Computational Analyses of High-Throughput Sequencing Data to Understand RNA Metabolism Defects Associated with ALS

Hembach, Katharina M

Computational Analyses of High-Throughput Sequencing Data to Understand RNA Metabolism Defects Associated with ALS

Dissertation

zur

Erlangung der naturwissenschaftlichen Doktorwürde

(Dr.sc.nat.)

vorgelegt der

Mathematisch-naturwissenschaflichen Fakultät

der

Universität Zürich

von

Katharina Maria Hembach

aus

Deutschland

Promotionskommission

Prof. Dr. Mark D. Robinson (Vorsitz und Leitung der Dissertation)
Prof. Dr. Magdalini Polymenidou (Leitung der Dissertation)
Prof. Dr. Mihaela Zavolan
Prof. Dr. Frédéric Allain

Zürich, 2020

To Alex

Contents

Part III Concluding Remarks

Summary

Amyotrophic lateral sclerosis (ALS) is a devastating neurodegenerative disease. It affects the upper and lower motor neurons and leads to progressive paralysis. So far, there is no cure and available treatments are very limited. The exact disease pathology is still not understood, but cytoplasmic protein inclusions have been identified as a hallmark of ALS. These inclusions mainly consists of either of two RNA binding proteins, TDP-43 and FUS. Besides, mutations in these two proteins have been associated with the inherited form of ALS. One important role of TDP-43 is the repression of cryptic exons, which are erroneously spliced-in exons that often cause frame shifts or premature termination codons. Mature mRNAs with cryptic exons are usually degraded fast leading to a disregulation of mRNA levels in affected neurons.

In my PhD, I wanted to gain insights into RNA metabolism defects associated with ALS. This included the development of bioinformatic tools for the analysis of RNA-seq data related to ALS. My thesis is comprised of four chapters.

The first chapter is a review about the RNA-seq technology. It describes statistical methods for differential analyses and recent developments in the field.

The second chapter presents ARMOR, a workflow for RNA-seq data analysis. ARMOR includes code for read preprocessing, mapping, transcript expression quantification and differential gene expression analysis. The workflow enables automated analyses of RNA-seq experiments and can easily be extended or adjusted.

The third chapter is a collaboration in which we tried to identify the role of FUS at the synaptic site. We identified the synaptic RNA targets of FUS with CLIP-seq. In an ALS-mouse model with increased levels of cytoplasmic FUS, we found age-dependent synapse specific RNA changes at presymptomatic stages. The altered RNAs included important components of GABAergic and glutamatergic synapses indicating a potential mechanism of early synaptic impairment in neurodegeneration.

The fourth chapter presents DISCERNS, an R package for the discovery of unanno-

tated splicing events in RNA-seq data using information from splice junction reads. I first simulated an RNA-seq data set with known alternative splicing events to find the best genome alignment tool for my pipeline. I show that STAR has better precision than hisat2. Next, I show that DISCERNS has less false positive predictions than StringTie, a transcriptome assembly method, on the simulated data. Finally, I applied DISCERNS to three published ALS-related RNA-seq data sets and found that it correctly predicted known cryptic exons. Additionally, DISCERNS predicted novel cryptic exons indicating that not all cryptic exon, which are regulated by TDP-43, are known yet.

Zusammenfassung

Amyotrophe Lateralsklerose (ALS) ist eine nicht heilbare neurodegenerative Krankheit. Sie schädigt die oberen und unteren Motorneuronen und führt zunächst zu Muskelschwäche und letztendlich zu vollständiger Lähmung. Bis jetzt gibt es keine Heilung und die aktuellen Behandlungsmöglichkeiten sind sehr begrenzt. Die genaue Krankheitspathologie ist noch nicht bekannt, aber cytoplasmatische Proteinaggregate wurden als ein typisches Merkmal der Erkrankung identifiziert. Diese Proteinaggregate bestehen hauptsächlich aus einem von zwei verschiedenen RNA-bindenden Proteinen, TDP-43 und FUS. Außerdem sind Mutationen in diesen beiden Proteinen mit der vererbbaren Form von ALS assoziiert. Eine wichtige Funktion von TDP-43 ist die Unterdrückung von kryptischen Exons. Kryptische Exons werden fälschlicherweise in die reife mRNA gespleißt und verursachen Leserasterverschiebungen oder vorzeitige Stoppcodons, weshalb die mRNA normalerwiese schnell abgebaut wird. Dies führt zu einer Fehlregulation der mRNA Mengen in den betroffenen Neuronen.

