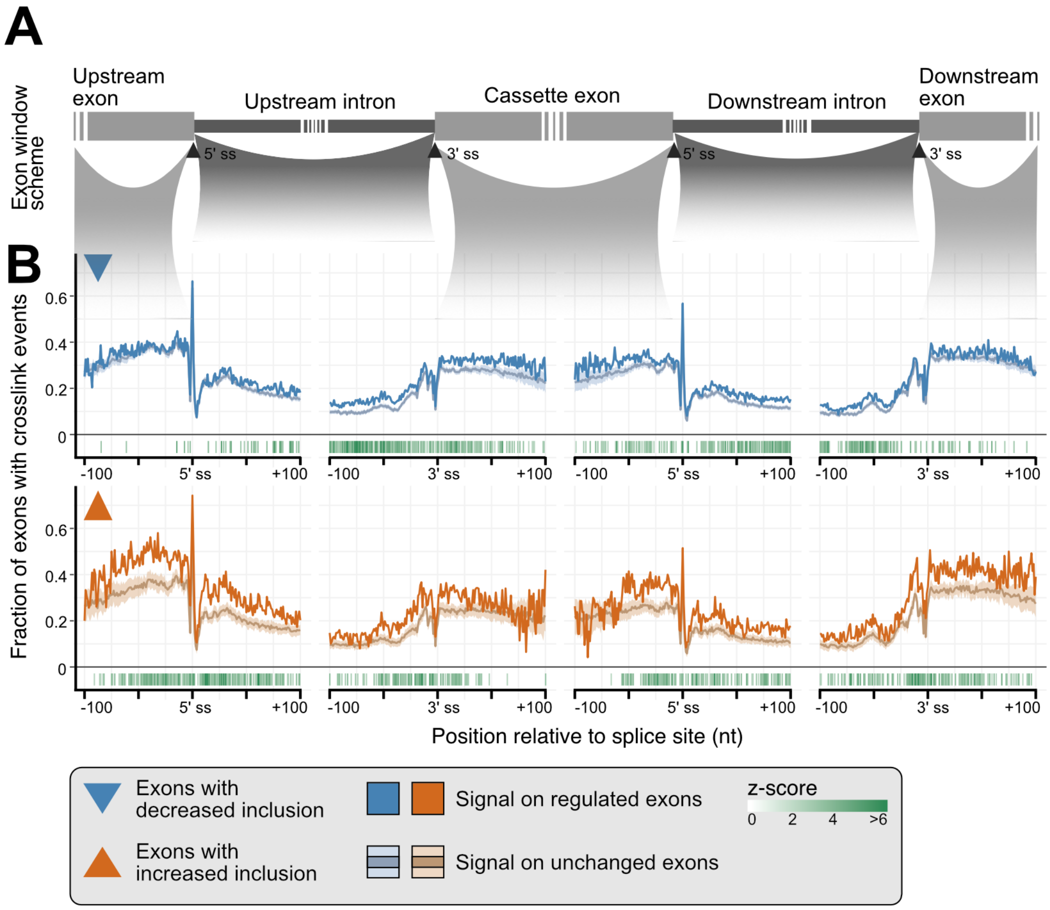

**Figure 3.27: SRSF6 RNA splicing maps.** (A) Scheme of the four selected windows in which crosslink events are counted. The windows are always centered around the indicated 3' or 5' splice site. (B) SRSF6 shows more crosslink events on alternative exons with decreased inclusion upon *SRSF6* KD (blue), compared to exons with increased inclusion. On exons with increased inclusion the flanking constitutive exons display more crosslink events (orange). The metaprofiles display the fraction of exons with a crosslink event at a given position in the indicated window. For each comparison mean and standard deviation of the respective PSI-matched background set is shown (light blue, light orange). Positions on which the signal for the regulated exons differs significantly from the respective background set are shown in green (adjusted $P$ value $< 0.05$). These positions are further color-coded by their z-score, so that darker values indicate stronger changes.

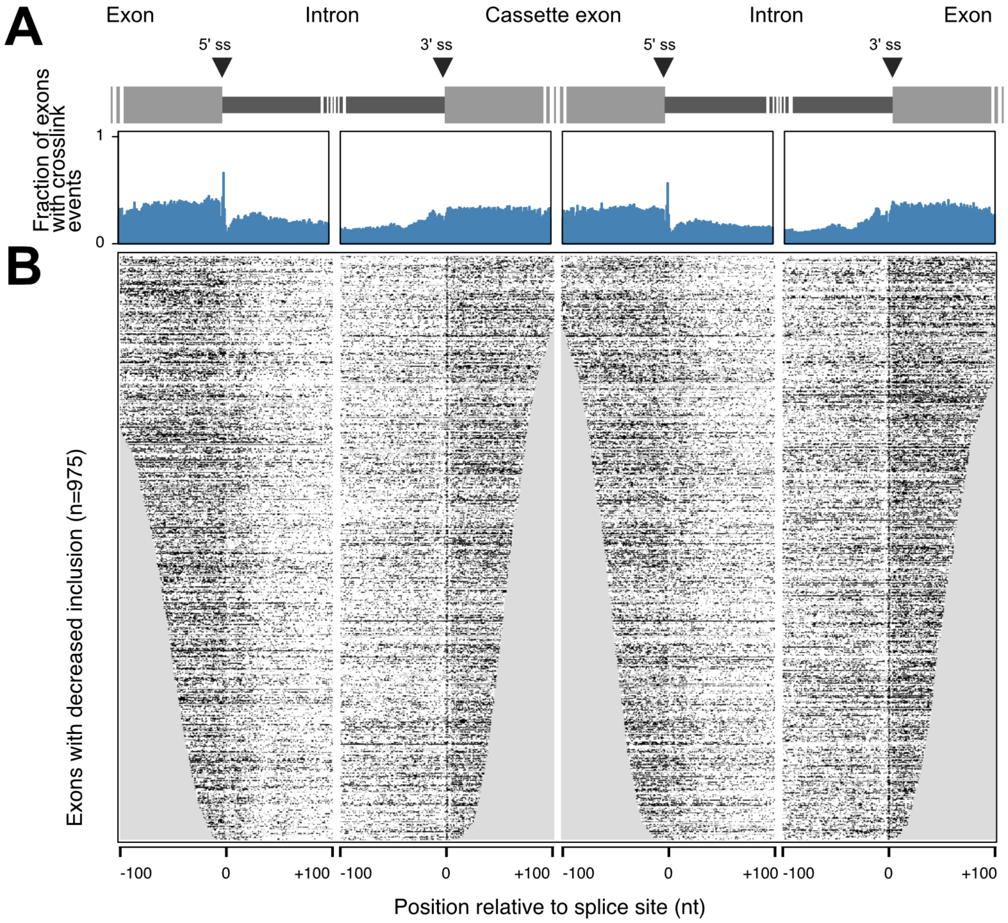

**Figure 3.28: SRSF6 crosslink events around downregulated cassette exons.** (A) Metaprofile that indicates the summed-up fraction of exons with crosslink events at a given position for 100-nt on either side of the indicated splice site (scheme shown on top). (B) Heatmap that shows the fraction of exons with crosslink events per position. Rows are sorted by length and the number of crosslink events. Grey positions indicate missing values because of exon length. These positions were set to 0 in the heatmap and thus are included in the metaprofile above.

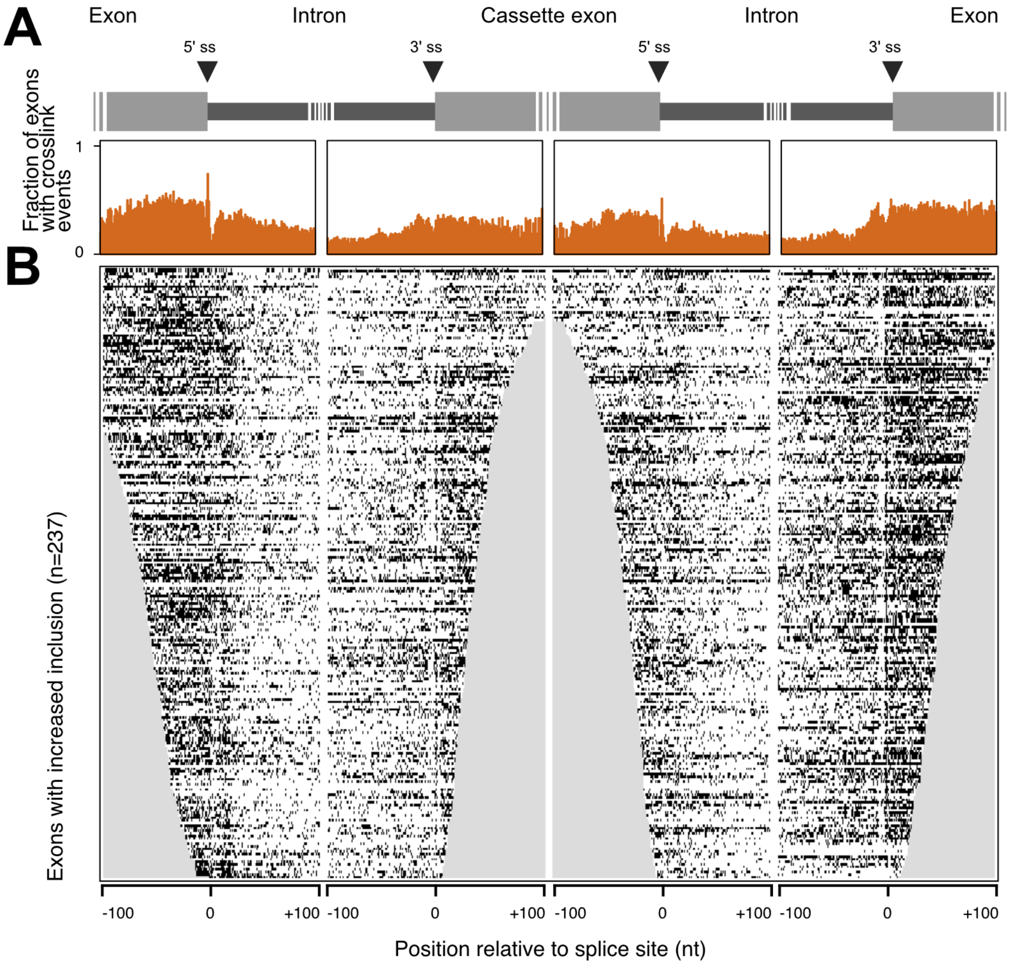

**Figure 3.29: SRSF6 crosslink events around upregulated cassette exons.** (A) Metaprofile that indicates the summed-up fraction of exons with crosslink events at a given position for 100-nt on either side of the indicated splice site (scheme shown on top). (B) Heatmap that shows the fraction of exons with crosslink events per position. Rows are sorted by length and the number of crosslink events. Grey positions indicate missing values because of exon length. These positions were set to 0 in the heatmap and thus are included in the metaprofile above.

### 3.2.5 SRSF6 regulates several susceptibility genes for type 1 and type 2 diabetes

In diabetes, pancreatic beta-cells lose their insulin-producing capacity, resulting in hyperglycemia and long-term complications for the patient. A major susceptibility gene for type 1 and type 2 diabetes is *GLIS3*. Interestingly, its downregulation decreases the expression of SRSF6. It was further found that SRSF6 itself regulates many genes involved in beta-cell function, such as insulin secretion (Juan-Mateu et al., 2018). This suggested a link between SRSF6 in its role as a key splicing regulator and the disease. In the above part we further refined this link by describing how SRSF6 regulates beta-cell function through direct binding. We next looked for specific diabetes susceptibility genes that are targeted by the characterized mechanism. A compiled list of T1D and T2D susceptibility genes was compiled from ImmunoBase and GWAS catalog. This list was first overlaid with the list of all direct SRSF6 targets identified above. We further narrowed the list of potential candidate genes down and asked for only those genes that harbored the SRSF6 binding site within the region of the cassette exon event (figure 3.30 A, B). We identified a total of 6 cassette exon events targeted by direct SRSF6 binding for the T1D and 22 events for the T2D susceptibility genes originating from five and 17 genes, respectively. A direct comparison of the binding site strength and the difference in the exon inclusion level change revealed no direct correlation (figure 3.30 C, D). For example, genes with a high or low $\Delta PSI$ value showed the same broad range of binding site strength scores. However, some genes might be subjected to a stronger SRSF6 dependent regulation, since the binding sites overlapping the CE events were supported by the identified binding motif. Among these was the cell-cycle regulators centromere protein O (*CENPO*) and the integrin subunit beta 3 binding protein (*ITGB3BP*) which both interact with the histone complexes (Foltz et al., 2006; Shattil et al., 1995). The cyclin dependent kinase 2 (*CDK2*) gene encodes for a signaling kinase which also participates in cell cycle regulation and its malfunction has been shown to be involved in caner (Chung and Bunz, 2010). A susceptibility gene for both T1D and T2D is breast cancer anti-estrogen resistance protein 1 (*BCAR1*) which is not only involved in cell-cycle regulation and cancer, but also controls cell apoptosis (Brinkman et al., 2000; Rufanova et al., 2009). These finding again confirm the essential role of SRSF6-based AS regulation in human beta-cells. SRSF6 directly effects multiple diabetes susceptibility genes and most of these genes seem to control cell-cycle and apoptosis. This further links SRSF6-mediated regulation to beta-cell death as it is the cases in for example T1D.

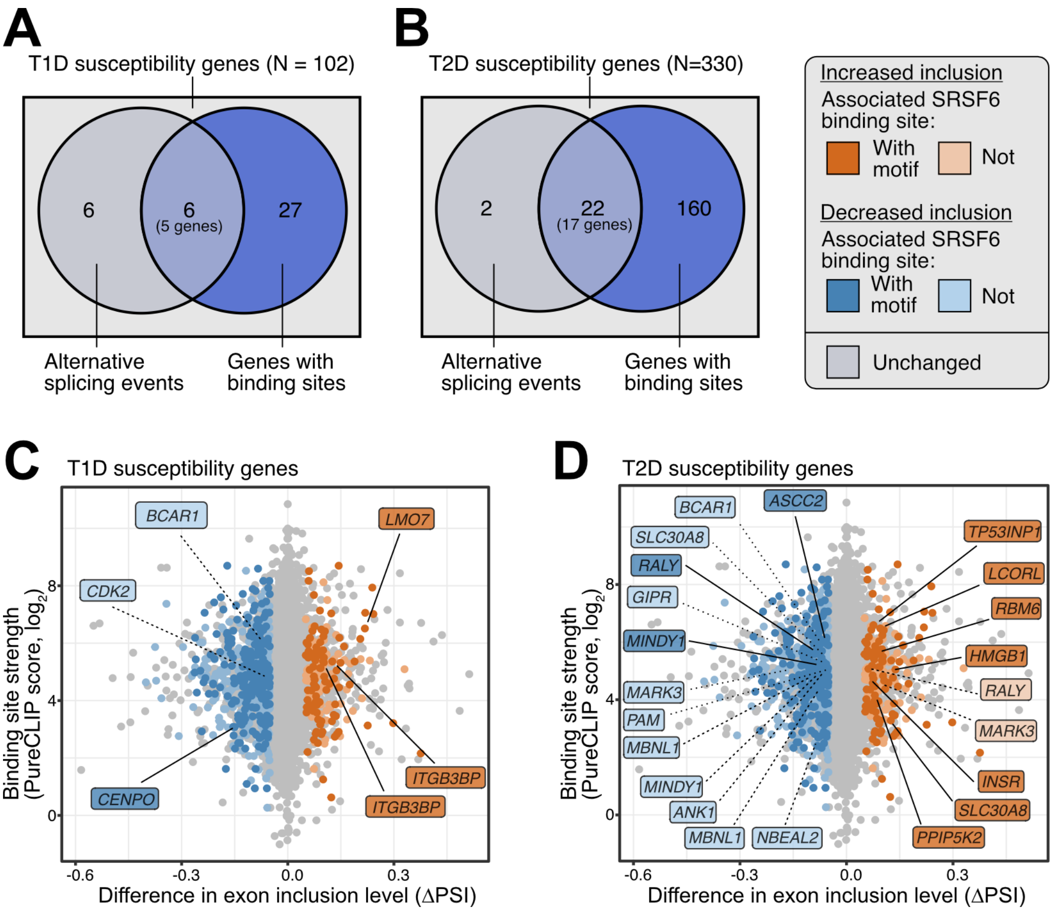

**Figure 3.30: SRSF6 directly binds to diabetes susceptibility genes.** SRSF6 regulates 6 and 22 splicing events in 5 T1D and 17 T2D susceptibility genes. (A, B) Venn diagrams show the overlap of alternative splicing (AS) events and genes with binding sites for T1D and T2D, respectively. (C, D) Scatterplots show the inclusion level difference ($\Delta PSI$) against the binding strength of the strongest binding site ($log_2$ transformed PureCLIP score) associated with the indicated gene. Genes with exons that show significantly increased of decreased inclusion values are color-coded in blue and orange, respectively. Darker shades indicate the presents of the GAA-rich binding motif directly at the AS event.

To validate the predicted changes, our colleagues from the Eizirik group used semi-quantitative RT-PCR to test for the AS changes. They performed independent experiments for nine susceptibility genes in EndoC-$\beta$H1 cells under control and *SRSF6* KD conditions (figure 3.31). They could validate the predicted changes for all T1D susceptibility genes in which we found the cassette exon event being a direct binding target of SRSF6. For *CENPO* and *ITGB3BP* the observed splicing changes were particularly strong, further strengthening the cell-cycle related influence of the SRSF6-controlled genes. With the RNA-binding motif protein 6 (RBM6) and the StAR related lipid transfer domain containing 10 (*STARD10*) they also validated two T2D susceptibility genes, which are involved in caner and cell metabolism regulation (Timmer et al., 1999; Scanlan et al., 1998).

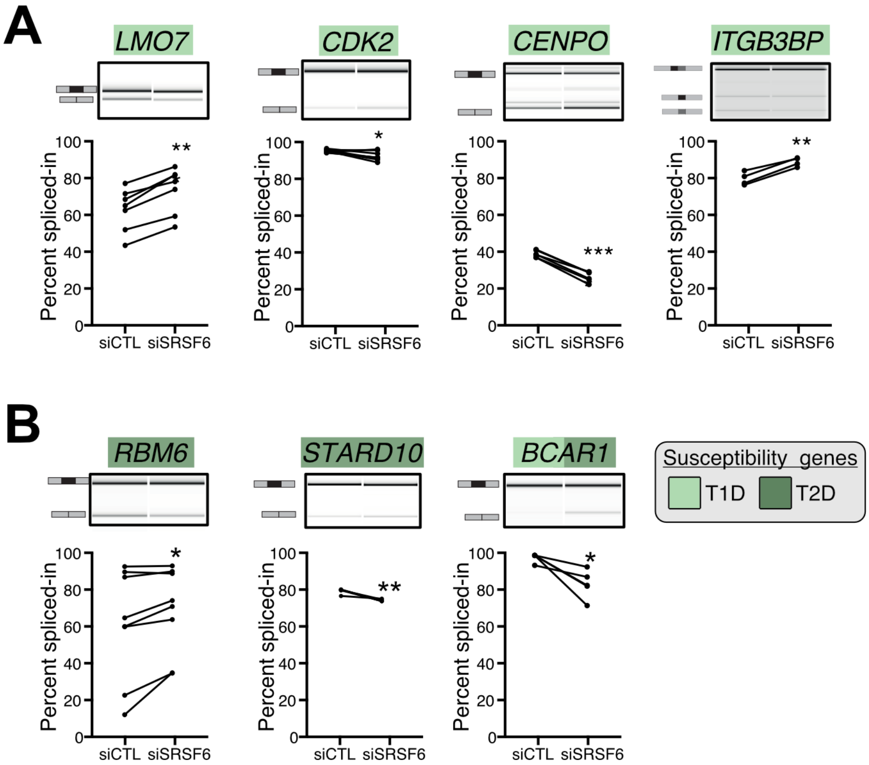

**Figure 3.31: Validated splicing changes in diabetes susceptibility genes.** Splicing changes in diabetes susceptibility genes were validated by semi-quantitative RT-PCR. Gel images are shown on top and paired data points for the quantification are shown blow with significance indicated by paired Students t-test (* $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$). (A) Validations of SRSF6-regulated exons in T1D. (B) Validations for SRSF6-regulated exons in T2D.