In meiner Doktorarbeit möchte ich die Fehler im RNA Metabolismus erforschen, die ALS verursachen. Ein Teil der Arbeit ist auch die Entwicklung von bioinformatischen Methoden zur Analyse von RNA-seq Daten mit Bezug auf ALS. Meine Dissertation besteht aus vier Kapiteln.

Das erste Kapitel ist eine Übersicht über die RNA-seq Technologie. Es beschreibt die statistischen Methoden der differenziellen Analyse und die jüngsten Entwicklungen auf diesem Gebiet.

Im zweiten Kapitel stelle ich ARMOR vor, einen Workflow für die Analyse von RNA-seq Daten. ARMOR beinhaltet Code um die sequenzierten Reads aufzubereiten, dem Referenzgenom zuzuordnen, die Transkriptexpression zu quantifizieren und die differenzielle Genexpression zu analysieren. Der Workflow ermöglicht die automatische Analyse von RNA-seq Experimenten und kann sehr einfach erweitert oder angepasst werden.

Das dritte Kapitel ist eine Kollaboration in der wir dir Rolle von FUS in Synapsen

untersuchen wollten. Wir haben die synaptischen RNA Bindungspartner von FUS mithilfe von CLIP-seq identifiziert. In einem ALS-Mausmodell mit einer erhöhten Menge von cytoplasmatischem FUS, haben wir altersabhängige synapsenspezifische RNA Veränderungen gefunden bevor die Mäuse Symptome gezeigt haben. Die veränderten RNAs beinhalteten wichtige Proteine von GABAergen und glutamatergen Synapsen, was darauf hindeutet, dass Synapsen schon früh während der Neurodegeneration beeinträchtigt sind.

Im vierten Kapitel stelle ich DISCERNS vor, ein R-Paket für das Finden von nicht annotierten Spleißereignissen in RNA-seq Daten mithilfe von Reads die Spleißverbindungen überspannen. Zunächst habe ich einen RNA-seq Datensatz mit bekannten Spleißstellen simuliert, um das beste Genomalignmentprogramm zu finden. Ich zeige, dass STAR eine höhere Präzision hat als hisat2. Als nächstes zeige ich, dass DISCERNS weniger falschpositive Vorhersagen auf den simulierten Daten macht als StringTie, ein Programm zum Zusammenfügen von Transkriptomen. Zuletzt habe ich DISCERNS auf drei publizierten ALS RNA-seq Datensätze angewandt und konnte zeigen, dass es bekannte kryptische Exons korrekt vorhergesagt hat. Außerdem hat DISCERNS neue kryptische Exons vorhergesagt, was darauf hindeutet, dass nicht alle kryptischen Exons, die durch TDP-43 reguliert werden, bereits bekannt sind.

Abbreviations

ALS	Amyotrophic lateral sclerosis
AS	Alternative splicing
cDNA	Complementary DNA
CLIP-seq	Cross-linking immunoprecipitation sequencing
DNA	Deoxyribonucleic acid
FDR	False discovery rate
FUS	Fused in sarcoma
mRNA	Messenger RNA
NMD	Nonsense mediated decay
PCR	Polymerase chain reaction
RNA	Ribonucleic acid
RNA-seq	RNA sequencing
TDP-43	TAR-DNA binding protein 43
UTR	Untranslated region

Part I.

Introduction and Background

1 Introduction

> *In the last few weeks I have frequently been asked "What causes genes to be split?". This is a deep evolutionary question and I could probably drone on for several hours, but a faster answer will be needed here. In many letters that Phil and I have received from schoolchildren this was the burning question. One surprising answer was proposed by a Reverend from New York. He sent a long and detailed letter explaining that it was "static electricity and deadly ozone gas" that caused the genes to be split. However, it was apparent that this was also his explanation for Crib Death and a variety of other ills that befall us. No doubt prompted by the phrase "splitting the atom" some letters from children, or perhaps their parents, suggested that Phil and I had ourselves split the genes and were curious to know how we had done it. During a conversation with a reporter today I was complimented as being "the youngest of this year's prize-winners". In fact, I am not but do have a youthful appearance. I wonder if my genes are not split as much as other people's.*

– Richard J. Roberts, Nobel laureate 1993 for the discovery of split genes, *speech at the Nobel Banquet, December 10, 1993*

1.1 Biological Background

Cells are the building blocks of life. All organisms are made up of cells and the number of cells varies from single cell organisms, such as yeast, to multi-cellular organisms, for example humans with an estimated 37 trillion cells [1]. On the molecular level, deoxyribonucleic acid (DNA) is the building block that stores the genome information of each cell [2]. A DNA molecule consists of sequences of four different bases: adenin

(A), thymnin (T), cytosin (C) and guanine (G). A is complementary to T and C is complementary to G. In cells, genomic DNA occurs as a double stranded right-handed helix consisting of two complementary DNA molecules. A DNA molecule has a direction: it starts with the 5' end and terminates in the 3' end. The numbers (5 and 3) refer to the number of the exposed carbon atom in the ribose ring of the first and last nucleotide in the DNA molecule. The double stranded DNA helix is stabilized by hydrogen bonds between each pair of bases in the two molecules. In DNA, the genomic information is encoded by the exact sequence of the four bases.

A genome consists of protein coding genes that are separated by intergenic regions that fulfill regulatory roles. Genes are comprised of exons and introns, where exons encode the protein amino acid sequence and introns are non-coding regions that separate exons within a gene. Protein production is a complicated process that involves various steps and intermediate products. The first step is transcription in which the DNA sequence of a gene is copied into a precursor messenger ribonucleic acid (pre-mRNA). RNA is made up of four nucleotides and carries the same information as DNA, however thymin is replaced by uracil (C) that also pairs with adenine. pre-mRNA is unstable and it needs to be stabilized to prevent fast degradation. The 5'end of the pre-mRNA is capped by the addition of a modified guanine nucleotide to prevent degradation. A polyadenylation signal at the 3'end of the pre-mRNA is recognized during transcription and a poly-A sequence, an RNA molecule consisting of 150-250 As in eukaryotes [3], is added to the pre-mRNA 3 end. The capped and polyadenylated pre-mRNA undergoes splicing.

Splicing is the process that generates the variety of proteins within a cell from a limited number of genes. During splicing, all intronic sequences are removed from the pre-mRNA and the exonic sequences are joined together. The spliced, capped and polyadenylated mRNA is termed mature mRNA. The spliceosome, an RNA-protein complex recognizes two splice sites, specific sequences in the mRNA (mostly AGGU), and excises the RNA in between, joining the two ends together. Splicing can also remove exons thereby creating alternative mature mRNAs from the same pre-mRNA molecule; called alternative splicing (Figure 1.1). Alternative splicing allows for the modulation of protein structure and function by generating different mRNAs from the same gene. Splicing is tightly regulated and used as an essential mechanism in cell differentiation, development and as reaction and compensatory mechanism to stress. To regulate splicing, specific sequences in the mRNA (cis-acting elements) are bound by splicing enhancers or silencers (trans-acting factors) to promote or inhibit splicing,