We additionally investigated how T1D and T2D susceptibility genes were affected in gene expression following the *SRSF6* KD (figure 3.32 A, B). Alongside the susceptibility genes described above, several other genes showed small but noticeable effects in gene expression. This might not be the driving force behind beta-cell reactions to SRSF6 levels but it indicated that transcript level changes were indeed present. This further highlights the complexity by which SRSF6 affects a plethora of different susceptibility genes. For example, most genes are direct binding targets to SRSF6, but the induced reaction to the transcriptome can be via differential expression, alternative splicing or a combination of both. On top of that, also several indirect effects might exit. For example, some SRSF6 regulated susceptibility genes might regulate additional susceptibility genes themselves. Such effect could be shown by the computation of protein-protein interaction networks (figure 3.32 C). The network identified for example, similar to *CDK2*, additional signaling kinases involved in cell-cycle regulations, like the cyclin dependent kinase inhibitor 1B (*CDKN1B*) and cyclin D1 (*CCND1*) (Goode et al., 2009). This suggests the presence of an even larger functional network of proteins that could be affected by SRSF6.

In summary, we showed that the splicing regulator SRSF6 has thousands of direct targets throughout the transcriptome in the context of pancreatic beta-cells. The strongest of these binding sites displayed a GAA-rich consensus binding motif and showed a specific and highly regulated binding mechanism to tightly control transcriptome changes. These changes are mostly by alternative splicing, with a preference for cassette exon skipping upon *SRSF6* KD. We furthermore saw that increased skipping of these exons was modulated by SRSF6 reinforcing neighboring up- and downstream exons. On exons with increased inclusion upon the KD SRSF6 normally binds directly to the cassette exon, resulting in a stabilizing effect. This mechanism seemed to be involved in the diabetic context, since several T1D and T2D susceptibility genes were regulated by the described system. In total this suggests that the GLIS3-regulated splicing regulator SRSF6 plays a key role in beta-cell function and thus contributes to the diabetes disease.

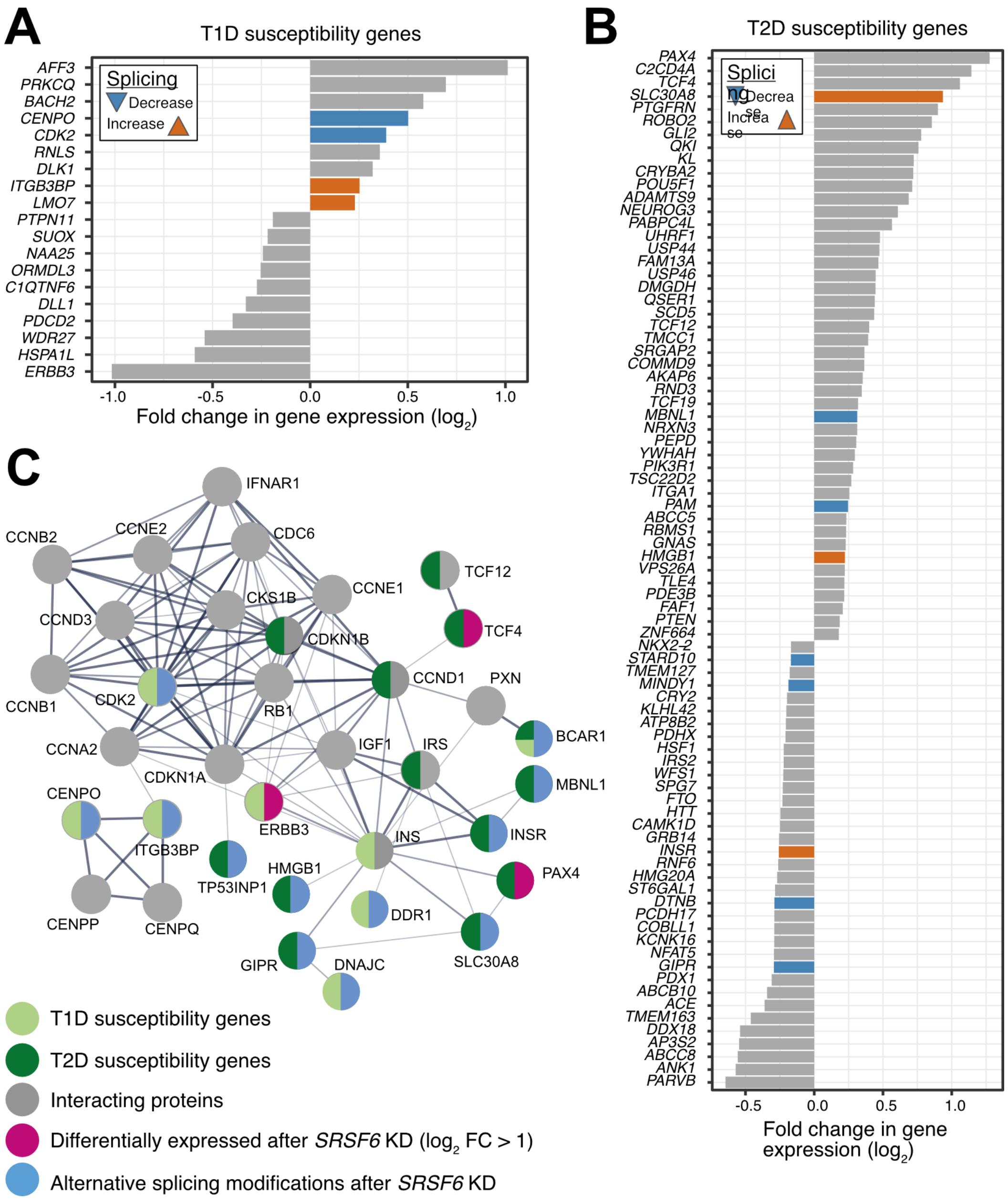

**Figure 3.32: SRSF6 impacts the expression of diabetes susceptibility genes.** (A, B) The expression of nine T1D (A) and 31 T2D (B) is affected by the *SRSF6* KD in EndoC-$\beta$H1 cells. Genes also affected by alternative splicing are color coded. The bar charts show the change in gene expression ($log_2$-transformed fold-change). (C) Protein-protein interaction network (PPI) for proteins encoded by SRSF6-regulated diabetes susceptibility genes. The PPI network was obtained using STRING, with all 29 T1D and T2D susceptibility genes as input which were significant in either the alternative splicing or the differential expression analysis.

# 4 | Discussion

The mRNA life cycle is a highly regulated and controlled process that is shaped to a large extend by the class of RNA-binding proteins (RBPs). They are involved in each phase of the mRNA life to such an extent that the mRNA is never seen alone (Müller-McNicoll and Neugebauer, 2013). A plethora of RBPs are coating the mRNA in the cytosol and nucleus forming large messenger ribonucleoprotein particles (mRNPs). This ranges from pre-mRNA processing, where for example SR and hnRNP proteins are both involved in splicing (Shi, 2017), to mRNA packaging for nuclear export (Singh et al., 2012) and finally degradation (Bicknell and Ricci, 2017). Such processes are indeed highly intertwined, since most RBPs interact with each other or participate in multiple regulatory processes. The cap-binding complex CBC for example is involved in transcription and splicing as well as translation. It binds to the 5'-cap structure during transcription but also stays bound to the mRNA and travels to the nuclear pore (Gonatopoulos-Pournatzis and Cowling, 2014; Narita et al., 2007). All of these processes are further controlled by a range of external and internal stimuli, adding to the highly dynamic form of the mRNA life (Zarnack et al., 2020).

In the last decade iCLIP has been established as a state-of-the-art experimental method to study the protein-RNA interaction of RBPs at single nucleotide resolution on a high-throughput scale (König et al., 2010). A very recent update described key optimizations of the protocol increasing the complexity of the produced sequencing libraries (Buchbender et al., 2020). A wide range of RBPs and complexes were studied using iCLIP, leading to very different research questions on different scales. For example, *in vitro* iCLIP experiments were used to the specific regulation of U2AF65 based on other RBPs (Sutandy et al., 2018). On the other side large multi-RBP complexes like the spliceosomal assembly were monitored as well (Briese et al., 2019). However, such studies usually feature a set of custom-tailored solutions on the computational processing of the iCLIP-derived sequencing reads. This ranges from a variety of different quality filtering standards, to the use of different peak calling tools and downstream postprocessing steps. To meet these needs, several processing pipelines exits for different derivates of CLIP-seq protocols (Bottini et al., 2018; Uhl et al., 2017). However, specific guidance for iCLIP analysis are missing. This highlighted the demand for a standardized iCLIP processing workflow to ensure reliable binding site detection, which also leads to better data comparability between groups and RBPs. This would ultimately result in a better understanding of complex regulatory processes with many RBPs involved.

In this thesis I described the process of computational iCLIP data processing step-by-step, giving a detailed workflow to generate highly reproducible results (Busch et al., 2020). In particular I described a way to boost the peak calling step with multiple replicates, while maintaining a high sensitivity and specificity. Upon this, binding sites were defined in uniform width so that downstream processing is simplified. This led to detailed descriptions of the binding spectrum of two splicing regulator proteins. First U2AF65 as part of the core spliceosome machinery was used to exemplify the different processing steps and its known binding properties were recaptured (Zarnack et al., 2013). Next, I characterized the so far unknown binding spectrum of SRSF6 in human pancreatic beta-cells (Alvelos et al., 2020). The identified binding motif matched a recent description in mouse cells (Müller-McNicoll et al., 2016). Further, I showed how the exact positioning of SRSF6 directly influences the splicing of cassette exons within known diabetes susceptibility genes. This ultimately revealed how SRSF6 shapes transcriptome changes and contributes to beta-cell survival and death.

## A pipeline for iCLIP data processing

A common problem to peak calling tools is that resulting binding sites differ in width. This makes it challenging to compare binding sites between multiple different RBPs and complicates downstream analysis that requires binding sites to be harmonized. I developed an iterative merging scheme that allows the combination of multiple PureCLIP-derived crosslink sites into binding sites of any desired width (Figure 3.4). One advantage of this strategy is that little prior knowledge of the binding site width is required, which is not the case for most other approaches (Chakrabarti et al., 2018). To run PARalyzer for example, one has to decide on a bandwidth for the kernel density estimation, which affects the width of the returned binding cluster (Corcoran et al., 2011). In contrast to these tools, I showed how the appropriate binding site width can be deduced from the crosslink event coverage itself (Figure 3.3). Such a unified width definition positively impacts downstream analyses like motif definition and the integration of orthogonal data such as gene annotations. Nevertheless, one has to be aware that this serves as an approximation to detect the most prevalent trend in the binding profile of the RBP under study. The same RBP might show different binding behaviors even within the same cell type, depending on for example the presence of additional cofactors. In the case of U2AF65 it was shown that the recruitment to the splice site is heavily depending on additional RBPs, such as FUBP1, PTBP1, CELF6 and PCBP1. Each of these cofactors affects U2AF65 binding and thus alters the landscape of possible binding sites, ultimately affecting the splicing outcome (Sutandy et al., 2018; Warf et al., 2009; Tavanez et al., 2012). Here a 9-nt window was found appropriate for U2AF65, given the computed crosslink event profiles (Figure 3.4). This also fitted to the known binding mode of the protein, which binds the polypyrimidine tract with a tandem of two RRM domains, each recognizing four nucleotides (Sickmier et al., 2006; Mackereth et al., 2011). It also nicely captured the preferred affinity of U2AF65 to intronic sequences, which has already been shown (Zarnack et al., 2013; Shao et al., 2014).

Most peak calling tools for CLIP data so far make use of only a single biological replicate. Only recently PureCLIP was extended to account for up to two biological replicates (available from version 1.3.0), which is still not enough for reliable iCLIP-based studies. Here I showed that the sensitivity of the peak calling can be enhanced by merging multiple replicate libraries prior to the peak calling (Figure 3.2). In a recent publication for example, I described the binding profile of MRKN1 using three iCLIP replicates, highlighting the need for an appropriate integration (Hildebrandt et al., 2019). Generating higher replicate numbers is common nowadays, since the iCLIP protocol improved to an extend that allows the straightforward generation of biological replicates (Haberman et al., 2017). In our study to describe the binding profile of SRSF6 as well as in the present case of U2AF65, even four biological replicates were used (Figures 3.6, 3.9). Merging replicates however, if not handled properly causes less specific crosslink sites to be called. By adding up signal across replicates one might, by chance create an artificial culmination of crosslink events. I explained how this bias can be controlled with the use of replicate-specific thresholds (Figure 3.6). The library size of each replicate was accounted for by computing an individual quantile-based cutoff. In the case of U2AF65 only 17.5% of the merged binding sites were found to not be reproducible by at least two replicates, leading to a total of 248,916 reproducible binding sites.

As an alternative to this, binding sites could be called individually on each replicate and merged again after the peak calling step in a split-and-combine fashion. This approach was used in a recent publication describing the binding profile of STRAP from eCLIP data with four biological replicates (Jin et al., 2020). On the other side replicate merging might not be desired if replicates differ tremendously from each other. This might be due to technical biases, but also due to the experimental design when for example an RBP is studied in different conditions. Despite this, one great advantage of binding site merging is the enhanced sensitivity for either low affinity binding sites or binding sites on lowly abundant transcripts. Since iCLIP always represents a snapshot of the cell, certainly not all genes are expressed and thus not all possible binding sites are detectable (Signor and Nuzhdin, 2018). However, with the increased sensitivity a higher proportion of the actual spectrum can be observed. This for example positively impacts approaches that try to predict binding sites on not expressed transcripts. Such tools take a given set of binding sites as input and extract sequential and structural features to make their predictions (Maticzka et al., 2014; Ghanbari and Ohler, 2020). In cases where an RBP performs multiple functions and might recognize different binding sites on different sets for transcripts, it is straightforward to see that the prediction would be hampered. Thus, it is important to initially describe the binding profile as holistically as possible.

In addition, I described how defined binding sites can be overlapped with orthogonal gene annotation data. These types of annotations are used in almost all high-throughput sequencing-based studies, for example in read mapping and alignment. Yet a common integration strategy is missing for iCLIP data. This starts with the choice of the annotation source. For human and mouse, GENCODE provides the full spectrum of annotated genes and putative isoforms (Frankish et al., 2019). Starting from such an inclusive data source coupled with specific filtering is preferred in most cases, compared to starting with an already fine-tuned and specialized annotation. This prevents missing out on potentially unexpected binding behaviors. Depending on the research question also more specific resources can be used as annotation base. FANTOM or MiRBase for example provide annotations specific for lncRNAs or micro-RNAs respectively, which could serve as a starting point for a targeted analysis (Forrest et al., 2014; Kozomara and Griffiths-Jones, 2010). Here the entire GENCODE annotation was used in the first place. It was then filtered for experimentally validated isoforms by RT-PCR, sequencing or by Havana manual curation (Howald et al., 2012). Such a filtering is of particular importance, since all downstream results depend on these decisions. Most commonly used approaches over-simplify this process by making arbitrary decisions early on in the analysis process. In many cases simply the longest annotated transcript is chosen as representative (Jin et al., 2020). In the case of a recent study describing the binding of U2AF65 based on additional cofactors, the analysis was restricted to annotated introns only (Sutandy et al., 2018). The above analysis on the U2AF65 binding spectrum revealed a similar trend, yet 10% of binding sites were still located outside of introns (Figure 3.8). On the other side also fine-tuned annotation resources exist for specific research questions. For example, in the case of the APPRIS database principal and alternative splice isoforms are annotated based on sequence, structure and conservation features (Rodriguez et al., 2013). The direct application of hierarchical rules is usually not desired when prior knowledge of the RBP is missing, since subsets of binding sites might remain unnoticed. In contrary specific rules might also be needed for certain binding sites to become visible. Many non-coding RNAs for example reside in the intron region of protein-coding genes and are thus missed when data analysis is tailored towards them (Uszczynska-Ratajczak et al., 2018; Olena and Patton, 2010). In a recent study the binding of PTBP1 was analyzed and the assignment to genomic features was done following a predefined hierarchy (Monzón-Casanova et al., 2020). In the present analysis of U2AF65 a hierarchical rule was used as well, but only after the application of a majority vote scheme. These two layers of resolving spurious annotations tend to introduce less errors, while still retaining most binding sites and thus lead to a better description of the binding preferences.

In summary, I showcased on the splicing regulator U2AF65 how accurate binding site definition can be done. The integration of merged replicates followed by reproducibility filtering allowed an increased sensitivity in the peak calling. This led to an accurate definition of binding sites, capturing a high proportion of the U2AF65 binding landscape. By avoiding most pitfalls when deciding on appropriate annotations, I could also show how the biological binding preference can be deduced from these binding sites.

## A novel SRSF6 binding motif

In diabetes, pancreatic beta-cells lose their insulin-producing capacity, which leaves patients with many short and long-term complications. Whereas in type 1 diabetes (T1D) an autoimmune response causes beta-cell death, patients with type 2 diabetes (T2D) suffer from gained insulin resistance (Weir and Bonner-Weir, 2013). GLIS3 is a major transcription factor and genetic variations of it are associated with susceptibility to both types of the disease (Dimitri et al., 2011; Senée et al., 2006; Taha et al., 2003). Beta-cells showed increased apoptosis upon *GLIS3* KD triggered by alternative splicing (AS) changes in the transcriptome (Barrett et al., 2009; Nogueira et al., 2013). Specifically, AS changes in the pro-apoptotic protein BIM lead to increased beta-cell death, via the activation the intrinsic mitochondrial pathway of apoptosis (Nogueira et al., 2013). These changes were shown to be caused by a decreased expression of the splicing regulator SRSF6, which is highly expressed in under normal conditions (Juan-Mateu et al., 2018). In summary, these studies performed by our collogues at the Eizirik group (ULB, Brussels) linked GLIS3 and SRSF6 expression to beta-cell survival. The exact mechanisms of SRSF6-induced AS changes however remained unexplained.

Here I described the impact of SRSF6 on the transcriptome by direct binding, through the analysis of iCLIP experiments of SRSF6 in EndoC-$\beta$H1 cells. I used the iCLIP workflow described above to carefully process four replicate experiments. More than 185,000 binding sites in nearly 9,000 genes could be identified. With about 16,000 genes being expressed in the EndoC-$\beta$H1 cells, this indicates that SRSF6 targets nearly 60% of all expressed genes in that cell type. This is in line with SRSF6 being a global splicing regulator affecting many splicing decisions (Screaton et al., 1995). Further, a preference for SRSF6 to bind to CDS from protein-coding genes was observed. Two earlier studies that described the splicing enhancer function of SRSF6 also showed a strong preference for exonic binding, predominantly in the CDS (Jensen et al., 2014; Müller-McNicoll et al., 2016). Furthermore, binding sites were also seen to cover the splice site region, thus extending beyond the exon-intron boundary. This is similar to the behavior of other members from the SR protein family. For example, binding to intronic splicing enhancers as well as recognition of the branch-point sequences are described in the literature (Lou et al., 1998; Änkö et al., 2012; Cho et al., 2011; Shen et al., 2004). As a secondary function, apart from splicing regulation, SRSF6 might also be involved in non-splicing-related functions, such as mRNA stability, export or translocation. This was also observed for other SR proteins and might explain the observed binding to 5' and 3' UTRs (Lemaire et al., 2002; Kim et al., 2014; Müller-McNicoll et al., 2016). Interestingly, a GO analysis of the bound targets revealed a similar picture (Figure 3.12). Splicing and mRNA processing-related terms were enriched, as well as numerous different regulatory terms, again pointing towards the multi-functional impact of SRSF6.