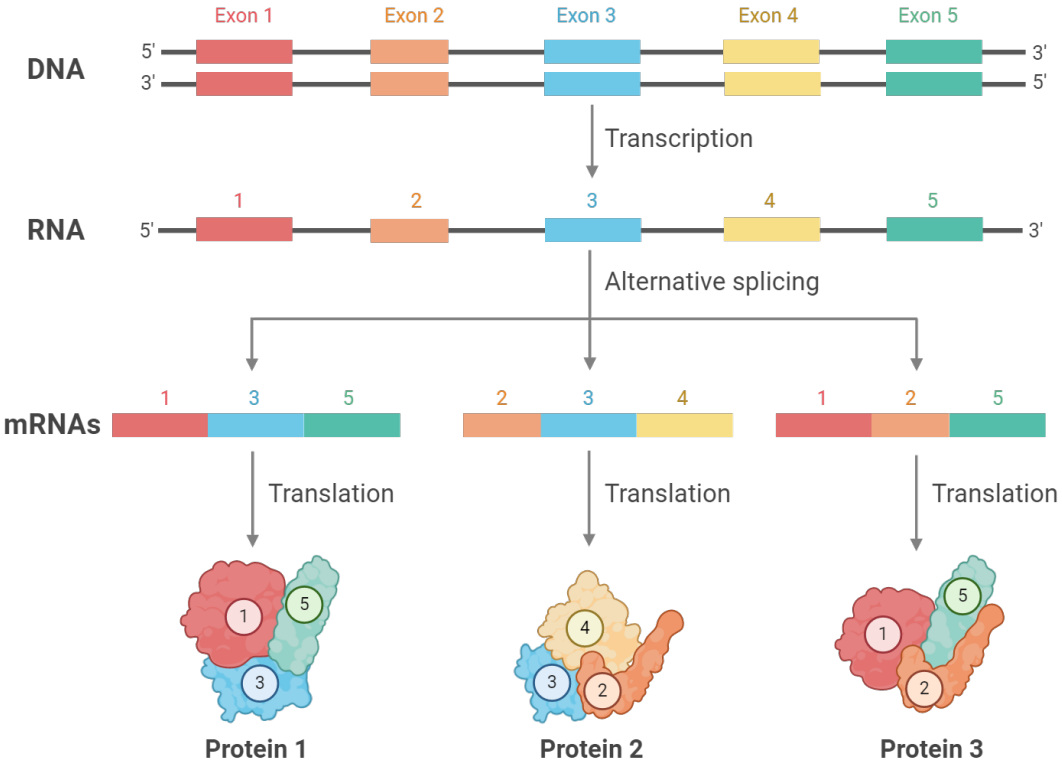


Figure 1.1: **From gene to protein.** Schematic of the processing of DNA to create proteins. Created with BioRender.com

respectively [4].

Mature mRNA is the template for the last step in protein production, which is translation. mRNA is exported to the cytoplasm where it is bound by ribosomes. Starting at the 5'end, a ribosome "translates" the mRNA nucleotide sequence into the protein amino acid sequence: a nucleotide triplet encodes for a specific amino acid, the building block of proteins. The polypeptide (amino acid chain) folds into a native three-dimensional structure simultaneous with translation to create the final functional protein. A single mRNA molecule can be translated more than once to generate multiple copies of the encoded protein. mRNA has a limited lifetime and is eventually degraded. mRNA half-life greatly varies in mammalian cells, ranging from minutes to more than a day, with a median mRNA half-life of 7-10 hours [5, 6].

mRNA only constitutes around 3-7% of total RNA by mass in a mammalian cell [7]. The most abundant RNA in every cell is ribosomal RNA (rRNA; 80-90%) followed by transfer RNA (tRNA; 10-15%). Non coding RNAs (ncRNA), such as micro RNAs (miRNA) or small nuclear RNAs (snRNA), are found in much lower amounts and they

are often cell type specific.

Many different diseases have been associated with the production of malformed proteins or defects in the regulation of protein production. Prominent examples are specific types of cancer or neurodegenerative diseases such as Alzheimer's or Amyotrophic lateral sclerosis (ALS).

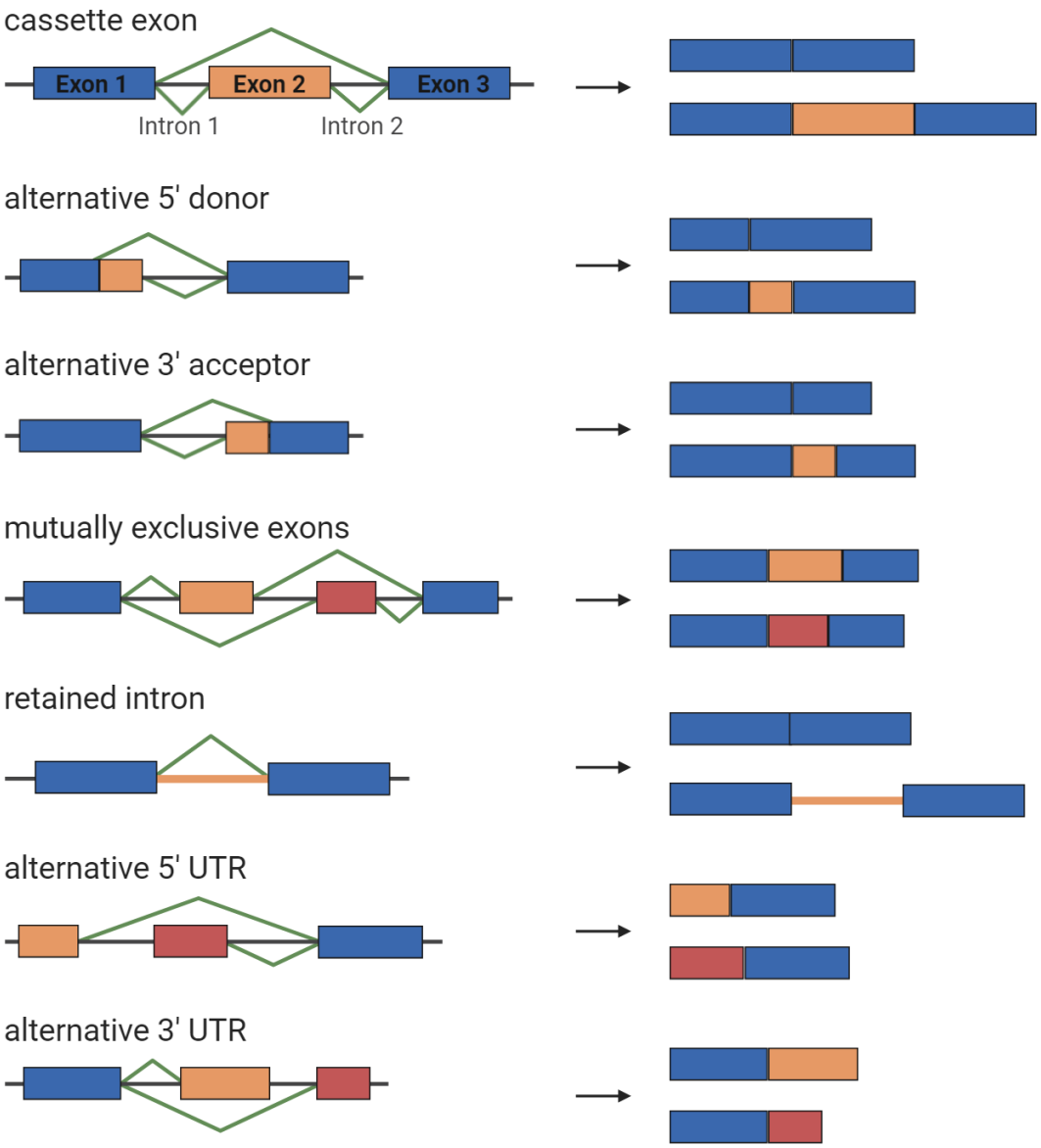


Figure 1.2: **Types of alternative splicing.** Schematic of the different types of alternative splicing. Exons are represented as colored boxes and introns as black lines. Created with BioRender.com

1.1.1 Alternative splicing

Alternative splicing events can be classified into different types. Figure 1.2 provides an overview of the most common types of alternative splicing. The most common form of alternative splicing are cassette exons, which are exons that are spliced out of the mature transcript. The usage of alternative 5' donor and 3' acceptor sites allows the modulation of exon lengths and causes minor changes to protein domains. Mutually exclusive exons are never observed together in a mature mRNA. Retained introns are introns that are not spliced out and included in the mature mRNA. Retained introns are often observed in cancerous tissue [8] and they are associated with cell cycle and cell differentiation [9]. Intron retention often leads to premature termination codons and the mature mRNAs are thus degraded by nonsense mediated decay (NMD), a cellular surveillance mechanisms that removes erroneous transcripts [10]. Alternative 5' and 3' untranslated regions (UTRs) are utilized to control mRNAs and translation. Alternative transcription initiation and/or alternative splicing are the major source of 5' UTR diversity [11]. The 5' UTR has important regulatory functions for transcription and translation initation [12]. Alternative splicing can also include or exclude non-coding exons and thereby regulate 5' UTR length. Alternative 5' UTRs can result in the usage of alternative initiation codons and thus cause alternative first exons in the mature mRNA. The 3' UTR contains regulatory regions that affect polyadenylation and mRNA modifications and thereby regulates mRNA stability, localization and translation. The splicing of alternative last exons results in alternative 3' UTRs and thus alternative polyadenylation, which is an important regulatory mechanism in eukaryotes and observed in about half of the human genes [12]. All examples in Figure 1.2 result in exactly two alternative mRNAs that are created by a specific splicing event. In reality, different splicing events can be combined and exon annotations can overlap causing more complicated events and possibly more than two mature transcripts per gene.

Microexons

Microexons are very short cassette exons, between three and 27 nucleotides long, that are mostly specific for neuronal tissue [13]. Microexons are highly conserved among vertebrates [14]. In mammals, the protein Serine/Arginine Repetitive Matrix 4 (SRRM4) regulates and promotes the inclusion of microexons in neurons via an enhancer of microexons (eMIC) domain in the C-terminus of the protein [14]. The inclusion of microex-

ons in mature mRNAs leads to the addition of one to nine amino acids in the translated protein. Thereby, microexons play important roles in neurogenesis and development as the additional amino acids modulate protein-protein interaction domains [15].