RBPs typically interact with specific motifs in the RNA sequence of their target transcripts. These motifs vary in length, position and composition between RBPs, giving rise to plenty of different RNA-binding possibilities. Previous studies described that SRSF6 and other SR proteins recognize short degenerate sequences of about 4 to 8-nt length (Änkö et al., 2012). In the present study I substantially refined the current knowledge of the binding specificity of SRSF6. Here in-depth sequence analysis was performed to describe the binding profile. Based on the defined binding sites, pentamer frequency analysis identified uridine- and GA-rich motifs (Figure 3.13). Directly at the binding site center uridine pentamers were seen to be most frequent, whereas GA-rich pentamers peaked outside of it. This showcased the effect of the UV-crosslinking bias which leads to binding sites being observed at uridine-rich positions (Sugimoto et al., 2012; Haberman et al., 2017; Chakrabarti et al., 2018). Such an enrichment is especially characteristic for proteins that interact with the RNA via RRM domains, such as SRSF6. These domains interact more frequently with uridine than with any other base (Corley et al., 2020). The GA-rich pentamers on the other side were enriched in the flanking regions displaying a clear positional pattern. Most prominently GAAGA and AAGAA showed a rising frequency from 100-nt upstream to 25-nt downstream of the binding sites (Figure 3.14). This was followed by a sharp drop in the frequencies, which indicates that SRSF6 specifically recognizes GA-rich sequence elements and positions towards the end of those (Figure 3.15). I further refined the binding motif to the resolution of triples, where two or more uninterrupted repetitions of the triplet GAA showed the strongest binding effect (Figure 3.16). The binding strength of sites with this motif grew progressively with increasing number of triplets. This suggests that multiple SRSF6 proteins assemble on a given exon to reach effective splicing enhancement. Such an assembly could be stabilized by protein-protein interactions via the RS domain (Figure 4.1). This domain is very versatile, but mainly found to be involved protein-protein interactions (Wu and Maniatis, 1993). However, it is also known that RS-domains in general might contact the RNA directly at the branchpoint or the 5' splice site (Shen et al., 2004; Cao et al., 2019). It would also be imaginable that the observed footprint resembles the contact of the two RRM domains from the same protein. Such an increased contact could lead to stronger binding potentially to ensure the desired splicing effect. However, it is typically hard to conclude such a fine-tuned mechanism without further structural validation. In a recent study for example iCLIP analysis and NMR structural biology were coupled to describe the binding of U2AF65 in detail (Kang et al., 2020). On the other hand, also, the binding of additional SR proteins or other RBPs that recognize the same motif could be favored, possibly via interactions with the RS-domain. Another possible scenario is that the repetition of the GAA motif provides a selective advantage for the recognition over more degenerative motifs. This could ensure reliable exon inclusion, but also offers the possibility for compensatory regulatory mechanisms among SR proteins (Pandit et al., 2013).

It is worth mentioning that the consensus motif for SRSF6 binding defined herein differs from those reported in previous studies. Mainly SELEX (systematic evolution of ligands by experimental enrichment), but also RNA immunoprecipitation experiments were used to describe the binding of SRSF6 in different cell types, but to my knowledge no conclusive consensus motif has been reported (Liu et al., 1998; Park et al., 2019; Screaton et al., 1995). The motif described above however perfectly fitted to a motif that was recently derived from SRSF6 iCLIP experiments in mouse cells (Müller-McNicoll et al., 2016). Additionally, the GAA-rich motif that I described has a higher similarity to binding site sequences from other SR proteins like SRSF1, SRSF4 and SRSF7 compared to previously defined motifs (Änkö et al., 2012; Änkö et al., 2010; Sanford et al., 2008; Zheng et al., 1997). This underpins the result, since SR proteins might frequently share functions and compensate each other in a highly regulated cross-talk. SRSF4 for example is known to partially compensate for a loss of SRSF6 and vice versa (Müller-McNicoll et al., 2016). Taken together, these observations suggest that the RNA sequence specificity of SRSF6 *in vivo* differs substantially from the *in vitro* specificity previously reported by SELEX experiments. This is of major importance for the scientific field, since SR protein motifs are commonly used, for example in mechanistic studies to predict the presence of exonic splicing enhancers (ESEs) (Cartegni et al., 2003).

Detailed insights of the SRSF6 binding motif also impact the understanding of human diseases. For example, transcriptomic genome-wide association studies are used to characterize somatic variants and single-nucleotide polymorphisms (SNPs) in patient groups (Tranchevent et al., 2017; Solovyev and

Shahmuradov, 2003). Thus, the association of a binding site and a SNP might shed new lights on regulatory processes in certain diseases. The findings described within this chapter thus have important implications for the interpretation of these variants and their putative role in splicing regulation. This reaches beyond the scope of diabetes alone, since SRSF6 is for example a known proto-oncogene and contributes to diverse forms of cancer (Karni et al., 2007).

## SRSF6 triggers AS changes by direct binding

Our colleagues from the Eizirik group previously reported that SRSF6 regulates a network of AS events in pancreatic beta-cells (Juan-Mateu et al., 2018). They showed that SRSF6 influences the splicing of the mRNA encoding apoptotic regulator proteins B-cell lymphoma 2 like 11 (BCL2L11/BIM) and B-cell lymphoma 2 associated X (BCL2 associated X/BAX), which are involved in pancreatic beta-cell apoptosis (Dooley et al., 2016; Schwerk and Schulze-Osthoff, 2005). They further showed that SRSF6 also acted upon members of the JNK signaling cascade, which plays a role in in pancreatic beta-cell death as well (Juan-Mateu et al., 2018; Gurzov and Eizirik, 2011; Fu et al., 2009; Cunha et al., 2012). Thus, SRSF6 influences insulin secretion by impacting beta-cell survival. Of note, in our analysis both BIM and BAX were found to be targets of SRSF6, pointing towards direct effects.

Here, I re-analyzed and integrated their RNA-Seq data with the above iCLIP data, to demonstrate that the majority of AS events are associated with direct binding of SRSF6. Initial AS analysis showed about 1,000 genes with significantly affected splicing isoforms due to the *SRSF6* KD. I confirmed that cassette exon events were the predominant form of AS changes, with a preference for increased exon skipping over exon inclusion (Figure 3.21). This suggests that SRSF6 under normal conditions promotes exon inclusion and might stabilize splicing by guiding splice site recognition. Moreover, my analysis showed that SRSF6-driven splicing regulation is highly context- and position-dependent, as previously described for other splicing factors (Bradley et al., 2015; Ke and Chasin, 2011; Pandit et al., 2013). For the integration of the iCLIP data with the observed splicing changes, differences in isoform abundance had to be accounted. This was controlled by computing PSI-matched background sets specifically for exons that decreased or increased their inclusion ratio upon *SRSF6* KD (Figure 3.26). By doing so it could be ensured that observations reflect genuine regulatory RNA-binding profiles, rather than intrinsic biases. Such biases commonly impair conclusions from a direct comparison of up- and down-regulated exons (Pandit et al., 2013). Based on this, an increased binding of SRSF6 within exons that are downregulated upon *SRSF6* KD could be observed. This again supports the idea that SRSF6 mainly contributes to exon inclusion. Similar observations have been made by early *in vitro* studies that reported SR proteins bind primarily to pre-mRNAs at ISEs and ESEs (Graveley et al., 1999). By doing so, they assist in the recruitment of spliceosomal components to the neighboring splice sites. Alternatively, SRSF6 might also interfere with the binding of splicing silencers, such as heterogeneous nuclear ribonucleoproteins (hnRNP). In particular, we observed SRSF6 binding to the polypyrimidine tract where it could compete with hnRNP (Long and Caceres, 2009; Zarnack et al., 2013). A precise description of the underlying mechanism by which SR proteins favor splice-site recognition is still questionable. Some studies for example described that SRSF1 guides the recruitment of the U1 snRNP protein U1-70k to facilitate donor splice site recognition (Cho et al., 2011; Wu and Maniatis, 1993). Further similar findings were made also for other SR proteins (Fu and Maniatis, 1992). It has been shown that besides U1 snRNP, also the U2 auxiliary factor 2 (U2AF) recruitment can be influenced by SRSF6 binding. SRSF6 binding is thought to increase the affinity of U1 and U2 to the respective 5' and 3' splice sites (Long and Caceres, 2009). Additionally, also the interference of SRSF6 with hnRNP on exonic sites is knonw (Zhu et al., 2001; Cáceres et al., 1994; Long and Caceres, 2009). It is possible that the recruitment of SRSF6 to the ESE is required to counteract a nearby ESS region which is for example bound by an hnRNP protein (Zhu et al., 2001). Thus, SRSF6 might block ESS regions to indirectly stabilize the recognition of the exon-intron boundary. Interestingly, we observed about 200 exons that displayed increased inclusion upon *SRSF6* KD, which suggests that their inclusion is weakened under normal conditions (Figure 3.27). Such upregulated exons showed strong SRSF6 binding directly on the flanking constitutive exons and the polypyrimidine tract. By binding to these flanking exons, SRSF6 might enforce the exon-intron boundary definition of these exons, while the alternative exon is only poorly recognized.

Taken together, this suggests that SRSF6 recognizes mainly ESE elements of alternative exons to reinforce potentially weaker splices sites and assists in exon inclusion (Figure 4.1). Thereby SRSF6 specifically recognizes a GA-rich binding motif. Similar, yet different binding motifs would also allow for further cross-regulation, which is known among the family of SR proteins (Lareau et al., 2007; Ni et al., 2007). Similar findings were also made for other splicing-regulatory proteins, which can enhance or repress alternative exon inclusion

(Sanford et al., 2009; Eperon et al., 1993; Han et al., 2011). This reinforces the idea that alternative splicing is a complex interplay of sequence elements and different splicing-regulatory proteins interacting with each other.

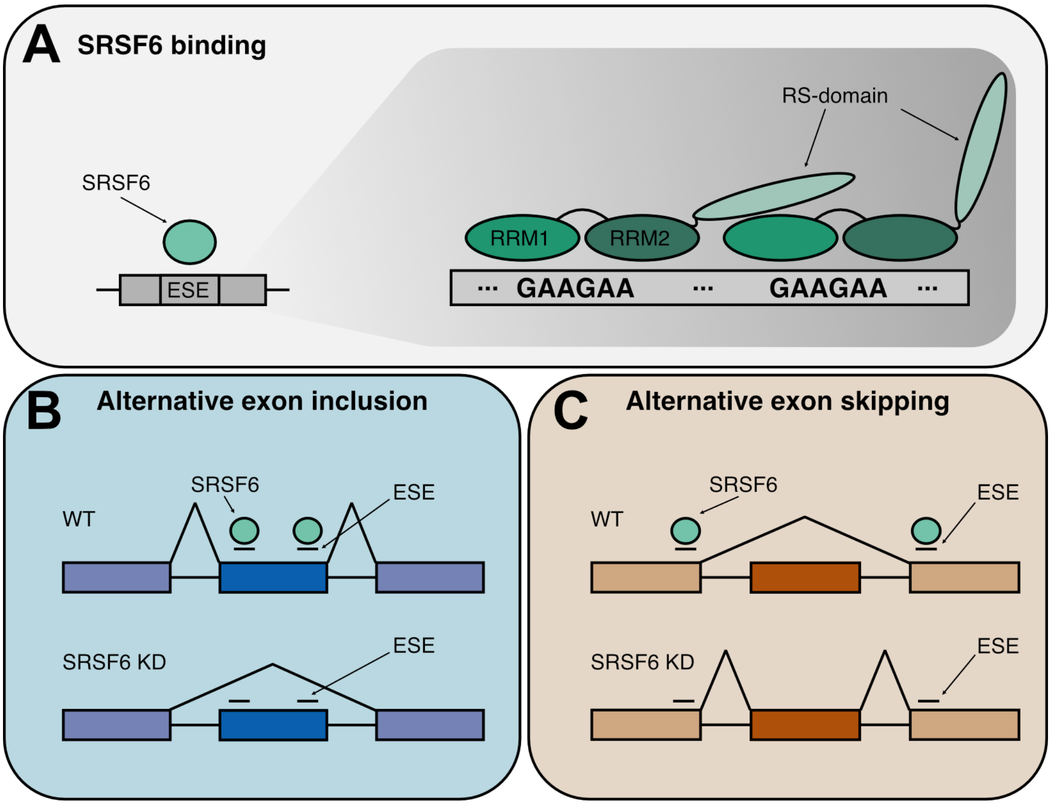

**Figure 4.1: Proposed schematic of the SRSF6-mediated AS regulation.** (A) SRSF6 binds to GA-rich sequence elements of exonic splicing enhancer (ESE) regions. The RNA contact is probably mediated by the RRM domains, while the RS domain might assist in protein-protein interactions, either with additional SRSF6 proteins or further splicing regulators. (B) SRSF6 enforces splice site recognition of the alternative exon by binding ESEs on the alternative exon. (C) SRSF6 promotes alternative exon skipping by reinforcement of flanking constitutive exons.

## SRSF6 regulates diabetes susceptibility genes

We were prompted to the influence of SRSF6 on diabetes by the susceptibility gene GLIS3, which encodes an important transcription factor for beta-cell maintenance (Nogueira et al., 2013; Juan-Mateu et al., 2018). Mutations on the gene itself lead to severe neonatal diabetes, whereas lowered expression increases the risk for T1D and T2D (Barrett et al., 2009; Cho et al., 2012; Dupuis et al., 2010; Steck et al., 2014; Winkler et al., 2014). GLIS3-depleted beta-cells in rats were shown to be associated with the inhibition of SRSF6, which in turn led to a splicing shift of the proapoptotic protein BCL2L11 (Nogueira et al., 2013). This protein is known to contribute to pancreatic beta-cell apoptosis upon for example high glucose induction (McKenzie et al., 2010). This was explicitly shown to be regulated by AS changes, that shift the equilibrium to the expression of the isoform BIM S which is the most apoptotic (Nogueira et al., 2013).

Interestingly, we observed that the KD of *SRSF6* does not lead to severe changes in the global transcription levels. Differential expression analysis revealed only minute changes in transcript abundance, with no preference for up- or down-regulation (Figure 3.20). The comparison with the observed AS changes also revealed no direct correlation, which suggests that no global regulatory pathway such as a widespread degradation of mis-spliced mRNAs via nonsense-mediated mRNA decay (NMD) is triggered (Figure 3.22). Usually NMD is a common consequence of AS, especially for the SR protein family (Ni et al., 2007; Lareau et al., 2007). We found that SRSF6 rather triggers AS changes to specifically alter certain isoforms. Thus, we reasoned that some of these targets might be diabetes susceptibility genes, potentially bridging the gap between SRSF6-mediated AS and the influence on diabetes. This hypothesis was approached by looking for direct SRSF6-regulated splicing events in a compiled list of known T1D and T2D susceptibility genes (Figure 3.30). Among these 400 candidates five T1D and 17 T2D well-documented diabetes susceptibility genes were identified (Pociot, 2017; Barrett et al., 2009; Bradfield et al., 2011). These susceptibility genes are in general predicted to mediate gene-environment interactions in diabetes, although further detailed molecular analysis would be required to confirm these statements.

Taken together, the present work extended the current knowledge of SRSF6-controlled susceptibility genes beyond the role of the known proapoptotic regulators BCL2L11 and BCL2 associated X (Juan-Mateu et al., 2018). We thereby highlighted the effect of SRSF6 as a downstream target of GLIS3, acting on beta-cell survival through the regulation of diabetes susceptibility genes via AS. I explored the genome-wide binding profile of the RNA-binding protein SRSF6 in human pancreatic beta-cells and integrated these results with observed AS changes triggered by the decreased expression of SRSF6. The fact that several diabetes susceptibility genes are directly targeted by this mechanism suggests the presence of a AS-regulated network with putative impact on diabetes risk.

# 5 | Conclusion and outlook

In the present thesis, I investigated the computational processing of iCLIP derived sequencing data to study protein-RNA interactions. In particular I investigated the binding profile of U2AF65 and SRSF6. In the first case a published dataset was used to describe the initial processing, whereas in the latter case conclusions on the impact of SRSF6 binding in the context of diabetes were drawn.

In the first part a pipeline to process iCLIP sequencing data was developed. On the example of the RNA binding protein U2AF65 I showed how quality control, peak calling, replicate integration and binding site annotation can be achieved in a reproducible manner. The described workflow allows the detection of protein-RNA interactions on lowly abundant transcripts, as well as on low affinity binding sites, thus revealing a larger proportion of the actual binding spectrum of a given RBP. Such an enhanced binding spectrum lowers the number of missed binding sites and therefore might also reveal unknown functions of RBPs. For example, an RBP might influence a broad range of transcripts by interaction with the splicing machinery, but also defines the translocation of a small set of transcripts by interaction with the UTR region. In such cases a coarse computation of binding sites might detect the splicing influence, but misses out on the secondary function. This has a potential large impact, since many RBPs express more than one function in the cell (Singh et al., 2015; Dreyfuss et al., 2002). Additionally, such a standardized pipeline opens up the possibility of a wide range of comparisons. For example, the binding of different RBPs might be compared to resolve the assembly of large protein complexes or regulatory mechanisms. On the other side, one might also analyze the change in the binding spectrum of an RBP between different cell types, conditions or induced mutations. The comparison of such differences within the binding spectrum of an RBP is a relatively new approach, for which a standardized binding site computation pipeline is a prerequisite. Thus, the results of the present thesis allow to enhance the resolution of the given experimental designs, while also allowing for the possibility of new research questions to be asked.

The second part of this thesis described the binding spectrum of SRSF6 in the context of diabetes. In human pancreatic beta-cells SRSF6 acts downstream of the susceptibility gene GLIS3. Here, I described how SRSF6 itself regulates the splicing of further diabetes susceptibility genes by direct binding in the pancreatic beta-cell line EndoC-$\beta$H1. I described that SRSF6 enhances the inclusion of certain exons via binding to ESE elements on the alternative exon. It was also observed that exon skipping can be promoted by binding to ESEs on flanking constitutive exons. This mechanism explained, at least in parts, how the beta-cell transcriptome is changed in T1D and T2D. Genes targeted by this mechanism were found to be associated with cell cycle regulation and apoptosis. The exact molecular modifications however require further functional validations. To treat diseases at the level of RNA splicing is relatively new, but increasingly researched to potentially cure complex diseases on the molecular level. For example, in the case of spinal muscle atrophy the splicing of the SMN (survival motor neuron) genes can be modulated to restore the healthy splicing pattern (Levin, 2019). Technically such approaches make use of antisense oligonucleotides (ASOs) to block specific binding sites. As a proof of principle, we already used ASOs in the related publication to modulate the splicing of the *LMO7* gene (Alvelos et al., 2020). In the case of SRSF6 binding sites on further susceptibility genes could be used as potential targets. ASOs could be designed specifically to the sequence of these binding sites and used to restore the downstream splicing defects in the SRSF6 targets. This would open up potential therapeutic opportunities, since the equilibrium of alternative splicing events could be altered by the blockage of these positions. This would ultimately change the transcriptional state of the cell potentially restoring beta-cell function. Many clinical approaches nowadays make use of bioinformatic tools and next-generation sequencing data to complement diagnostics for complex diseases. In particular the prediction of *cis*-acting sequence variants is of major importance. Services like the Human Splicing Finder, or ESEfinder search for splicing regulatory elements (SRE) based on known binding sequences of splicing RBPs, such as SRSF6 (Desmet et al., 2009; Cartegni et al., 2003). Here, the improved understanding of the SRSF6 binding motif is of great advantage. It is commonly known that the quality of the prediction of SREs varies within clinical cases or scientific contexts (Baralle and Buratti, 2017). Thus a better understanding of SRSF6 binding could globally affect the prediction of SREs and therefore impacts the treatment of diseases, even beyond the scope of diabetes.

# Supplementary Material

## In-house R scripts

### Binding site merge and resize

The following scripts contain the R code for the binding site definition. The central function ”mergePeaks” is used to summarize crosslink sites into binding site of a user-defined width. These sites can be filtered for different properties by using the functions ”pMinPos”, ”pPureCenter”, ”nSitesPureClip”, ”pMaxCenter”. All of these functions were designed to be for internal use. The end user can should call the wrapper-function ”peaksToBindingSites” for convenience.

**Listing 5.1: Functions for the merge and resize routine.**

```
#' Merge and resize peaks called by PureCLIP. Single nt peaks are merged into binding sites of equal width, by either extending or iteratively reducing the size of the merged peak object.
#'
#' @param peaks GRanges-Object (the PureCLIP output)
#' @param peakSize numeric (desired output peak size)
#' @param peaksSize_lowerThreshold numeric (define what merged regions will be retained)
#' @param clipSignalPlus RLE (single nuclotide crosslink events of + strand)
#' @param clipSignalMinus RLE (single nuclotide crosslink events of - strand)
#' @return GRanges-Object of merged and resized peaks
```

```r
#' @example peaksMergeAndResize(peaks = p, peakSize = 9, peaksSize_lowerThreshold = 2, clipSignalPlus = plus, clipSignalMinus = minus)
mergePeaks <- function(peaks, peaksSize, csPlus,
                       csMinus, pMinSize){
  ps1 = keepStandardChromosomes(peaks,pruning.mode = "coarse")
  # merged peaks with gap width that is 1nt smaller than the desired output binding site width
  ps2 = reduce(ps1, min.gapwidth = peaksSize - 1)
  # remove merged regions smaller than threshold (usually 1 or 2)
  ps3 = ps2[width(ps2) > pMinSize]
  names(ps3) = 1:length(ps3)
  # merge and resize routine
  pfrPlus <- GRanges()
  pfrMinus <- GRanges()
  ptpPlus <- subset(ps3, strand == "+")
  ptpMinus <- subset(ps3, strand == "-")
  # initiate resize
  Counter = 0
  while (TRUE) {
    # quit if no more regions to check
    if (length(ptpMinus) == 0 & length(ptpPlus) == 0) {
      break
    } else {
      # handle positive strand
      if (length(ptpPlus) != 0) {
        # get max xlink position of each peak
        pmPlus = as.matrix(csPlus[ptpPlus])
        pmPlus[is.na(pmPlus)] = -Inf
        pmPlus = max.col(pmPlus, ties.method = "first")
        # make new peaks centered arround max position
        cpPlus = ptpPlus
        start(cpPlus) = start(cpPlus) + pmPlus - 1
        end(cpPlus) = start(cpPlus)
        cpPlus = cpPlus + ((peaksSize - 1)/2)
        # store peaks
        pfrPlus = c(pfrPlus, cpPlus)
        # remove peak regions from rest of possible regions
        cpPlus = as(cpPlus + ((peaksSize - 1)/2),
                    "GRangesList")
        # update peak regions that are left for processing
        ptpPlus = unlist(psetdiff(ptpPlus, cpPlus))
      }
      # handle negative strand
      if (length(ptpMinus) != 0) {
        pmMinus = as.matrix(csMinus[ptpMinus])
        pmMinus[is.na(pmMinus)] = -Inf
        pmMinus = max.col(pmMinus, ties.method = "last")
        # make new peaks centered arround max position
        cpMinus = ptpMinus
        start(cpMinus) = start(cpMinus) + pmMinus - 1
        end(cpMinus) = start(cpMinus)
        cpMinus = cpMinus + ((peaksSize - 1)/2)
        # store peaks
        pfrMinus = c(pfrMinus, cpMinus)
        # remove peak regions from rest of possible regions
        cpMinus = as(cpMinus + ((peaksSize - 1)/2),
                     "GRangesList")
        # update peak regions that are left for processing
        ptpMinus = unlist(psetdiff(ptpMinus, cpMinus))
      }
      Counter = Counter + 1
    }
  }
  ps4 = c(pfrPlus, pfrMinus)
  ps4 = sortSeqlevels(ps4)
  ps4 = sort(ps4)
  return(ps4)
}

#' Filter that retains only peaks that harbour a defined number of single nt crosslink events.
#'
#' @param peaks GRanges-Object (the merged and resized peaks)
#' @param minXlinksPerPeak number (threshold for the number of xlink events)
#' @param csPlus RLE (single nuclotide crosslink events of + strand)
#' @param csMinus RLE (single nuclotide crosslink events of - strand)
#' @example pMinPos(peaks = p, minXlinksPerPeak = 3, csPlus = plus, csMinus = minus)
pMinPos <- function(peaks, minXlinksPerPeak, csPlus, csMinus){
  # split by strand plus
  pcPlus = peaks[strand(peaks) == "+"]
  pcPlusMat = as.matrix(csPlus[pcPlus])
  pcPlus = pcPlus[apply((pcPlusMat > 0), 1, sum) > minXlinksPerPeak]
  # split by strand minus
  pcMinus = peaks[strand(peaks) == "-"]
  pcMinusMat = as.matrix(csMinus[pcMinus])
  pcMinus = pcMinus[apply((pcMinusMat > 0), 1, sum) > minXlinksPerPeak]
  # combine sort return
  pCurr = c(pcPlus, pcMinus)
  pCurr = sortSeqlevels(pCurr)
  pCurr = sort(pCurr)
  return(pCurr)
}

#' Filter that removes all binding sites where the center position was not an initially called PureCLIP crosslink site.
#'
#' @param peaks GRanges-Object (the merged and resized peaks)
#' @param pureClipOriginal GRanges-Object (the PureCLIP output)
#' @example pPureCenter(peaks = p, pureClipOriginal = pPureClip)
pPureCenter <- function(peaks, pureClipOriginal){
  pCurr = peaks - (unique(width(peaks)) - 1)/2
  pCurr = peaks[queryHits(findOverlaps(pCurr, pureClipOriginal))]
  # combine sort return
  pCurr = sortSeqlevels(pCurr)
  pCurr = sort(pCurr)
  return(pCurr)
}

#' Filter that retains only those binding sites that harbor a defined number of initially called PureCLIP sites
#'
#' @param peaks GRanges-Object (the merged and resized peaks)
#' @param pureClipOriginal GRanges-Object (the PureCLIP output)
#' @param minSitesPure number (minimum PureCLIP site threshold)
#' @example nSitesPureClip(peaks = p, pureClipOriginal = pOriginal, minSitesPure = 2)
nSitesPureClip <- function(peaks, pureClipOriginal,
                           minSitesPure){
  overlaps = findOverlaps(peaks, pureClipOriginal)
  freq = table(queryHits(overlaps))
  idx = as.numeric(names(freq[freq >= minSitesPure]))
  pCurr = peaks[idx]

  # combine sort return
  pCurr = sortSeqlevels(pCurr)
  pCurr = sort(pCurr)
}

#' Filter that removes all binding sites whose center position does not harbor the highest crosslink event count
```

```r
#'
#' @param peaks GRanges-Object (the merged and resized peaks)
#' @param csPlus RLE (single nuclotide crosslink events of + strand)
#' @param csMinus RLE (single nuclotide crosslink events of - strand)
#' @example pMaxCenter(peaks = p, csPlus = plus, csMinus = minus)
pMaxCenter <- function(peaks, csPlus, csMinus){
  # find max per peak
  peaksPlus = peaks[strand(peaks) == "+"]
  peaksMinus = peaks[strand(peaks) == "-"]
  pcPlusMat = as.matrix(csPlus[peaksPlus])
  pcMinusMat = as.matrix(csMinus[peaksMinus])
  pcPlusCount = apply(pcPlusMat, 1, max)
  pcMinusCount = apply(pcMinusMat, 1, max)
  # remove not max peaks
  pcPlus = peaksPlus[pcPlusCount == pcPlusMat[,((unique(width(peaks)) - 1) / 2) + 1]]
  pcMinus = peaksMinus[pcMinusCount == pcMinusMat[,((unique(width(peaks)) - 1) / 2) + 1]]
  # combine sort return
  pCurr = c(pcPlus, pcMinus)
  pCurr = sortSeqlevels(pCurr)
  pCurr = sort(pCurr)
  return(pCurr)
}

#' Wrapper function that subsequently calls: mergePeaks, pMinPos, pPureCenter, pMaxCenter, nSitesPureClip and reports basic statistics about how many peaks are removed in each step
#'
#' @param peaks GRanges-Object (the PureCLIP output)
#' @param peakSize numeric (desired output peak size)
#' @param pMinSize numeric (define what merged regions will be retained)
#' @param minSitesPure number (minimum PureCLIP site threshold)
#' @param minXlinksPerPeak number (threshold for the number of xlink events)
#' @param csPlus RLE (single nuclotide crosslink events of + strand)
#' @param csMinus RLE (single nuclotide crosslink events of - strand)
#' @example peaksToBindingSites(peaks = peaksFiltered, peaksSize = 9,
#' pMinSize = 2, minXlinksPerPeak = 2, minSitesPure = 2,
#' csPlus = csPlus, csMinus = csMinus)
```

```r
peaksToBindingSites <- function(peaks, peaksSize,
                                pMinSize,
                                minSitesPure, minXlinksPerPeak,
                                csPlus, csMinus){
  peaks1 = mergePeaks(peaks = peaks,
                      peaksSize = peaksSize,
                      pMinSize = pMinSize,
                      csPlus = csPlus,
                      csMinus = csMinus)
  peaks2 = pMinPos(peaks = peaks1,
                   minXlinksPerPeak = minXlinksPerPeak,
                   csPlus = csPlus,
                   csMinus = csMinus)
  peaks3 = pPureCenter(peaks = peaks2,
                       pureClipOriginal = peaks)
  peaks4 = pMaxCenter(peaks = peaks3,
                      csPlus = csPlus,
                      csMinus = csMinus)
  peaks5 = nSitesPureClip(peaks = peaks4,
                          pureClipOriginal = peaks,
                          minSitesPure = minSitesPure)
  reportDf = data.frame(
    pStep = c("InitialPeaks", "Merge",
             "MinPosFilter", "CenterPureClipFilter",
             "CenterIsMaxFilter", "minSitesPure"),
    nPeaks = c(length(peaks), length(peaks1),
               length(peaks2),length(peaks3),
               length(peaks4), length(peaks5)))
  result = list(reportDf = reportDf, peaks = peaks5)
  return(result)
}
```

## SRSF6 binding site definition

The following script was used to compute the binding sites specifically for SRSF6. The function ”peaksToBindingSites” is called to compute 9-nt wide binding sites in the first place. These sites are then filtered for replicate reproducibility and overlapped with gene and transcript annotations, as it is described in the respective section of the script.

**Listing 5.2: SRSF6 Binding Site Definition.**

```
### ==============================================================
### Load required packages
### --------------------------------------------------------------
library(rtracklayer)
library(GenomicRanges)
library(GenomicFeatures)
library(dplyr)


### ==============================================================
### Import PureCLIP output
### --------------------------------------------------------------
peaksInitial = "./data/PureCLIP_output.bed"
peaksInitial = import(con = peaksInitial, format = "BED")


### ==============================================================
### Filter peaks by PureCLIP score -> glaobal 5% cutoff
### --------------------------------------------------------------
cutoff = quantile(peaksInitial$score, probs = seq(0,1, by = 0.05))
peaksFiltered = peaksInitial[peaksInitial$score >= cutoff[2]]
rtracklayer::export(peaksFiltered, "./data/peaksFilteredGlobal.bed
    ", format = "BED")


### ==============================================================
### Summarize peaks into 9nt wide binding sites
### --------------------------------------------------------------
clipSignalPlus = "./data/clip_merge_plus.bw"
clipSignalPlus = import.bw(clipSignalPlus, as = "Rle")
clipSignalMinus = "./data/clip_merge_minus.bw"
clipSignalMinus = abs(import.bw(clipSignalMinus, as = "Rle"))

peaksProcessed = peaksToBindingSites(peaks = peaksFiltered,
                                     peaksSize = 9,
                                     peaksSize_lowerThreshold = 2,
                                     minXlinksPerPeak = 2,
                                     minPureClipSites = 2,
                                     clipSignalPlus = clipSignalPlus,
                                     clipSignalMinus = clipSignalMinus)

rtracklayer::export(peaksProcessed$peaks, "./data/peaksMerged.bed", format = "BED")

### =================================================================
### Filter binding sites for reproducibility on replicate level
### -----------------------------------------------------------------
# import replicates as RLE
folder = "./data/clip_all_replicates/"
files = list.files(folder, pattern = ".bw$", full.names = TRUE)
files = lapply(files, function(x){
  c = abs(import(x, format = "BigWig", as = "Rle"))
  std = standardChromosomes(c)
  c = c[names(c) %in% std]
})
names(files) = 1:8

# describe the datasete
infos = data.frame(
  ID = 1:8,
  condition = rep("wt",8),
  strand = rep(c("-","+"),4),
  replicate = c(1,1,2,2,3,3,4,4))

# split peaks by strand
peaks = peaksProcessed$peaks
peaksPlus = peaks[strand(peaks) == "+"]
peaksMinus = peaks[strand(peaks) == "-"]

# count crosslinks on peaks -> plus strand
signalPlus = files[names(files) %in% infos$ID[infos$strand == "+"]]
countsPlus = sapply(1:length(signalPlus), function(x){
  sum(signalPlus[[x]][peaksPlus])
})
mcols(peaksPlus) = countsPlus

# count crosslinks on peaks -> minus strand
signalMinus = files[names(files) %in% infos$ID[infos$strand == "-"]]
```

```
countsMinus = sapply(1:length(signalMinus), function(x){
  sum(signalMinus[[x]][peaksMinus])
})
mcols(peaksMinus) = countsMinus

# combine counts of both strands
peaksCount = c(peaksPlus,peaksMinus)
peaksCount = sortSeqlevels(peaksCount)
peaksCount = sort(peaksCount)
colnames(mcols(peaksCount)) = c("rep1", "rep2", "rep3", "rep4")

# compute replicate specific count threshold
df = as.data.frame(mcols(peaksCount))
cutoff = apply(df, 2, function(x){
  quantile(x, probs = seq(0,1, by = 0.1))
})
cutoff = data.frame(q = cutoff[3,], variable = colnames(cutoff))
cutoff$q[cutoff$q < 2] = 2

# summarize replicate reproducibility
support = t(apply(mcols(peaksCount), 1, function(x){
  ifelse(x >= cutoff$q, 1, 0)
}))

mcols(peaksCount)$support = rowSums(support)
peaksReproducible = peaksCount[peaksCount$support >= 3]


### ==============================================================
### Filter genome annotation
### --------------------------------------------------------------
# Import annotation file
annotationFile = "./data/gencode.v29.annotation.gtf"
anno = import(annotationFile, format = "GTF")

# Filter feature level annotation
anno = anno[anno$level != 3]

# Filter transcript level annotation
anno$transcript_support_level[is.na(anno$transcript_support_level)] = 0
anno$transcript_support_level[anno$transcript_support_level == "NA"] = 10
anno = anno[anno$transcript_support_level == 0
      | anno$transcript_support_level == 1
      | anno$transcript_support_level == 2
```

```
            | anno$transcript_support_level == 3 ]

# Create txdb databse from filtered annotations
annoDb = makeTxDbFromGRanges(anno)

genesAll = genes(annoDb)
idx = match(genesAll$gene_id, anno$gene_id)
elementMetadata(genesAll) = cbind(elementMetadata(genesAll), elementMetadata(anno)[idx,])


### ==============================================================
### Assign each binding site to the hosting gene
### --------------------------------------------------------------
genesTargets = subsetByOverlaps(genesAll, peaksReproducible)
genesTargetsProt = genesTargets[genesTargets$gene_type == "protein_coding"]
peaksProt = subsetByOverlaps(peaksReproducible, genesTargetsProt)
peaksFinal = peaksProt[countOverlaps(peaksProt, genesTargetsProt) == T]


### ==============================================================
### Specify location of each binding site in the transcript
### --------------------------------------------------------------
# count binding site overlap with transcript parts
cdseq = cds(annoDb) %>% countOverlaps(peaksFinal,.)
intrns = unlist(intronsByTranscript(annoDb)) %>% countOverlaps(peaksFinal,.)
utrs3 = unlist(threeUTRsByTranscript(annoDb)) %>% countOverlaps(peaksFinal,.)
utrs5 = unlist(fiveUTRsByTranscript(annoDb)) %>% countOverlaps(peaksFinal,.)
overlapCounts = data.frame(cds = cdseq, intron = intrns, utr3 = utrs3, utr5 = utrs5)

# init the hierarchical rule for ties
rule = c("intron", "cds", "utr3", "utr5")

# applying the majority vote and rule
overlapCounts = overlapCounts[, rule] %>% as.matrix %>%
  cbind.data.frame(., outside = ifelse(rowSums(overlapCounts) == 0, 1, 0) )
names = colnames(overlapCounts)
reg = apply(overlapCounts, 1, function(x){ names[which.max(x)] })

```

```
# add final regions to binding sites object
mcols(peaksFinal)$region = reg

# remove binding sites outside of annotated regions
# (this is a special case where binding sites are located within an annotated
# gene, but outside of any of it’s transcripts)
peaksFinal = peaksFinal[peaksFinal$region != "outside"]


### ==============================================================
### re-annotate final binding sites with initial PureCLIP scores
### --------------------------------------------------------------
overlaps = findOverlaps(peaksFinal, peaksInitial)
matchDF = data.frame(qHits = queryHits(overlaps),
                     sHits = subjectHits(overlaps),
                     score = peaksInitial$score[subjectHits(overlaps)])
scores = group_by(matchDF, qHits) %>%
  summarize(pSum = sum(score),
            pMax = max(score),
            pMean = mean(score)) %>%
  as.data.frame

mcols(peaksFinal)$pSum = scores$pSum
mcols(peaksFinal)$pMean = scores$pMean
mcols(peaksFinal)$pMax = scores$pMax
```

## SRSF6 binding motif definition

The following script holds the code to compute the SRSF6 binding motif. It requires the binding sites computed with the scripts explained above as input. Internally binding sites are represented as GenomicRanges objects. For the kmer counting binding sites are represented by their central position. Kmer counting is based on the stri_ locate_ all function from the stringi package. For convenience several wrapper functions are given in the second code section.