Cryptic exons

Cryptic exons are cassette exons that are not annotated in databases and that are never observed in samples from healthy individuals, but only in disease or after knockdown of splicing repressors [16]. The splicing mechanism of cryptic exons is identical to cassette exons and they use canonical splice sites, however, cryptic exons are completely repressed under normal conditions and in healthy cells. Cryptic exons can lead to premature termination codons and the resulting mRNAs are degraded by NMD which in turn can lead to decreased levels of the encoded proteins. The immediate degradation of transcripts with cryptic exons makes them difficult to detect. Examples of splicing repressors that prevent the inclusion of deleterious cryptic exons in mature mRNAs are polypyrimidine tract-binding protein 1 (PTBP1) and polypyrimidine tract-binding protein 2 (PTBP2) [17], as well as Transactive response DNA binding protein 43 kDa (TDP-43). Specific non-conserved cryptic exons have been found in brain tissue from amyotrophic lateral sclerosis (ALS) patients with TDP-43 loss of function [18].

1.2 Transcriptomics

The collection of all transcribed RNA molecules within a cell is called the transcriptome. Transcriptomics refers to research methods that study the transcriptome of a sample. The first transcriptomic methods were developed 30 years ago. However, these methods were quickly overtaken by DNA microarrays and later RNA sequencing (RNA-seq). Nowadays, transcriptomics ranges from bulk methods that analyse large collections of cells to single-cell resolution methods that are able to sequence the complete transcriptome of a single cell or even a nucleus.

1.2.1 RNA sequencing

RNA sequencing (RNA-seq), also termed next generation sequencing or high-throughput sequencing, is the technology that revolutionized genomic studies and enable the large

scale high-throughput comparison of many different samples and experimental groups. RNA-seq was developed in 2008 and enables the sequencing of millions of RNAs from a sample of interest [19–21]. RNA-seq can be used for discovery of novel transcripts and genes, because it is not limited by the type and origin of detectable RNA sequences, as are older transcriptomic methods such as microarrays. Microarrays consist of short DNA molecules that are fixed on the array surface. The DNA sequences are arranged in probes, which comprise millions of copies with identical sequence. Depending on microarray type, a single gene is represented by one or multiple unique probe sequences. Microarrays are only able to measure the expression of genes/transcripts that are represented by a probe. Therefore, microarrays are limited by their probe design in two ways. Firstly, microarrays can only measure genes with known sequence that are represented by at least one probe. Secondly, probe sequences are sometimes not long enough to distinguish between different mRNAs that originate from the same gene. If none of the probes covers the exon-exon junction that distinguishes two splice variants, differential transcript usage or novel splice variants cannot be measured. RNA-seq reads, however, can cover unknown genes and novel transcripts. Nowadays, RNA-seq is fast, easy and comparably cheap and the standard tool for any transcriptomic analysis.