**Listing 5.3: Definition of the SRSF6 binding motif.**

```
### ==============================================================
### Load required packages
### --------------------------------------------------------------
library(Biostrings)
library(BSgenome.Hsapiens.UCSC.hg38)
library(GenomicRanges)
library(GenomicFeatures)
library(stringi)
library(reshape2)
library(ggplot2)
library(ggrepel)

### ==============================================================
### Load processed binding sites
### --------------------------------------------------------------
load("./data/bindingSites.rda")
peaksFinal = bs

### ==============================================================
### Subsample binding sites per region
### --------------------------------------------------------------
set.seed(1234)
sampCds = subset(peaksFinal, region == "cds") %>%
  sample(., 5000)
sampIntron = subset(peaksFinal, region == "intron") %>%
  sample(., 5000)

### ==============================================================
### Calculate pentamer frequency for each subsample
### --------------------------------------------------------------
# pentamer freq in 9nt binding site center windows
df1 = countKmerFreq(sampCds, kmerSize = 5)
df2 = countKmerFreq(sampIntron, kmerSize = 5)
dfCenter = data.frame(kmer = rownames(df1),
                      exon = rowMeans(df1),
                      intron = rowMeans(df2),
                      region = "center")

df1 = countKmerFreq(flank(sampCds, width = 20, start = T),
                    kmerSize = 5)
df2 = countKmerFreq(flank(sampIntron, width = 20, start = T),
                    kmerSize = 5)
dfUpstream = data.frame(kmer = rownames(df1),
                        exon = rowMeans(df1),
                        intron = rowMeans(df2),
                        region = "upstream")

df1 = countKmerFreq(flank(sampCds, width = 20, start = F),
                    kmerSize = 5)
df2 = countKmerFreq(flank(sampIntron, width = 20, start = F),
                    kmerSize = 5)
dfDownstream = data.frame(kmer = rownames(df1),
                          exon = rowMeans(df1),
                          intron = rowMeans(df2),
                          region = "downstream")

### ==============================================================
### Example plot for exon vs. intron pentamer frequencies
### --------------------------------------------------------------
p = ggplot(dfCenter, aes(x = exon, y = intron)) +
  geom_abline() +
  geom_point(color = "darkgrey") +
  geom_label_repel(data = subset(dfCenter,
                                 exon > 0.03 | intron > 0.03),
                   aes(x = exon, y = intron, label = kmer),
                   color = "black", nudge_y = .001,
                   min.segment.length = unit(0, ’lines’),
                   size = 2, segment.size = .5, alpha = .5,
                   segment.alpha = .3) +
  ggtitle("Center - 9nt")


### ==============================================================
### Calculate triplet frequency for GAA/ UUC
### --------------------------------------------------------------
seqSet = RNAStringSet(getSeq(Hsapiens, peaksFinal - 4 + 30 ))
freq = oligonucleotideFrequency(seqSet, 3) %>% as.data.frame()

# extract triplets of interest
selKmers = c("GAA", "UUC")
selKmers = freq[colnames(freq) %in% selKmers]
```

```

# group triplet frequency with PureCLIP score
df = data.frame(region = peaksFinal$region, pSum = peaksFinal$pSum)
df = cbind(df, selKmers)
df = melt(df, id = c("region", "pSum"))


### ============================================================
### Calculate triplet frequency for GAA/ UUC with gap sizes (0,1,2)
### ------------------------------------------------------------
seqSet = RNAStringSet(getSeq(Hsapiens, peaksFinal - 4 + 30)) %>%
  as.character()

# single motif
d0 = data.frame(region = peaksFinal$region,
                pSum = peaksFinal$pSum,
                count = stri_count_regex(str = seqSet,
                                         pattern = "UUC"),
                pattern = "UUC")
d1 = data.frame(region = peaksFinal$region,
                pSum = peaksFinal$pSum,
                count = stri_count_regex(str = seqSet,
                                         pattern = "GAA"),
                pattern = "GAA")
dfSingleMotif = rbind(d0,d1)

# multiple motifs - no gap
d2 = data.frame(region = peaksFinal$region,
                pSum = peaksFinal$pSum,
                count = stri_count_regex(str = seqSet,
                                         pattern = "GAAGAA"),
                pattern = "GAAGAA")
d3 = data.frame(region = peaksFinal$region,
                pSum = peaksFinal$pSum,
                count = stri_count_regex(str = seqSet,
                                         pattern = "GAAGAAGAA"),
                pattern = "GAAGAAGAA")
dfNoGap = rbind(d2,d3)

# multiple motifs - with single gaps
d4 = data.frame(region = peaksFinal$region,
                pSum = peaksFinal$pSum,
                count = stri_count_regex(str = seqSet,
                                         pattern = "GAA.GAA"),
                pattern = "GAA.GAA")
d5 = data.frame(region = peaksFinal$region,
                pSum = peaksFinal$pSum,
                count = stri_count_regex(str = seqSet,
                                         pattern = "GAA.GAA.GAA"),
                pattern = "GAA.GAA.GAA")
dfSingleGap = rbind(d4,d5)

# multiple motifs - with double gaps
d6 = data.frame(region = peaksFinal$region,
                pSum = peaksFinal$pSum,
                count = stri_count_regex(str = seqSet,
                                         pattern = "GAA..GAA"),
                pattern = "GAA..GAA")
d7 = data.frame(region = peaksFinal$region,
                pSum = peaksFinal$pSum,
                count = stri_count_regex(str = seqSet,
                                         pattern = "GAA..GAA..GAA"),
                pattern = "GAA..GAA..GAA")
dfDoubleGap = rbind(d6,d7)
```

### Listing 5.4: Functions for the kmer counting

```
# vectorize base R seq function
v.seq <- Vectorize(seq.default, vectorize.args = c("from", "to"))

#' counts frequency of kmers over the given object
#'
#' @param bs GRanges-Object (range to count in)
#' @param kmerSize number (width of the kmer)
#' @return matrix (with nrows = length(bs) and ncols = length(kmerSize))
#' @example countKmerFreq(peaks, 5)
countKmerFreq <- function(bs, kmerSize){
  intBs = bs
  frameSize = unique(width(intBs))
  # extract sequences per region
  seq = as.character(RNAStringSet(getSeq(Hsapiens, intBs)))
  # compute all possible pentamers of the given size
  kmers = names(oligonucleotideFrequency(RNAString("AA"),kmerSize))

  # get start end position of every kmer in every sequence
  # returns a list of lists
  countsKmer = lapply(kmers, function(x){
    loc = stri_locate_all(seq, regex = x, omit_no_match = T)
  })

  # reformat start end postions into a matrix format
  # each line of the matrix corresponds to the sequence of one binding site
  # each entry corresponds to the number of times a kmer was found at the respective position
  kmerMatrix = lapply(countsKmer, function(k){
    currK = k[!isEmpty(k)]
    if (length(currK) > 0) {
      currKPos = sapply(currK, function(x){
        y = as.data.frame(x)
        v = rep(0, frameSize)
        repl = v.seq(y$start, y$end)
        v[as.numeric(repl)] = 1
        return(v)
      })
      currKPos = rowSums(currKPos)
    }
    if (length(currK) == 0) {
      currKPos = rep(0,frameSize)
    }
    return(currKPos)
  })

  # format final matrix
  kmerMatrix = t(do.call("cbind",kmerMatrix))
  rownames(kmerMatrix) = kmers
  kmerMatrix = kmerMatrix / length(intBs)
  return(kmerMatrix)
}
```

## Alternative splicing of SRSF6

The following script holds code for the processing of the AS events called by rMATs. The rMATs output is transferred to a GenomicRanges representation for internal usage. The GenomicRanges object is extended to a GenomicRangesList to capture also the location of the flanking up- and down-stream exons of a given exon skipping event.

**Listing 5.5: Processing of AS events.**

```
### ==============================================================
### Load required packages
### --------------------------------------------------------------
library(GenomicRanges)
library(GenomicFeatures)

### ==============================================================
### Convert rMATS results into GRanges-Objects
### --------------------------------------------------------------
se = read.table("./data/rMATs_output.txt", header = TRUE)
# create central object
events = GRanges(
  seqnames = se$chr,
  ranges = IRanges(start = se$exonStart_0base + 1,
                   end = se$exonEnd),
  strand = se$strand,
  geneID = se$GeneID,
  geneSymbol = se$geneSymbol,
  ijc_s1 = se$IJC_SAMPLE_1 %>% as.character,
  sjc_s1 = se$SJC_SAMPLE_1 %>% as.character,
  ijc_s2 = se$IJC_SAMPLE_2 %>% as.character,
  sjc_s2 = se$SJC_SAMPLE_2 %>% as.character,
  pval = se$PValue,
  fdr = se$FDR,
  incDiff = se$IncLevelDifference,
  incLvl_s1 = se$IncLevel1,
  incLvl_s2 = se$IncLevel2
)
# set tracing IDs
names(events) = se$ID
# summarize counts over all replicates
events = combineReplicateJunctionCounts(events)


### ==============================================================
### Convert rMATS results into GRanges-Objects
### ---------------------------------------------------------------
# create central objects with center,up and downstream ranges
eventsCenter = GRanges(
  seqnames = se$chr,
  ranges = IRanges(start = se$exonStart_0base + 1,
                   end = se$exonEnd),
  strand = se$strand,
  geneID = se$GeneID,
  geneSymbol = se$geneSymbol,
  ijc_s1 = se$IJC_SAMPLE_1 %>% as.character,
  sjc_s1 = se$SJC_SAMPLE_1 %>% as.character,
  ijc_s2 = se$IJC_SAMPLE_2 %>% as.character,
  sjc_s2 = se$SJC_SAMPLE_2 %>% as.character,
  pval = se$PValue,
  fdr = se$FDR,
  incDiff = se$IncLevelDifference,
  incLvl_s1 = se$IncLevel1,
  incLvl_s2 = se$IncLevel2,
  id = se$ID,
  location = "center"
)
eventsUpstream = GRanges(
  seqnames = se$chr,
  ranges = IRanges(start = se$upstreamES + 1,
                   end = se$upstreamEE),
  strand = se$strand,
  geneID = se$GeneID,
  geneSymbol = se$geneSymbol,
  ijc_s1 = se$IJC_SAMPLE_1 %>% as.character,
  sjc_s1 = se$SJC_SAMPLE_1 %>% as.character,
  ijc_s2 = se$IJC_SAMPLE_2 %>% as.character,
  sjc_s2 = se$SJC_SAMPLE_2 %>% as.character,
  pval = se$PValue,
  fdr = se$FDR,
  incDiff = se$IncLevelDifference,
  incLvl_s1 = se$IncLevel1,
  incLvl_s2 = se$IncLevel2,
  id = se$ID,
  location = "upstream"
)
eventsDownstream = GRanges(
  seqnames = se$chr,
  ranges = IRanges(start = se$downstreamES + 1,
                   end = se$downstreamEE),
  strand = se$strand,
  geneID = se$GeneID,
```

```r
  geneSymbol = se$geneSymbol,
  ijc_s1 = se$IJC_SAMPLE_1 %>% as.character,
  sjc_s1 = se$SJC_SAMPLE_1 %>% as.character,
  ijc_s2 = se$IJC_SAMPLE_2 %>% as.character,
  sjc_s2 = se$SJC_SAMPLE_2 %>% as.character,
  pval = se$PValue,
  fdr = se$FDR,
  incDiff = se$IncLevelDifference,
  incLvl_s1 = se$IncLevel1,
  incLvl_s2 = se$IncLevel2,
  id = se$ID,
  location = "downstream"
)


### ================================================================
### Combine flanking exons and central exons into a GRanges-List
### ----------------------------------------------------------------
### Each list entry represents a single AS event. Each list has exactly three entries (upstream exon, central exon, downstrem exon).

# handle plus strand
centerPlus = as(subset(eventsCenter, strand == "+"), "GRangesList")
upstreamPlus = as(subset(eventsUpstream, strand == "+"), "GRangesList")
downstreamPlus = as(subset(eventsDownstream, strand == "+"), "GRangesList")
# combine lists and set ids
exonsPlus = pc(centerPlus, upstreamPlus, downstreamPlus)
exonsPlus = as(lapply(1:length(exonsPlus), function(x){
  g = unlist(exonsPlus[x])
  g$id = x
  g$pos = c("center", "upstream", "downstream")
  return(g)
}), "GRangesList")
# add tracing id
names(exonsPlus) = subset(eventsCenter, strand == "+")$id

# handle minus strand
centerMinus = as(subset(eventsCenter, strand == "-"), "GRangesList")
upstreamMinus = as(subset(eventsUpstream, strand == "-"), "GRangesList")
```

```
downstreamMinus = as(subset(eventsDownstream, strand == "-"), "GRangesList")
# combine lists and set ids
exonsMinus = pc(centerMinus, upstreamMinus, downstreamMinus)
exonsMinus = as(lapply(1:length(exonsMinus), function(x){
  g = unlist(exonsMinus[x])
  g$id = x
  g$pos = c("center", "downstream", "upstream")
  return(g)
}), "GRangesList")
# add tracing id
names(exonsMinus) = subset(eventsCenter, strand == "-")$id

# combine strands
exonSet = c(exonsMinus, exonsPlus)
exonSet = sort(exonSet)
exonSet = exonSet[order(names(exonSet))]
# store results
rtracklayer::export(exonSet, "./data/exonSet.bed", format = "BED")
rtracklayer::export(events, "./data/events.bed", format = "BED")


### ==============================================================
### Filter for overlaping events
### --------------------------------------------------------------
# remove events with low junction read counts
eventsFilter0 = events
sel = apply(as.matrix(mcols(eventsFilter0)[c(3:6)]), 1,
            function(x) length(x[x > 10]) >= 3)
eventsFilter1 = events[sel,]
eventsFilter2 = eventsFilter1
# iteratively remove overlapping events
while (max(countOverlaps(eventsFilter2)) > 1) {
  print(paste0("start ", max(countOverlaps(eventsFilter2))))
  ols = findOverlaps(eventsFilter2, drop.self = F) %>%
    as.data.frame()
  idx = sapply(1:nrow(ols), function(x){
    ols$subjectHits[ols$queryHits == x]
  })
  idx = idx[!duplicated(idx)]
  eventsFilter2 = sapply(1:length(idx), function(x){
    curr.idx = unlist(idx[x])
    curr.ranges = eventsFilter2[curr.idx]
    if (length(curr.ranges) > 1) {
      # select by lowest fdr exon
      curr.ranges = curr.ranges[which.min(curr.ranges$fdr)]
    }
    return(curr.ranges)
  })
  # update output
  eventsFilter2 = unlist(GRangesList(eventsFilter2))
  eventsFilter2 = eventsFilter2[!duplicated(eventsFilter2)]
  print(paste0("end ", max(countOverlaps(eventsFilter2))))
}
# store results
rtracklayer::export(eventsFilter2,
                    con = "./data/eventsFilter2.bed",
                    format = "BED")

# remove overlapping results from exon set as well
exonSet = exonSet[names(exonSet) %in% names(eventsFilter2)]


### ===============================================================
### Filter for significantly AS exons
### ---------------------------------------------------------------
# apply significance thresholds
eventsFilter2$meanLog2Exp =
  log2(
    rowMeans(
      as.matrix(
        mcols(eventsFilter2[,c(3:6)]))))
eventsReg = subset(eventsFilter2, abs(incDiff) > 0.05 &
                     fdr < 0.05 &
                     meanLog2Exp > 5)
eventsNot = eventsFilter2[!eventsFilter2 %in% eventsReg]

# export and save processed AS events
rtracklayer::export(eventsReg,
                    con = "./data/eventsReg.bed",
                    format = "BED")
save(eventsReg, file = "./data/eventsReg.rda")
save(eventsNot, file = "./data/eventsNot.rda")
save(exonSet, file = "./data/exonSet.rda")
```

**Listing 5.6: Functions for AS event processing.**

```
#' summarize replicate counts for junction and inclusion levels
#'
#' @param x Granges object
#' @return GRanges object with modified meta columns
#' @example combineReplicateJunctionCounts(events)
```

```
combineReplicateJunctionCounts <- function(x){
  x$ijc_s1 = x$ijc_s1 %>%
    strsplit(., split = ",") %>%
    as.data.frame() %>%
    apply(., 1, as.numeric) %>%
    rowSums
  x$sjc_s1 = x$sjc_s1 %>%
    strsplit(., split = ",") %>%
    as.data.frame() %>%
    apply(., 1, as.numeric) %>%
    rowSums
  x$ijc_s2 = x$ijc_s2 %>%
    strsplit(., split = ",") %>%
    as.data.frame() %>%
    apply(., 1, as.numeric) %>%
    rowSums
  x$sjc_s2 = x$sjc_s2 %>%
    strsplit(., split = ",") %>%
    as.data.frame() %>%
    apply(., 1, as.numeric) %>%
    rowSums
  x$incLvl_s1 = as.character(x$incLvl_s1) %>%
    strsplit(., split = ",") %>%
    as.data.frame() %>%
    apply(., 1, as.numeric) %>%
    rowMeans
  x$incLvl_s2 = as.character(x$incLvl_s2) %>%
    strsplit(., split = ",") %>%
    as.data.frame() %>%
    apply(., 1, as.numeric) %>%
    rowMeans
  return(x)
}
```

## SRSF6 RNA-splicing map

The following script holds the code to generate the RNA-splicing map. The GenomicRangesList from the above script is required as input. The splice sites of each exons are used as central point to span a symmetric window in which crosslinked position are counted in. The result is a matrix, which is plotted as heatmap as well as a summary profile.