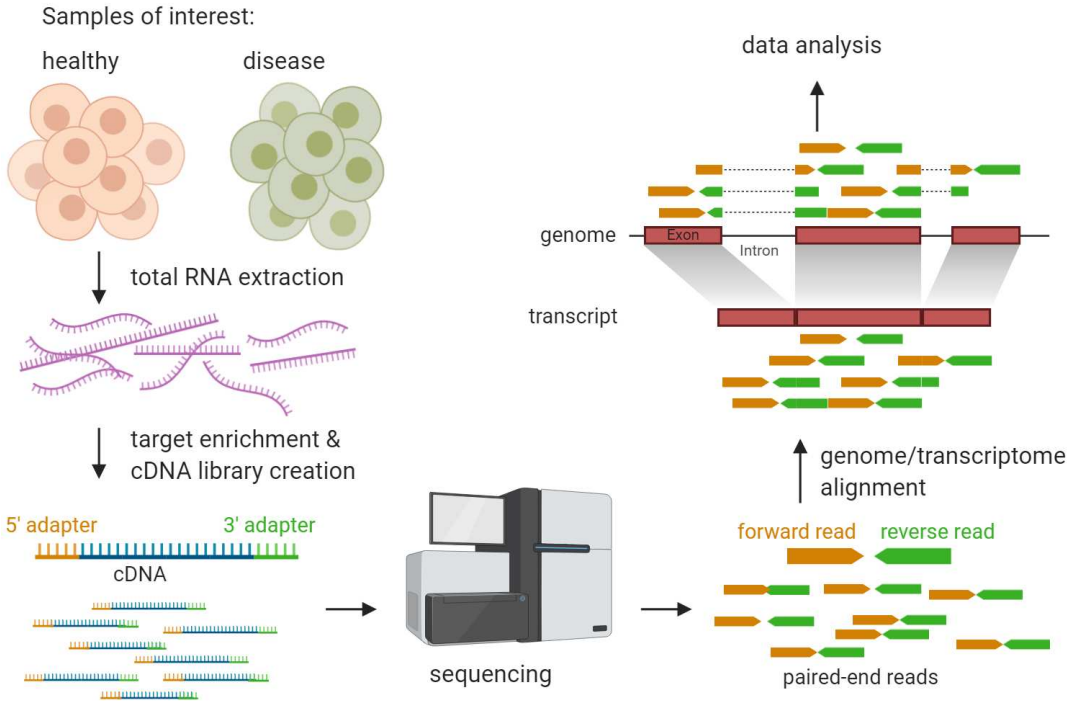


Figure 1.3: **RNA sequencing protocol.** Schematic of a typical RNA sequencing protocol and read processing. Created with BioRender.com

RNA-seq protocol

A standard RNA-seq protocol is outlined in figure 1.3. Total RNA is extracted from all samples of interest. Depending on the experiment, samples are for example from healthy controls and disease patients, from wildtype and gene knockout models, or from controls and chemically treated cells. After RNA extraction, the target RNAs, most likely mRNA, is enriched via rRNA depletion or poly-A selection. Target RNAs are fragmented and reverse transcribed to complementary DNA (cDNA) because cDNA is more stable than RNA and can be easily amplified via polymerase chain reaction (PCR). Adapter sequences are added to the 5' and 3' ends of all cDNA molecules and the cDNA library is amplified. The library is sequenced via high-throughput sequencing, resulting in millions of reads with the nucleotide sequences of the cDNA templates and associated quality scores. Depending on the type of sequencing, reads are single-end (only one end of the cDNA is sequenced) or paired-end (both ends of the cDNA are sequenced). The read data is bioinformatically analysed: depending on the research question, all reads are aligned to the reference genome or transcriptome. Possible applications of RNA-seq include differential gene expression, transcriptome assembly or variant calling.

RNA-seq analysis pipeline

Different questions can be answered with RNA-seq, but most commonly, RNA-seq experiments are targeted towards a differential analysis. This includes differential gene expression (DGE) analyses, where the goal is to compare the gene expression between two or more experimental groups. Differential transcript expression (DTE) analyses if a specific transcript changes in expression and differential transcript usage (DTU) asks if the proportions of all expressed transcripts per gene change between conditions, independent of the overall gene expression. RNA-seq can also be used to also study alternative splicing (AS). Independent of the aim of the experiment, most RNA-seq analysis pipelines follow a common order, which I will outline shortly.

Preprocessing RNA-seq reads are first preprocessed to ensure reliable and comparable read quality across all samples of an experiment. The base calling quality of a read often deteriorates at the read ends [22] and thus positions after a nucleotide with low quality are removed. Possible adapter sequences resulting from short cDNA fragments are also detected and removed during preprocessing. Preprocessing also general quality controls

such as overall read GC content and detection of overrepresented sequences. The most common tool for quality control is FastQC [23] and examples of tools for preprocessing are Cutadapt [24] or Flexbar [25].

Read mapping After preprocessing, the reads are mapped to the reference genome or transcriptome, depending on the research question. Splice-aware alignment tools, such as STAR [26] or hisat2 [27], are able to map reads to their original position in the reference genome and even across splice-junctions when a read covers the boundary between two different exons in the mature mRNA.

Feature counting/estimation The third step is the quantification of features of interest. Depending on the experiment, this can be genes, transcripts or exons. There exist different quantification tools that count the number of reads that map to each feature of interest. Gene or exon counts can be computed from genomic read alignments, by simply counting the number of reads that fall within or overlap with the genomic coordinates of genes or exons. The most commonly used tools are featureCounts [28] and HTSeq [29]. So called alignment-free quantification tools us an alternative approach. These tools combine the transcriptome read alignment with the transcript quantification step. The most prominent transcriptome abundance quantificators are Salmon [30] and kallisto [31]. Instead of read alignment, Salmon and kallisto compare the k-mers of each read to the k-mer content of each transcript and each read is assigned to all compatible transcripts. For transcript quantification, the expectation maximisation (EM) algorithm is used to estimate the expected number of reads of each transcript. The resulting transcript quantifications can be used for DTU or DTE analyses or the transcript estimates can be summed up at the gene level to generate gene counts for DGE. Alignment-free quantification is much faster than genome alignment and it was shown that these approaches generate comparable or even more accurate gene counts than genome alignment followed by traditional gene counting [32, 33]. Salmon also implements an alternative strategy that combines the two approaches: reads are first aligned to the transcriptome, followed by transcript abundance estimation.