**Listing 5.7: Functions for the RNA splicing map.**

```r
### ==============================================================
### Load required packages
### ==============================================================
library(ggplot2)
library(GenomicFeatures)
library(GenomicRanges)
library(matrixStats)

### ==============================================================
### Load required datasets
### ==============================================================
# Load iCLIP data as single nucleotide crosslink events
clp.plus = import("./data/clip_merge_plus.bw",
                  format = "BigWig", as = "Rle")
clp.minus = import("./data/clip_merge_minus.bw",
                   format = "BigWig", as = "Rle")
# Load AS events
load(file = "./data/eventsReg.rda")
load(file = "./data/eventsNot.rda")
load(file = "./data/exonSet.rda")


### ==============================================================
### Compute RNA-Map for Up-regualted exons
### ==============================================================
# select upregualted exons and exon sets
exnUp = eventsReg[eventsReg$incDiff > 0]
mcols(exnUp)$meanInc = rowMeans(as.matrix(mcols(exnUp[,10:11])))
exnUpFlankUpstream = exonSet[names(exonSet) %in% names(exnUp)] %>%
  unlist %>% subset(., pos == "upstream")
exnUpFlankDownstream = exonSet[names(exonSet) %in% names(exnUp)]
    %>%
  unlist %>% subset(., pos == "downstream")
# sort and combine regualted exon regions
exnUp = sortSeqlevels(exnUp)
exnUp = sort(exnUp)
```

```
exnUpFlankUpstream = sortSeqlevels(exnUpFlankUpstream)
exnUpFlankUpstream = sort(exnUpFlankUpstream)
exnUpFlankDownstream = sortSeqlevels(exnUpFlankDownstream)
exnUpFlankDownstream = sort(exnUpFlankDownstream)
mcols(exnUpFlankUpstream) = mcols(exnUp)
mcols(exnUpFlankDownstream) = mcols(exnUp)

# select background exons and exon sets
exnNot = eventsNot
mcols(exnNot)$meanInc = rowMeans(as.matrix(mcols(exnNot[,10:11])))
exnNotFlankUpstream = exonSet[names(exonSet) %in% names(eventsNot)] %>%
  unlist %>% subset(., pos == "upstream")
exnNotFlankDownstream = exonSet[names(exonSet) %in% names(eventsNot)] %>%
  unlist %>% subset(., pos == "downstream")
# sort and combine background exon regions
exnNot = sortSeqlevels(exnNot)
exnNot = sort(exnNot)
exnNotFlankUpstream = sortSeqlevels(exnNotFlankUpstream)
exnNotFlankUpstream = sort(exnNotFlankUpstream)
exnNotFlankDownstream = sortSeqlevels(exnNotFlankDownstream)
exnNotFlankDownstream = sort(exnNotFlankDownstream)
mcols(exnNotFlankUpstream) = mcols(exnNot)
mcols(exnNotFlankDownstream) = mcols(exnNot)

# iteratively sample background count frequency from PSI match background distribution
sampleBcCenter = sampleBackground(exn = exnUp, bc = exnNot)
sampleBcFlankUpstream = sampleBackground(exn = exnUpFlankUpstream,
                                         bc = exnNotFlankUpstream)
sampleBcFlankDownstream = sampleBackground(exn = exnUpFlankDownstream,
                                           bc = exnNotFlankDownstream)
# summarize the background coverage
df1 = data.frame(mean = rowMeans(sampleBcCenter$matr1),
                 sd = rowSds(sampleBcCenter$matr1),
                 position = "ss5")
df2 = data.frame(mean = rowMeans(sampleBcCenter$matr2),
                 sd = rowSds(sampleBcCenter$matr2),
                 position = "ss3")
df3 = data.frame(mean = rowMeans(sampleBcFlankUpstream$matr1),
                 sd = rowSds(sampleBcFlankUpstream$matr1),
                 position = "up-ss5")
df4 = data.frame(mean = rowMeans(sampleBcFlankDownstream$matr2),
                 sd = rowSds(sampleBcFlankDownstream$matr2),
                 position = "down-ss3")
df.bc = rbind(df1,df2,df3,df4)
df.bc$pos = -100:100

# compute coverage for upregulated exons
df1 = makeCounts(obj = exnUp, signal.p = clp.plus, signal.m = clp.minus,
                 range = 101, which = "ss5")
df1 = cbind.data.frame(df1, calcPval(bc.matrix = sampleBcCenter$matr1,
                                     test.vector = df1$counts))
df2 = makeCounts(obj = exnUp, signal.p = clp.plus, signal.m = clp.minus,
                 range = 101, which = "ss3")
df2 = cbind.data.frame(df2, calcPval(bc.matrix = sampleBcCenter$matr2,
                                     test.vector = df2$counts))
df3 = makeCounts(obj = exnUpFlankUpstream,
                 signal.p = clp.plus,
                 signal.m = clp.minus,
                 range = 101, which = "ss5")
df3 = cbind.data.frame(df3, calcPval(bc.matrix = sampleBcFlankUpstream$matr1,
                                     test.vector = df3$counts))
df3$position = "up-ss5"
df4 = makeCounts(obj = exnUpFlankDownstream,
                 signal.p = clp.plus,
                 signal.m = clp.minus,
                 range = 101, which = "ss3")
df4 = cbind.data.frame(df4,
                       calcPval(bc.matrix = sampleBcFlankDownstream$matr2,
                                test.vector = df4$counts))
df4$position = "down-ss3"
df.real = rbind(df1,df2,df3,df4)

# adjust zscores to a readable scale
df.real$zscore[(df.real$zscore) > 6] = 6
df.real$zscore[df.real$zscore < 0] = 0
df.zscore = subset(df.real, padj < 0.05)

# define plotting order
df.real$position = factor(df.real$position,
                          levels = c("up-ss5", "ss3", "ss5", "down-ss3"))
```

```r
df.bc$position = factor(df.bc$position,
                        levels = c("up-ss5", "ss3", "ss5", "down-ss3"))
df.zscore$position = factor(df.zscore$position,
                            levels = c("up-ss5", "ss3", "ss5", "down-ss3"))

# adjust computed results for plotting
upDfReal = cbind.data.frame(df.real,
                            type = "signal",
                            regulation = "up")
upDfBc = cbind.data.frame(df.bc,
                          type = "signal",
                          regulation = "up")
upDfZscore = cbind.data.frame(df.zscore,
                              type = "zscore",
                              regulation = "up")


### ==================================================================
### Compute RNA-Map for Down-regualted exons
### ==================================================================
# select upregualted exons and exon sets
exnDown = eventsReg[eventsReg$incDiff < 0]
mcols(exnDown)$meanInc = rowMeans(as.matrix(mcols(exnDown[,10:11])))
exnDownFlankUpstream = exonSet[names(exonSet) %in% names(exnDown)] %>%
  unlist %>% subset(., pos == "upstream")
exnDownFlankDownstream = exonSet[names(exonSet) %in% names(exnDown)] %>%
  unlist %>% subset(., pos == "downstream")
# sort and combine regualted exon regions
exnDown = sortSeqlevels(exnDown)
exnDown = sort(exnDown)
exnDownFlankUpstream = sortSeqlevels(exnDownFlankUpstream)
exnDownFlankUpstream = sort(exnDownFlankUpstream)
exnDownFlankDownstream = sortSeqlevels(exnDownFlankDownstream)
exnDownFlankDownstream = sort(exnDownFlankDownstream)
mcols(exnDownFlankUpstream) = mcols(exnDown)
mcols(exnDownFlankDownstream) = mcols(exnDown)


# iteratively sample background count frequency from PSI match background distribution
sampleBcCenter = sampleBackground(exn = exnDown, bc = exnNot)
```

```
sampleBcFlankUpstream = sampleBackground(exn = exnDownFlankUpstream,
                                         bc = exnNotFlankUpstream)
sampleBcFlankDownstream = sampleBackground(exn = exnDownFlankDownstream,
                                           bc = exnNotFlankDownstream)
# summarize the background coverage
df1 = data.frame(mean = rowMeans(sampleBcCenter$matr1),
                 sd = rowSds(sampleBcCenter$matr1),
                 position = "ss5")
df2 = data.frame(mean = rowMeans(sampleBcCenter$matr2),
                 sd = rowSds(sampleBcCenter$matr2),
                 position = "ss3")
df3 = data.frame(mean = rowMeans(sampleBcFlankUpstream$matr1),
                 sd = rowSds(sampleBcFlankUpstream$matr1),
                 position = "up-ss5")
df4 = data.frame(mean = rowMeans(sampleBcFlankDownstream$matr2),
                 sd = rowSds(sampleBcFlankDownstream$matr2),
                 position = "down-ss3")
df.bc = rbind(df1,df2,df3,df4)
df.bc$pos = -100:100

# compute coverage for upregulated exons
df1 = makeCounts(obj = exnDown, signal.p = clp.plus, signal.m = clp.minus,
                 range = 101, which = "ss5")
df1 = cbind.data.frame(df1, calcPval(bc.matrix = sampleBcCenter$matr1,
                                     test.vector = df1$counts))
df2 = makeCounts(obj = exnDown,
                 signal.p = clp.plus,
                 signal.m = clp.minus,
                 range = 101, which = "ss3")
df2 = cbind.data.frame(df2, calcPval(bc.matrix = sampleBcCenter$matr2,
                                     test.vector = df2$counts))
df3 = makeCounts(obj = exnDownFlankUpstream,
                 signal.p = clp.plus,
                 signal.m = clp.minus,
                 range = 101, which = "ss5")
df3 = cbind.data.frame(df3, calcPval(bc.matrix = sampleBcFlankUpstream$matr1,
                                     test.vector = df3$counts))
df3$position = "up-ss5"
df4 = makeCounts(obj = exnDownFlankDownstream,
                 signal.p = clp.plus,
                 signal.m = clp.minus,
                 range = 101, which = "ss3")
df4 = cbind.data.frame(df4,
                       calcPval(bc.matrix = sampleBcFlankDownstream$matr2,
                                test.vector = df4$counts))
df4$position = "down-ss3"
df.real = rbind(df1,df2,df3,df4)

# adjust zscores to a readable scale
df.real$zscore[(df.real$zscore) > 6] = 6
df.real$zscore[df.real$zscore < 0] = 0
df.zscore = subset(df.real, padj < 0.05)

# define plotting order
df.real$position = factor(df.real$position,
                          levels = c("up-ss5", "ss3", "ss5", "down-ss3"))
df.bc$position = factor(df.bc$position,
                        levels = c("up-ss5", "ss3", "ss5", "down-ss3"))
df.zscore$position = factor(df.zscore$position,
                            levels = c("up-ss5", "ss3", "ss5", "down-ss3"))

# adjust computed results for plotting
downDfReal = cbind.data.frame(df.real,
                              type = "signal",
                              regulation = "down")
downDfBc = cbind.data.frame(df.bc,
                            type = "signal",
                            regulation = "down")
downDfZscore = cbind.data.frame(df.zscore,
                                type = "zscore",
                                regulation = "down")

### ==================================================================
### Plot the final map
### ==================================================================
p = ggplot() +
  geom_ribbon(data = downDfBc,
              aes(x = pos, ymin = mean - sd, ymax = mean + sd),
              alpha = .6, fill = "LightSteelBLue") +
  geom_line(data = downDfBc,
            aes(x = pos, y = mean),
```

```r
          color = "LightSteelBLue", size = .7) +
geom_line(data = downDfReal,
          aes(x = pos, y = counts), size = .7,
          color = "SteelBLue") +
geom_ribbon(data = upDfBc,
            aes(x = pos, ymin = mean - sd, ymax = mean + sd),
            alpha = .6, fill = "#e2be9c") +
geom_line(data = upDfBc,
          aes(x = pos, y = mean),
          color = "#e2be9c", size = .7) +
geom_line(data = upDfReal,
          aes(x = pos, y = counts),
          size = .7, color = "Chocolate") +
facet_wrap(~regulation+position, ncol = 4) +
xlab("Position in alternatively spliced exon") +
ylab("mean counts per nucleotide") +
geom_point(data = upDfZscore,
           aes(x = pos, y = -.1, color = zscore),
           size = 3, shape = 73) +
geom_point(data = downDfZscore,
           aes(x = pos, y = -.1, color = zscore),
           size = 3, shape = 73) +
scale_color_gradient(low = "white", high = "#2E8B57") +
theme_bw() +
ggtitle("RNA-Maps") +
geom_hline(yintercept = 0, color = "darkgrey")
```

# Bibliography

Alvelos, M. I., Brüggemann, M., Sutandy, F. R., Juan-Mateu, J., Colli, M. L., Busch, A., Lopes, M., Castela, Â., Aartsma-Rus, A., König, J., et al. (2020). The RNA-binding profile of the splicing factor SRSF6 in immortalized human pancreatic $\beta$-cells. *Life science alliance*, 4(3).

Änkö, M.-L. (2014). Regulation of gene expression programmes by serinearginine rich splicing factors. *Seminars in cell & developmental biology*, 32:11 – 21.

Änkö, M.-L., Morales, L., Henry, I., Beyer, A., and Neugebauer, K. M. (2010). Global analysis reveals SRp20 and SRp75-specific mRNPs in cycling and neural cells. *Nature structural & molecular biology*, 17(8):962.

Änkö, M.-L., Müller-McNicoll, M., Brandl, H., Curk, T., Gorup, C., Henry, I., Ule, J., and Neugebauer, K. M. (2012). The RNA-binding landscapes of two SR proteins reveal unique functions and binding to diverse RNA classes. *Genome biology*, 13(3):R17.

Auweter, S. D., Oberstrass, F. C., and Allain, F. H.-T. (2006). Sequence-specific binding of single-stranded RNA: is there a code for recognition? *Nucleic acids research*, 34(17):4943–4959.

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

Baralle, D. and Buratti, E. (2017). RNA splicing in human disease and in the clinic. *Clinical science*, 131(5):355–368.

Barbosa-Morais, N. L., Irimia, M., Pan, Q., Xiong, H. Y., Gueroussov, S., Lee, L. J., Slobodeniuc, V., Kutter, C., Watt, S., Çolak, R., et al. (2012). The evolutionary landscape of alternative splicing in vertebrate species. *Science*, 338(6114):1587–1593.

Barrett, J. C., Clayton, D. G., Concannon, P., Akolkar, B., Cooper, J. D., Erlich, H. A., Julier, C., Morahan, G., Nerup, J., Nierras, C., et al. (2009). Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nature genetics*, 41(6):703.

Baruzzo, G., Hayer, K. E., Kim, E. J., Di Camillo, B., FitzGerald, G. A., and Grant, G. R. (2017). Simulation-based comprehensive benchmarking of RNA-seq aligners. *Nature methods*, 14(2):135.

Beckmann, B. M., Horos, R., Fischer, B., Castello, A., Eichelbaum, K., Alleaume, A.-M., Schwarzl, T., Curk, T., Foehr, S., Huber, W., et al. (2015). The RNA-binding proteomes from yeast to man harbour conserved enigmRBPs. *Nature communications*, 6(1):1–9.

Berglund, J. A., Abovich, N., and Rosbash, M. (1998). A cooperative interaction between U2AF65 and mBBP/SF1 facilitates branchpoint region recognition. *Genes & development*, 12(6):858–867.

Bicknell, A. A. and Ricci, E. P. (2017). When mRNA translation meets decay. *Biochemical society transactions*, 45(2):339–351.

Blencowe, B. J., Bowman, J. A., McCracken, S., and Rosonina, E. (1999). SR-related proteins and the processing of messenger RNA precursors. *Biochemistry and cell biology*, 77(4):277–291.

Bottini, S., Hamouda-Tekaya, N., Tanasa, B., Zaragosi, L.-E., Grandjean, V., Repetto, E., and Trabucchi, M. (2017). From benchmarking HITS-CLIP peak detection programs to a new method for identification of miRNA-binding sites from Ago2-CLIP data. *Nucleic acids research*, 45(9):e71–e71.

Bottini, S., Pratella, D., Grandjean, V., Repetto, E., and Trabucchi, M. (2018). Recent computational developments on CLIP-seq data analysis and microRNA targeting implications. *Briefings in bioinformatics*, 19(6):1290–1301.

Bourgeois, C. F., Lejeune, F., and Stévenin, J. (2004). Broad specificity of SR (serine/arginine) proteins in the regulation of alternative splicing of pre-messenger RNA. *Progress in nucleic acid research and molecular biology*, 78:37.

Bradfield, J. P., Qu, H.-Q., Wang, K., Zhang, H., Sleiman, P. M., Kim, C. E., Mentch, F. D., Qiu, H., Glessner, J. T., Thomas, K. A., et al. (2011). A genome-wide meta-analysis of six type 1 diabetes cohorts identifies multiple associated loci. *PLoS genetics*, 7(9):e1002293.

Bradley, T., Cook, M. E., and Blanchette, M. (2015). SR proteins control a complex network of RNA-processing events. *RNA*, 21(1):75–92.

Briese, M., Haberman, N., Sibley, C. R., Faraway, R., Elser, A. S., Chakrabarti, A. M., Wang, Z., König, J., Perera, D., Wickramasinghe, V. O., et al. (2019). A systems view of spliceosomal assembly and branchpoints with iCLIP. *Nature structural & molecular biology*, 26(10):930–940.

Brinkman, A., Van Der Flier, S., Kok, E. M., and Dorssers, L. C. (2000). BCAR1, a human homologue of the adapter protein p130Cas, and antiestrogen resistance in breast cancer cells. *Journal of the national cancer institute*, 92(2):112–120.

Buchbender, A., Mutter, H., Sutandy, F. R., Körtel, N., Hänel, H., Busch, A., Ebersberger, S., and König, J. (2020). Improved library preparation with the new iCLIP2 protocol. *Methods*, 178:33–48.

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research*, 47(D1):D1005–D1012.

Burd, C. G. and Dreyfuss, G. (1994). Conserved structures and diversity of functions of RNA-binding proteins. *Science*, 265(5172):615–621.

Busch, A., Brüggemann, M., Ebersberger, S., and Zarnack, K. (2020). iCLIP data analysis: A complete pipeline from sequencing reads to RBP binding sites. *Methods*, 178:49–62.

Cabrera, O., Berman, D. M., Kenyon, N. S., Ricordi, C., Berggren, P.-O., and Caicedo, A. (2006). The unique cytoarchitecture of human pancreatic islets has implications for islet cell function. *Proceedings of the national academy of sciences*, 103(7):2334–2339.

Cáceres, J. F., Screaton, G. R., and Krainer, A. R. (1998). A specific subset of SR proteins shuttles continuously between the nucleus and the cytoplasm. *Genes & development*, 12(1):55–66.

Cáceres, J. F., Stamm, S., Helfman, D. M., and Krainer, A. R. (1994). Regulation of alternative splicing in vivo by overexpression of antagonistic splicing factors. *Science*, 265(5179):1706–1709.

Cao, C., Zhang, Y., Jia, Q., Wang, X., Zheng, Q., Zhang, H., Song, R., Li, Y., Luo, A., Hong, Q., et al. (2019). An exonic splicing enhancer mutation in DUOX2 causes aberrant alternative splicing and severe congenital hypothyroidism in Bama pigs. *Disease models & mechanisms*, 12(1).

Cartegni, L. and Krainer, A. R. (2002). Disruption of an SF2/ASF-dependent exonic splicing enhancer in SMN2 causes spinal muscular atrophy in the absence of SMN1. *Nature genetics*, 30(4):377–384.

Cartegni, L., Wang, J., Zhu, Z., Zhang, M. Q., and Krainer, A. R. (2003). ESEfinder: a web resource to identify exonic splicing enhancers. *Nucleic acids research*, 31(13):3568–3571.

Castello, A., Fischer, B., Frese, C. K., Horos, R., Alleaume, A.-M., Foehr, S., Curk, T., Krijgsveld, J., and Hentze, M. W. (2016). Comprehensive identification of RNA-binding domains in human cells. *Molecular cell*, 63(4):696–710.

Chakrabarti, A. M., Haberman, N., Praznik, A., Luscombe, N. M., and Ule, J. (2018). Data science issues in studying protein-RNA interactions with CLIP technologies. *Annual review of biomedical data science*, 1:235–261.

Chen, B., Yun, J., Kim, M. S., Mendell, J. T., and Xie, Y. (2014). PIPE-CLIP: a comprehensive online tool for CLIP-seq data analysis. *Genome biology*, 15(1):R18.

Chen, J. and Weiss, W. (2015). Alternative splicing in cancer: implications for biology and therapy. *Oncogene*, 34(1):1–14.

Cho, S., Hoang, A., Sinha, R., Zhong, X.-Y., Fu, X.-D., Krainer, A. R., and Ghosh, G. (2011). Interaction between the RNA binding domains of Ser-Arg splicing factor 1 and U1-70K snRNP protein determines early spliceosome assembly. *Proceedings of the national academy of sciences*, 108(20):8233–8238.

Cho, Y. S., Chen, C.-H., Hu, C., Long, J., Ong, R. T. H., Sim, X., Takeuchi, F., Wu, Y., Go, M. J., Yamauchi, T., et al. (2012). Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east Asians. *Nature genetics*, 44(1):67.

Chung, J. H. and Bunz, F. (2010). CDK2 is required for p53-independent G2/M checkpoint control. *PLoS genetics*, 6(2):e1000863.

Corcoran, D. L., Georgiev, S., Mukherjee, N., Gottwein, E., Skalsky, R. L., Keene, J. D., and Ohler, U. (2011). PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data. *Genome biology*, 12(8):1–16.

Corley, M., Burns, M. C., and Yeo, G. W. (2020). How RNA-Binding Proteins Interact with RNA: Molecules and Mechanisms. *Molecular cell*, 78(1):9–29.

Cramer, P., Bushnell, D. A., and Kornberg, R. D. (2001). Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution. *Science*, 292(5523):1863–1876.

Crick, F. H. (1958). On protein synthesis. In *Symp soc exp biol*, volume 12, page 8.

Cunha, D. A., Igoillo-Esteve, M., Gurzov, E. N., Germano, C. M., Naamane, N., Marhfour, I., Fukaya, M., Vanderwinden, J.-M., Gysemans, C., Mathieu, C., et al. (2012). Death protein 5 and p53-upregulated modulator of apoptosis mediate the endoplasmic reticulum stress-mitochondrial dialog triggering lipotoxic rodent and human $\beta$-cell apoptosis. *Diabetes*, 61(11):2763–2775.

David, C. J., Chen, M., Assanah, M., Canoll, P., and Manley, J. L. (2010). HnRNP proteins controlled by c-Myc deregulate pyruvate kinase mRNA splicing in cancer. *Nature*, 463(7279):364–368.

David, C. J. and Manley, J. L. (2010). Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged. *Genes & development*, 24(21):2343–2364.

De, S. and Gorospe, M. (2017). Bioinformatic tools for analysis of CLIP ribonucleoprotein data. *Wiley interdisciplinary reviews: RNA*, 8(4):e1404.

Desmet, F.-O., Hamroun, D., Lalande, M., Collod-Béroud, G., Claustres, M., and Béroud, C. (2009). Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic acids research*, 37(9):e67–e67.

Dimitri, P., Warner, J. T., Minton, J., Patch, A., Ellard, S., Hattersley, A., Barr, S., Hawkes, D., Wales, J., and Gregory, J. (2011). Novel GLIS3 mutations demonstrate an extended multisystem phenotype. *European journal of endocrinology*, 164(3):437–443.

Dinarello, C. A. (2000). Proinflammatory cytokines. *Chest*, 118(2):503–508.

Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., et al. (2012). Landscape of transcription in human cells. *Nature*, 489(7414):101–108.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21.

Dodt, M., Roehr, J. T., Ahmed, R., and Dieterich, C. (2012). FLEXBAR: flexible barcode and adapter processing for next-generation sequencing platforms. *Biology*, 1(3):895–905.

Dooley, J., Tian, L., Schonefeldt, S., Delghingaro-Augusto, V., Garcia-Perez, J. E., Pasciuto, E., Di Marino, D., Carr, E. J., Oskolkov, N., Lyssenko, V., et al. (2016). Genetic predisposition for beta cell fragility underlies type 1 and type 2 diabetes. *Nature genetics*, 48(5):519–527.

Drewe-Boss, P., Wessels, H.-H., and Ohler, U. (2018). omniCLIP: probabilistic identification of protein-RNA interactions from CLIP-seq data. *Genome biology*, 19(1):183.

Dreyfuss, G., Kim, V. N., and Kataoka, N. (2002). Messenger-RNA-binding proteins and the messages they carry. *Nature reviews molecular cell biology*, 3(3):195–205.

Dujardin, G., Lafaille, C., Petrillo, E., Buggiano, V., Acuña, L. I. G., Fiszbein, A., Herz, M. A. G., Moreno, N. N., Muñoz, M. J., Alló, M., et al. (2013). Transcriptional elongation and alternative splicing. *Biochimica et biophysica acta (BBA)-gene regulatory mechanisms*, 1829(1):134–140.

Dupuis, J., Langenberg, C., Prokopenko, I., Saxena, R., Soranzo, N., Jackson, A. U., Wheeler, E., Glazer, N. L., Bouatia-Naji, N., Gloyn, A. L., et al. (2010). New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nature genetics*, 42(2):105–116.

Edlich, F. (2018). BCL-2 proteins and apoptosis: Recent insights and unknowns. *Biochemical and biophysical research communications*, 500(1):26–34.

Eizirik, D. L., Pasquali, L., and Cnop, M. (2020). Pancreatic $\beta$-cells in type 1 and type 2 diabetes mellitus: different pathways to failure. *Nature reviews endocrinology*, pages 1–14.

Eizirik, D. L., Sammeth, M., Bouckenooghe, T., Bottu, G., Sisino, G., Igoillo-Esteve, M., Ortis, F., Santin, I., Colli, M. L., Barthson, J., et al. (2012). The human pancreatic islet transcriptome: expression of candidate genes for type 1 diabetes and the impact of pro-inflammatory cytokines. *PLoS genetics*, 8(3):e1002552.

Eperon, I., Ireland, D., Smith, R., Mayeda, A., and Krainer, A. (1993). Pathways for selection of 5' splice sites by U1 snRNPs and SF2/ASF. *The EMBO journal*, 12(9):3607–3617.

Foltz, D. R., Jansen, L. E., Black, B. E., Bailey, A. O., Yates, J. R., and Cleveland, D. W. (2006). The human CENP-A centromeric nucleosome-associated complex. *Nature cell biology*, 8(5):458–469.

Forrest, A. R., Kawaji, H., Rehli, M., Baillie, J. K., De Hoon, M. J., Haberle, V., Lassmann, T., Kulakovskiy, I. V., Lizio, M., Itoh, M., et al. (2014). A promoter-level mammalian expression atlas. *Nature*, 507(7493):462–470.

Frankish, A., Diekhans, M., Ferreira, A.-M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J. M., Sisu, C., Wright, J., Armstrong, J., et al. (2019). GENCODE reference annotation for the human and mouse genomes. *Nucleic acids research*, 47(D1):D766–D773.

Frendewey, D. and Keller, W. (1985). Stepwise assembly of a pre-mRNA splicing complex requires U-snRNPs and specific intron sequences. *Cell*, 42(1):355–367.

Fu, N. Y., Sukumaran, S. K., Kerk, S. Y., and Victor, C. Y. (2009). Bax$\beta$: a constitutively active human Bax isoform that is under tight regulatory control by the proteasomal degradation mechanism. *Molecular cell*, 33(1):15–29.

Fu, X.-D. and Maniatis, T. (1992). The 35-kDa mammalian splicing factor SC35 mediates specific interactions between U1 and U2 small nuclear ribonucleoprotein particles at the 3'splice site. *Proceedings of the national academy of sciences*, 89(5):1725–1729.

Galante, P. A. F., Sakabe, N. J., Kirschbaum-Slager, N., and de Souza, S. J. (2004). Detection and evaluation of intron retention events in the human transcriptome. *RNA*, 10(5):757–765.

Ghanbari, M. and Ohler, U. (2020). Deep neural networks for interpreting RNA-binding protein target preferences. *Genome research*, 30(2):214–226.

Gonatopoulos-Pournatzis, T. and Cowling, V. H. (2014). Cap-binding complex (CBC). *Biochemical journal*, 457(2):231–242.

Goode, E. L., Fridley, B. L., Vierkant, R. A., Cunningham, J. M., Phelan, C. M., Anderson, S., Rider, D. N., White, K. L., Pankratz, V. S., Song, H., et al. (2009). Candidate gene analysis using imputed genotypes: cell cycle single-nucleotide polymorphisms and ovarian cancer risk. *Cancer epidemiology and prevention biomarkers*, 18(3):935–944.

Grant, C. E., Bailey, T. L., and Noble, W. S. (2011). FIMO: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018.

Graveley, B. R., Hertel, K. J., and Maniatis, T. (1999). SR proteins are 'locators' of the RNA splicing machinery. *Current biology*, 9(1):R6–R7.

Gu, Z., Eils, R., and Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, 32(18):2847–2849.

Gurzov, E. N. and Eizirik, D. L. (2011). Bcl-2 proteins in diabetes: mitochondrial pathways of $\beta$-cell death and dysfunction. *Trends in cell biology*, 21(7):424–431.

Haberman, N., Huppertz, I., Attig, J., König, J., Wang, Z., Hauer, C., Hentze, M. W., Kulozik, A. E., Le Hir, H., Curk, T., et al. (2017). Insights into the design and interpretation of iCLIP experiments. *Genome biology*, 18(1):1–21.

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano Jr, M., Jungkamp, A.-C., Munschauer, M., et al. (2010). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, 141(1):129–141.

Hakonarson, H. and Grant, S. F. (2011). Genome-wide association studies (GWAS): impact on elucidating the aetiology of diabetes. *Diabetes/ metabolism research and reviews*, 27(7):685–696.

Han, J., Ding, J.-H., Byeon, C. W., Kim, J. H., Hertel, K. J., Jeong, S., and Fu, X.-D. (2011). SR proteins induce alternative exon skipping through their activities on the flanking constitutive exons. *Molecular and cellular biology*, 31(4):793–802.

Havlioglu, N., Wang, J., Fushimi, K., Vibranovski, M. D., Kan, Z., Gish, W., Fedorov, A., Long, M., and Wu, J. Y. (2007). An intronic signal for alternative splicing in the human genome. *PloS one*, 2(11):e1246.

Hentze, M. W., Castello, A., Schwarzl, T., and Preiss, T. (2018). A brave new world of RNA-binding proteins. *Nature reviews molecular cell biology*, 19(5):327.

Hildebrandt, A., Brüggemann, M., Rücklé, C., Boerner, S., Heidelberger, J. B., Busch, A., Hänel, H., Voigt, A., Möckel, M. M., Ebersberger, S., et al. (2019). The RNA-binding ubiquitin ligase MKRN1 functions in ribosome-associated quality control of poly (A) translation. *Genome biology*, 20(1):1–20.

Howald, C., Tanzer, A., Chrast, J., Kokocinski, F., Derrien, T., Walters, N., Gonzalez, J. M., Frankish, A., Aken, B. L., Hourlier, T., et al. (2012). Combining RT-PCR-seq and RNA-seq to catalog all genic elements encoded in the human genome. *Genome research*, 22(9):1698–1710.

Huber, W., Von Heydebreck, A., Sültmann, H., Poustka, A., and Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18(suppl_1):S96–S104.

Huppertz, I., Attig, J., DAmbrogio, A., Easton, L. E., Sibley, C. R., Sugimoto, Y., Tajnik, M., König, J., and Ule, J. (2014). iCLIP: protein-RNA interactions at nucleotide resolution. *Methods*, 65(3):274–287.

Ionescu-Tirgoviste, C., Gagniuc, P. A., Gubceac, E., Mardare, L., Popescu, I., Dima, S., and Militaru, M. (2015). A 3D map of the islet routes throughout the healthy human pancreas. *Scientific reports*, 5(1):1–14.

Jankowsky, E. and Harris, M. E. (2015). Specificity and nonspecificity in RNA-protein interactions. *Nature reviews molecular cell biology*, 16(9):533–544.

Jensen, M. A., Wilkinson, J. E., and Krainer, A. R. (2014). Splicing factor SRSF6 promotes hyperplasia of sensitized skin. *Nature structural & molecular biology*, 21(2):189–197.

Jeong, S. (2017). SR proteins: binders, regulators, and connectors of RNA. *Molecules and cells*, 40(1):1.

Jin, L., Chen, Y., Crossman, D. K., Datta, A., Vu, T., Mobley, J. A., Basu, M. K., Scarduzio, M., Wang, H., Chang, C., et al. (2020). STRAP regulates alternative splicing fidelity during lineage commitment of mouse embryonic stem cells. *Nature communications*, 11(1):1–18.

Juan-Mateu, J., Alvelos, M. I., Turatsinze, J.-V., Villate, O., Lizarraga-Mollinedo, E., Grieco, F. A., Marroquí, L., Bugliani, M., Marchetti, P., and Eizirik, D. L. (2018). SRp55 regulates a splicing network that controls human pancreatic $\beta$-cell function and survival. *Diabetes*, 67(3):423–436.

Kang, H.-S., Sánchez-Rico, C., Ebersberger, S., Sutandy, F. R., Busch, A., Welte, T., Stehle, R., Hipp, C., Schulz, L., Buchbender, A., et al. (2020). An autoinhibitory intramolecular interaction proof-reads RNA recognition by the essential splicing factor U2AF2. *Proceedings of the national academy of sciences*, 117(13):7140–7149.

Kargapolova, Y., Levin, M., Lackner, K., and Danckwardt, S. (2017). sCLIP an integrated platform to study RNA-protein interactomes in biomedical research: identification of CSTF2tau in alternative processing of small nuclear RNAs. *Nucleic acids research*, 45(10):6074–6086.

Karni, R., de Stanchina, E., Lowe, S. W., Sinha, R., Mu, D., and Krainer, A. R. (2007). The gene encoding the splicing factor SF2/ASF is a proto-oncogene. *Nature structural & molecular biology*, 14(3):185–193.

Kashima, T. and Manley, J. L. (2003). A negative element in SMN2 exon 7 inhibits splicing in spinal muscular atrophy. *Nature genetics*, 34(4):460–463.

Ke, S. and Chasin, L. A. (2011). Context-dependent splicing regulation: exon definition, co-occurring motif pairs and tissue specificity. *RNA biology*, 8(3):384–388.

Kim, E., Magen, A., and Ast, G. (2007). Different levels of alternative splicing among eukaryotes. *Nucleic acids research*, 35(1):125–131.

Kim, J., Park, R., Chen, J., Jeong, S., and Ohn, T. (2014). Splicing factor SRSF3 represses the translation of programmed cell death 4 mRNA by associating with the 5'-UTR region. *Cell death & differentiation*, 21(3):481–490.

König, J., Zarnack, K., Luscombe, N. M., and Ule, J. (2012). Protein-RNA interactions: new genomic technologies and perspectives. *Nature reviews genetics*, 13(2):77–83.

König, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D. J., Luscombe, N. M., and Ule, J. (2010). iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nature structural & molecular biology*, 17(7):909.

Kornblihtt, A. R. (2007). Coupling transcription and alternative splicing. *Advances in experimental medicine and biology*, 623:175–189.

Kozomara, A. and Griffiths-Jones, S. (2010). miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic acids research*, 39(suppl_1):D152–D157.

Krakau, S., Richard, H., and Marsico, A. (2017). PureCLIP: capturing target-specific protein-RNA interaction footprints from single-nucleotide CLIP-seq data. *Genome biology*, 18(1):1–17.

Kucukural, A., Özadam, H., Singh, G., Moore, M. J., and Cenik, C. (2013). ASPeak: an abundance sensitive peak detection algorithm for RIP-Seq. *Bioinformatics*, 29(19):2485–2486.

Kurosaki, T. and Maquat, L. E. (2016). Nonsense-mediated mRNA decay in humans at a glance. *Journal of cell science*, 129(3):461–467.

Kyburz, A., Friedlein, A., Langen, H., and Keller, W. (2006). Direct interactions between subunits of CPSF and the U2 snRNP contribute to the coupling of pre-mRNA 3' end processing and splicing. *Molecular cell*, 23(2):195–205.

Lareau, L. F., Inada, M., Green, R. E., Wengrod, J. C., and Brenner, S. E. (2007). Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. *Nature*, 446(7138):926–929.

Lawrence, M., Gentleman, R., and Carey, V. (2009). rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics*, 25(14):1841–1842.

Lawrence, M., Huber, W., Pages, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M. T., and Carey, V. J. (2013). Software for computing and annotating genomic ranges. *PLoS computational biology*, 9(8):e1003118.

Lee, F. C. and Ule, J. (2018). Advances in CLIP technologies for studies of protein-RNA interactions. *Molecular cell*, 69(3):354–369.

Lee, Y. and Rio, D. C. (2015). Mechanisms and regulation of alternative pre-mRNA splicing. *Annual review of biochemistry*, 84:291–323.

Lemaire, R., Prasad, J., Kashima, T., Gustafson, J., Manley, J. L., and Lafyatis, R. (2002). Stability of a PKCI-1-related mRNA is controlled by the splicing factor ASF/SF2: a novel function for SR proteins. *Genes & development*, 16(5):594–607.

Lerner, M. R. and Steitz, J. A. (1979). Antibodies to small nuclear RNAs complexed with proteins are produced by patients with systemic lupus erythematosus. *Proceedings of the national academy of sciences*, 76(11):5495–5499.

Levin, A. A. (2019). Treating disease at the RNA level with oligonucleotides. *New england journal of medicine*, 380(1):57–70.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078–2079.

Li, Q., Brown, J. B., Huang, H., Bickel, P. J., et al. (2011). Measuring reproducibility of high-throughput experiments. *The annals of applied statistics*, 5(3):1752–1779.

Licatalosi, D. D., Mele, A., Fak, J. J., Ule, J., Kayikci, M., Chi, S. W., Clark, T. A., Schweitzer, A. C., Blume, J. E., Wang, X., et al. (2008). HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, 456(7221):464–469.

Lin, S. and Fu, X.-D. (2007). SR proteins and related factors in alternative splicing. *Advances in experimental medicine and biology*, 623:107–122.

Liu, H.-X., Zhang, M., and Krainer, A. R. (1998). Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins. *Genes & development*, 12(13).

Liu, X. (2019). SLC family transporters. In *Drug Transporters in Drug Disposition, Effects and Toxicity*, pages 101–202. Springer.

Long, J. C. and Caceres, J. F. (2009). The SR protein family of splicing factors: master regulators of gene expression. *Biochemical journal*, 417(1):15–27.

Lou, H., Neugebauer, K. M., Gagel, R. F., and Berget, S. M. (1998). Regulation of alternative polyadenylation by U1 snRNPs and SRp20. *Molecular and cellular biology*, 18(9):4977–4985.

Lovci, M. T., Ghanem, D., Marr, H., Arnold, J., Gee, S., Parra, M., Liang, T. Y., Stark, T. J., Gehman, L. T., Hoon, S., et al. (2013). Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. *Nature structural & molecular biology*, 20(12):1434.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, 15(12):550.

Lunde, B. M., Moore, C., and Varani, G. (2007). RNA-binding proteins: modular design for efficient function. *Nature reviews molecular cell biology*, 8(6):479–490.

Mackereth, C. D., Madl, T., Bonnal, S., Simon, B., Zanier, K., Gasch, A., Rybin, V., Valcárcel, J., and Sattler, M. (2011). Multi-domain conformational selection underlies pre-mRNA splicing regulation by U2AF. *Nature*, 475(7356):408–411.

Manley, J. L. and Krainer, A. R. (2010). A rational nomenclature for serine/arginine-rich protein splicing factors (SR proteins). *Genes & development*, 24(11):1073–1074.

Maticzka, D., Lange, S. J., Costa, F., and Backofen, R. (2014). GraphProt: modeling binding preferences of RNA-binding proteins. *Genome biology*, 15(1):1–18.

McKenzie, M. D., Jamieson, E., Jansen, E. S., Scott, C. L., Huang, D. C., Bouillet, P., Allison, J., Kay, T. W., Strasser, A., and Thomas, H. E. (2010). Glucose induces pancreatic islet cell apoptosis that requires the BH3-only proteins Bim and Puma and multi-BH domain protein Bax. *Diabetes*, 59(3):644–652.

Monzón-Casanova, E., Matheson, L. S., Tabbada, K., Zarnack, K., Smith, C. W., and Turner, M. (2020). Polypyrimidine tract-binding proteins are essential for B cell development. *Elife*, 9:e53557.

Müller-McNicoll, M., Botti, V., de Jesus Domingues, A. M., Brandl, H., Schwich, O. D., Steiner, M. C., Curk, T., Poser, I., Zarnack, K., and Neugebauer, K. M. (2016). SR proteins are NXF1 adaptors that link alternative RNA processing to mRNA export. *Genes & development*, 30(5):553–566.

Müller-McNicoll, M. and Neugebauer, K. M. (2013). How cells get the message: dynamic assembly and function of mRNA-protein complexes. *Nature reviews genetics*, 14(4):275–287.

Narita, T., Yung, T. M., Yamamoto, J., Tsuboi, Y., Tanabe, H., Tanaka, K., Yamaguchi, Y., and Handa, H. (2007). NELF interacts with CBC and participates in 3' end processing of replication-dependent histone mRNAs. *Molecular cell*, 26(3):349–365.

Ni, J. Z., Grate, L., Donohue, J. P., Preston, C., Nobida, N., OBrien, G., Shiue, L., Clark, T. A., Blume, J. E., and Ares, M. (2007). Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. *Genes & development*, 21(6):708–718.

Niranjanakumari, S., Lasda, E., Brazas, R., and Garcia-Blanco, M. A. (2002). Reversible cross-linking combined with immunoprecipitation to study RNA-protein interactions in vivo. *Methods*, 26(2):182–190.

Nogueira, T. C., Paula, F. M., Villate, O., Colli, M. L., Moura, R. F., Cunha, D. A., Marselli, L., Marchetti, P., Cnop, M., Julier, C., et al. (2013). GLIS3, a susceptibility gene for type 1 and type 2 diabetes, modulates pancreatic beta cell apoptosis via regulation of a splice variant of the BH3-only protein Bim. *PLoS genetics*, 9(5):e1003532.

Ochoa, D., Hercules, A., Carmona, M., Suveges, D., Gonzalez-Uriarte, A., Malangone, C., Miranda, A., Fumis, L., Carvalho-Silva, D., Spitzer, M., et al. (2020). Open Targets Platform: supporting systematic drug-target identification and prioritisation. *Nucleic acids research*.

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic acids research*, 44(D1):D733–D745.

Olena, A. F. and Patton, J. G. (2010). Genomic organization of microRNAs. *Journal of cellular physiology*, 222(3):540–545.

Overbergh, L., Valckx, D., Waer, M., and Mathieu, C. (1999). Quantification of murine cytokine mRNAs using real time quantitative reverse transcriptase PCR. *Cytokine*, 11(4):305–312.

Padgett, R. A., Konarska, M. M., Grabowski, P. J., Hardy, S. F., and Sharp, P. A. (1984). Lariat RNA's as intermediates and products in the splicing of messenger RNA precursors. *Science*, 225(4665):898–903.

Pagès, H., Aboyoun, P., Gentleman, R., and DebRoy, S. (2017). Biostrings: Efficient manipulation of biological strings. *R package version*, 2(0).

Pandit, S., Zhou, Y., Shiue, L., Coutinho-Mansfield, G., Li, H., Qiu, J., Huang, J., Yeo, G. W., Ares Jr, M., and Fu, X.-D. (2013). Genome-wide analysis reveals SR protein cooperation and competition in regulated splicing. *Molecular cell*, 50(2):223–235.

Park, J. W., Jung, S., Rouchka, E. C., Tseng, Y.-T., and Xing, Y. (2016). rMAPS: RNA map analysis and plotting server for alternative exon regulation. *Nucleic acids research*, 44(W1):W333–W338.

Park, S., Brugiolo, M., Akerman, M., Das, S., Urbanski, L., Geier, A., Kesarwani, A. K., Fan, M., Leclair, N., Lin, K.-T., et al. (2019). Differential functions of splicing factors in mammary transformation and breast cancer metastasis. *Cell reports*, 29(9):2672–2688.

Pociot, F. (2017). Type 1 diabetes genome-wide association studies: not to be lost in translation. *Clinical & translational immunology*, 6(12):e162.

Proudfoot, N. (2000). Connecting transcription to messenger RNA processing. *Trends in biochemical sciences*, 25(6):290–293.

Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842.

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., and Mesirov, J. P. (2011). Integrative genomics viewer. *Nature biotechnology*, 29(1):24–26.

Röder, P. V., Wu, B., Liu, Y., and Han, W. (2016). Pancreatic regulation of glucose homeostasis. *Experimental & molecular medicine*, 48(3):e219–e219.

Rodriguez, J. M., Maietta, P., Ezkurdia, I., Pietrelli, A., Wesselink, J.-J., Lopez, G., Valencia, A., and Tress, M. L. (2013). APPRIS: annotation of principal and alternative splice isoforms. *Nucleic acids research*, 41(D1):D110–D117.

Rogers, J. and Wall, R. (1980). A mechanism for RNA splicing. *Proceedings of the national academy of sciences*, 77(4):1877–1879.

Rot, G., Wang, Z., Huppertz, I., Modic, M., Lenče, T., Hallegger, M., Haberman, N., Curk, T., von Mering, C., and Ule, J. (2017). High-resolution RNA maps suggest common principles of splicing and polyadenylation regulation by TDP-43. *Cell reports*, 19(5):1056–1067.

Rufanova, V. A., Alexanian, A., Wakatsuki, T., Lerner, A., and Sorokin, A. (2009). Pyk2 mediates endothelin-1 signaling via p130Cas/BCAR3 cascade and regulates human glomerular mesangial cell adhesion and spreading. *Journal of cellular physiology*, 219(1):45–56.

Sanford, J. R., Coutinho, P., Hackett, J. A., Wang, X., Ranahan, W., and Caceres, J. F. (2008). Identification of nuclear and cytoplasmic mRNA targets for the shuttling protein SF2/ASF. *PloS one*, 3(10):e3369.

Sanford, J. R., Wang, X., Mort, M., VanDuyn, N., Cooper, D. N., Mooney, S. D., Edenberg, H. J., and Liu, Y. (2009). Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts. *Genome research*, 19(3):381–394.

Sapra, A. K., Änkö, M.-L., Grishina, I., Lorenz, M., Pabis, M., Poser, I., Rollins, J., Weiland, E.-M., and Neugebauer, K. M. (2009). SR protein family members display diverse activities in the formation of nascent and mature mRNPs in vivo. *Molecular cell*, 34(2):179–190.

Scanlan, M. J., Chen, Y.-T., Williamson, B., Gure, A. O., Stockert, E., Gordan, J. D., Türeci, Ö., Sahin, U., Pfreundschuh, M., and Old, L. J. (1998). Characterization of human colon cancer antigens recognized by autologous antibodies. *International journal of cancer*, 76(5):652–658.

Schaal, T. D. and Maniatis, T. (1999). Selection and characterization of pre-mRNA splicing enhancers: identification of novel SR protein-specific enhancer sequences. *Molecular and cellular biology*, 19(3):1705–1719.

Schwerk, C. and Schulze-Osthoff, K. (2005). Regulation of apoptosis by alternative pre-mRNA splicing. *Molecular cell*, 19(1):1–13.

Screaton, G. R., Caceres, J. F., Mayeda, A., Bell, M. V., Plebanski, M., Jackson, D. G., Bell, J. I., and Krainer, A. R. (1995). Identification and characterization of three members of the human SR family of pre-mRNA splicing factors. *The EMBO journal*, 14(17):4336–4349.

Senée, V., Chelala, C., Duchatelet, S., Feng, D., Blanc, H., Cossec, J.-C., Charon, C., Nicolino, M., Boileau, P., Cavener, D. R., et al. (2006). Mutations in GLIS3 are responsible for a rare syndrome with neonatal diabetes mellitus and congenital hypothyroidism. *Nature genetics*, 38(6):682–687.

Shao, C., Yang, B., Wu, T., Huang, J., Tang, P., Zhou, Y., Zhou, J., Qiu, J., Jiang, L., Li, H., et al. (2014). Mechanisms for U2AF to define 3' splice sites and regulate alternative splicing in the human genome. *Nature structural & molecular biology*, 21(11):997–1005.

Shattil, S. J., O'Toole, T., Eigenthaler, M., Thon, V., Williams, M., Babior, B. M., and Ginsberg, M. H. (1995). Beta 3-endonexin, a novel polypeptide that interacts specifically with the cytoplasmic tail of the integrin beta 3 subunit. *The journal of cell biology*, 131(3):807–816.

Shen, H., Kan, J. L., and Green, M. R. (2004). Arginine-serine-rich domains bound at splicing enhancers contact the branchpoint to promote prespliceosome assembly. *Molecular cell*, 13(3):367–376.

Shen, S., Park, J. W., Lu, Z.-x., Lin, L., Henry, M. D., Wu, Y. N., Zhou, Q., and Xing, Y. (2014). rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proceedings of the national academy of sciences*, 111(51):E5593–E5601.

Shi, Y. (2017). Mechanistic insights into precursor messenger RNA splicing by the spliceosome. *Nature reviews molecular cell biology*, 18(11):655.

Sickmier, E. A., Frato, K. E., Shen, H., Paranawithana, S. R., Green, M. R., and Kielkopf, C. L. (2006). Structural basis for polypyrimidine tract recognition by the essential pre-mRNA splicing factor U2AF65. *Molecular cell*, 23(1):49–59.

Signor, S. A. and Nuzhdin, S. V. (2018). The Evolution of Gene Expression in cis and trans. *Trends in genetics*, 34(7):532–544.

Singh, G., Kucukural, A., Cenik, C., Leszyk, J. D., Shaffer, S. A., Weng, Z., and Moore, M. J. (2012). The cellular EJC interactome reveals higher-order mRNP structure and an EJC-SR protein nexus. *Cell*, 151(4):750–764.

Singh, G., Pratt, G., Yeo, G. W., and Moore, M. J. (2015). The clothes make the mRNA: past and present trends in mRNP fashion. *Annual review of biochemistry*, 84:325–354.

Smith, T., Heger, A., and Sudbery, I. (2017). UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome research*, 27(3):491–499.

Solovyev, V. V. and Shahmuradov, I. A. (2003). PromH: promoters identification using orthologous genomic sequences. *Nucleic acids research*, 31(13):3540–3545.

Steck, A. K., Dong, F., Wong, R., Fouts, A., Liu, E., Romanos, J., Wijmenga, C., Norris, J. M., and Rewers, M. J. (2014). Improving prediction of type 1 diabetes by testing non-HLA genetic variants in addition to HLA markers. *Pediatric diabetes*, 15(5):355–362.

Stewart, M. (2019). Polyadenylation and nuclear export of mRNAs. *Journal of biological chemistry*, 294(9):2977–2987.

Sugimoto, Y., König, J., Hussain, S., Zupan, B., Curk, T., Frye, M., and Ule, J. (2012). Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions. *Genome biology*, 13(8):R67.

Sutandy, F. R., Ebersberger, S., Huang, L., Busch, A., Bach, M., Kang, H.-S., Fallmann, J., Maticzka, D., Backofen, R., Stadler, P. F., et al. (2018). In vitro iCLIP-based modeling uncovers how the splicing factor U2AF2 relies on regulation by cofactors. *Genome research*, 28(5):699–713.

Sutandy, F. R., Hildebrandt, A., and König, J. (2016). Profiling the binding sites of RNA-binding proteins with nucleotide resolution using iCLIP. In *Post-Transcriptional Gene Regulation*, pages 175–195. Springer.

Taha, D., Barbar, M., Kanaan, H., and Williamson Balfe, J. (2003). Neonatal diabetes mellitus, congenital hypothyroidism, hepatic fibrosis, polycystic kidneys, and congenital glaucoma: a new autosomal recessive syndrome? *American journal of medical genetics part A*, 122(3):269–273.

Tavanez, J. P., Madl, T., Kooshapur, H., Sattler, M., and Valcárcel, J. (2012). hnRNP A1 proofreads 3' splice site recognition by U2AF. *Molecular cell*, 45(3):314–329.

Tenenbaum, S. A., Carson, C. C., Lager, P. J., and Keene, J. D. (2000). Identifying mRNA subsets in messenger ribonucleoprotein complexes by using cDNA arrays. *Proceedings of the national academy of sciences*, 97(26):14085–14090.

Timmer, T., Terpstra, P., van den Berg, A., Veldhuis, P. M., Ter Elst, A., Voutsinas, G., Hulsbeek, M. M., Draaijers, T. G., Looman, M. W., Kok, K., et al. (1999). A comparison of genomic structures and expression patterns of two closely related flanking genes in a critical lung cancer region at 3p21. 3. *European journal of human genetics*, 7(4):478–486.

Tranchevent, L.-C., Aubé, F., Dulaurier, L., Benoit-Pilven, C., Rey, A., Poret, A., Chautard, E., Mortada, H., Desmet, F.-O., Chakrama, F. Z., et al. (2017). Identification of protein features encoded by alternative exons using Exon Ontology. *Genome research*, 27(6):1087–1097.

Uhl, M., Houwaart, T., Corrado, G., Wright, P. R., and Backofen, R. (2017). Computational analysis of CLIP-seq data. *Methods*, 118:60–72.

Ule, J., Jensen, K., Mele, A., and Darnell, R. B. (2005). CLIP: a method for identifying protein-RNA interaction sites in living cells. *Methods*, 37(4):376–386.

Ule, J., Jensen, K. B., Ruggiu, M., Mele, A., Ule, A., and Darnell, R. B. (2003). CLIP identifies Nova-regulated RNA networks in the brain. *Science*, 302(5648):1212–1215.

Ule, J., Stefani, G., Mele, A., Ruggiu, M., Wang, X., Taneri, B., Gaasterland, T., Blencowe, B. J., and Darnell, R. B. (2006). An RNA map predicting Nova-dependent splicing regulation. *Nature*, 444(7119):580–586.

Uren, P. J., Bahrami-Samani, E., Burns, S. C., Qiao, M., Karginov, F. V., Hodges, E., Hannon, G. J., Sanford, J. R., Penalva, L. O., and Smith, A. D. (2012). Site identification in high-throughput RNA-protein interaction data. *Bioinformatics*, 28(23):3013–3020.

Uszczynska-Ratajczak, B., Lagarde, J., Frankish, A., Guigó, R., and Johnson, R. (2018). Towards a complete map of the human long non-coding rna transcriptome. *Nature reviews genetics*, 19(9):535–548.

Vagner, S., Vagner, C., and Mattaj, I. W. (2000). The carboxyl terminus of vertebrate poly (A) polymerase interacts with U2AF 65 to couple 3'-end processing and splicing. *Genes & development*, 14(4):403–413.

Van Nostrand, E. L., Freese, P., Pratt, G. A., Wang, X., Wei, X., Xiao, R., Blue, S. M., Chen, J.-Y., Cody, N. A., Dominguez, D., et al. (2020). A large-scale binding and functional map of human RNA-binding proteins. *Nature*, 583(7818):711–719.

Van Nostrand, E. L., Gelboin-Burkhart, C., Wang, R., Pratt, G. A., Blue, S. M., and Yeo, G. W. (2017). CRISPR/Cas9-mediated integration enables TAG-eCLIP of endogenously tagged RNA binding proteins. *Methods*, 118:50–59.

Wahl, M. C., Will, C. L., and Lührmann, R. (2009). The spliceosome: design principles of a dynamic RNP machine. *Cell*, 136(4):701–718.

Wang, Y., Liu, J., Huang, B., Xu, Y.-M., Li, J., Huang, L.-F., Lin, J., Zhang, J., Min, Q.-H., Yang, W.-M., et al. (2015). Mechanism of alternative splicing and its regulation. *Biomedical reports*, 3(2):152–158.

Warf, M. B., Diegel, J. V., von Hippel, P. H., and Berglund, J. A. (2009). The protein factors MBNL1 and U2AF65 bind alternative RNA structures to regulate splicing. *Proceedings of the national academy of sciences*, 106(23):9203–9208.

Weir, G. C. and Bonner-Weir, S. (2013). Islet $\beta$ cell mass in diabetes and how it relates to function, birth, and death. *Annals of the New York academy of sciences*, 1281(1):92.

Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., DiCuccio, M., Edgar, R., Federhen, S., et al. (2007). Database resources of the national center for biotechnology information. *Nucleic acids research*, 36(1):D13–D21.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer.

Winkler, C., Krumsiek, J., Buettner, F., Angermüller, C., Giannopoulou, E. Z., Theis, F. J., Ziegler, A.-G., and Bonifacio, E. (2014). Feature ranking of type 1 diabetes susceptibility genes improves prediction of type 1 diabetes. *Diabetologia*, 57(12):2521–2529.

Wu, J. Y. and Maniatis, T. (1993). Specific interactions between proteins implicated in splice site selection and regulated alternative splicing. *Cell*, 75(6):1061–1070.

Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics: a journal of integrative biology*, 16(5):284–287.

Zahler, A. M., Neugebauer, K. M., Lane, W. S., and Roth, M. B. (1993). Distinct functions of SR proteins in alternative pre-mRNA splicing. *Science*, 260(5105):219–222.

Zarnack, K., Balasubramanian, S., Gantier, M. P., Kunetsky, V., Kracht, M., Schmitz, M. L., and Sträßer, K. (2020). Dynamic mRNP Remodeling in Response to Internal and External Stimuli. *Biomolecules*, 10(9):1310.

Zarnack, K., König, J., Tajnik, M., Martincorena, I., Eustermann, S., Stévant, I., Reyes, A., Anders, S., Luscombe, N. M., and Ule, J. (2013). Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements. *Cell*, 152(3):453–466.

Zarnegar, B. J., Flynn, R. A., Shen, Y., Do, B. T., Chang, H. Y., and Khavari, P. A. (2016). irCLIP platform for efficient characterization of protein-RNA interactions. *Nature methods*, 13(6):489–492.

Zhao, J., Ohsumi, T. K., Kung, J. T., Ogawa, Y., Grau, D. J., Sarma, K., Song, J. J., Kingston, R. E., Borowsky, M., and Lee, J. T. (2010). Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Molecular cell*, 40(6):939–953.

Zheng, Z.-M., He, P.-J., and Baker, C. C. (1997). Structural, functional, and protein binding analyses of bovine papillomavirus type 1 exonic splicing enhancers. *Journal of virology*, 71(12):9096–9107.

Zhou, Q. and Melton, D. A. (2018). Pancreas regeneration. *Nature*, 557(7705):351–358.

Zhou, Z. and Fu, X.-D. (2013). Regulation of splicing by SR proteins and SR protein-specific kinases. *Chromosoma*, 122(3):191–207.

Zhu, J., Mayeda, A., and Krainer, A. R. (2001). Exon identity established through differential antagonism between exonic splicing silencer-bound hnRNP A1 and enhancer-bound SR proteins. *Molecular cell*, 8(6):1351–1361.

# Acknowledgements

First and foremost, I would like to thank my PhD supervisor Dr. Kathi Zarnack. Thanks for the great opportunity to join your group. Thank you for the amazing support and advice along the way. Your passion for science and motivation to explain the same thing over and over again was truly inspiring. I also wish to thank my co-supervisor Prof. Dr. Ingo Ebersberger, who constantly encouraged me in my learning process and shaped my understanding of science and bioinformatics like few others.

My sincere thanks also goes to all collaborators, especially to Dr. Julian Konig for advice and help in almost all projects. Thanks to Dr. Stefanie Ebersberger and Dr. Anke Busch for the collaboration on the iCLIP processing pipeline project. A special thank goes to the group of Prof. Dr. Décio L. Eizirik, in particular Ines Alvelos for the amazing collaborative work on the SRSF6 project.

Another warm thank goes to all members of the Zarnack group for the great atmosphere and nice learning environment. Thanks to the former group members Antonella and Samarth to get me started in the group. Thanks also to all current members of the group for the fruitful discussions, especially Mario and my PhD colleagues You and Melina. Its been a nice and fun time.

Lastly, I would like to thank my parents, family and friends for their support and encouragement.

# Erklärung

Ich erkläre hiermit, dass ich mich bisher keiner Doktorprüfung im Mathematisch-Naturwissenschaftlichen Bereich unterzogen habe.

Frankfurt am Main, den ____________ ________________________________

Mirko Brüggemann

# Versicherung

Ich versichere hiermit, dass die vorgelegte Doktorarbeit über "**Computational studies on RNA processing in higher eukaryotes**" selbständig angefertigt und mich anderer Hilfsmittel als der in ihr angegebenen nicht bedient habe, insbesondere, dass alle Entlehnungen aus anderen Schriften mit Angabe der betreffenden Schrift gekennzeichnet sind.

Ich versichere, die Grundsätze der guten wissenschaftlichen Praxis beachtet, und nicht die Hilfe einer kommerziellen Promotionsvermittlung in Anspruch genommen zu haben.

Frankfurt am Main, den ____________ ____________________________________

Mirko Brüggemann