Data analysis The last step is the count modeling and final data analysis. The RNA-seq read counts from replicate samples are commonly modeled using a negative binomial distribution. This is based on the assumption that the cDNA fragments that are being sequenced are randomly drawn from the pool of all expressed RNAs. Due to random

sampling, highly expressed genes are more likely to be represented in the pool of sequence reads, whereas lowly expressed genes are more likely to be missed. Log-transformed read counts of lowly expressed genes have higher variance between replicate samples whereas the counts of highly expressed genes have lower variance. This relationship between gene expression and count variance between replicates is termed mean-variance relationship and is inherent to RNA-seq count data. A second result of the random sampling during sequencing is that lowly expressed genes become indistinguishable from noise. Therefore, extremely lowly expressed genes in an experiment are often filtered.

The two most common tools for DGE analysis are edgeR [34] and DESeq2 [35]. Both methods model the gene counts of all samples with a generalized linear model (GLM) using the negative binomial distribution. The usage of GLMs allows for the modeling of more complex experimental designs than simple two-group comparisons, including covariates [36]. The tools compare the null hypothesis, that the relative abundance of a gene is identical in two groups, against the alternative hypothesis, that the relative abundance is different in the two groups. More generally, the tested parameter typically corresponds to the gene expression log fold-change of the two groups and the null hypothesis is that the parameter is zero. The results of a DGE analysis is the list with differential expressed genes, the associated p-values and log2 fold-changes between treatment and control.

Alternative splicing analysis

RNA-seq data can not only be used for differential analyses, but also for alternative splicing (AS) detection. Splice-junction reads contain information about the order of exons in the original mRNA molecule. By comparing the observed splice-junction with the annotated set of splice-junctions per gene, we can discovery novel exon combinations that have not been annotated.

There are three different types of computational AS analysis methods that can be distinguished. The first type are AS detection tools that identify known alternative splicing events in a given RNA-seq sample. The second type are tools that quantify a given list of AS events in a sample. Typically, a percentage spliced-in or percent splicing index (PSI) value is computed per AS event and sample [37, 38]. The PSI value is the fraction of reads that support one specific version of the AS event. The third type of AS analysis tools predict novel AS events that have not been annotated. Most commonly, AS prediction methods use splice graphs to represent genome structure [39,

40]. Splice graphs are directed acyclic graphs that describe all possible splicing variants of a gene. The nodes represent splice-sites, ordered from 5' to 3', and edges represent splice junctions (introns) that connect the splice sites. A transcript corresponds to a path through the graph.

Transcriptome assembly methods reconstruct the transcriptome of an RNA-seq sample. Methods for both *de novo* assembly and genome-guided assembly, using reference gene annotations as a guide, exist. These tools are, in theory, able to also identify unannotated AS events as they intend to assemble the whole transcriptome of a sample.

The first set of ALS-related cryptic exons from Ling *et al.* [18] where identified by manual screening of novel exons annotated by Cufflinks [41]. The authors searched for novel exons that were highly abundant in the TDP-43 knockout samples but not the control samples. More recent publications screened for cryptic exons by first filtering novel splice junctions (obtained from genome alignment) and then comparing the number of supporting reads in control and knockout conditions. CryptSplice [42] quantifies the 5' splice-site coverage of novel junctions and uses a beta binomial test to find junctions specific for an experimental condition. The CryptEx pipeline [43] defines the coordinates of potential cryptic exons based on splice-junctions that splice into an annotated intron and tests for differential usage with DEXseq.

Microexons are more complicated to identify in RNA-seq data than other AS events, because of their short length. There is uncertainty with respect to short splice junction overhangs, because they could be caused by wrong read alignments. The default parameters of STAR [26], for example, will remove all splice junctions with an overhang of less than 12 bases on both sides. In 2014 Irimia *et al.* [13] published the Vertebrate Alternative Splicing and Transcription Tools (Vast-tools), which is a tool set for the identification and comparison of alternative splicing in RNA-seq data. Vast-tools includes a database of potential exon-microexons-exon splice junctions (Vast-DB), based on novel splice-junctions identified from published RNA-seq data sets. To identify microexons in an RNA-seq sample, Vast-tools first splits reads into short (50 nucleotides) fragments and aligns them to the reference genome without allowing splice-junctions. All unmapped read fragments are then aligned to the genomic sequences of the hypothetical exon-microexon-exon combinations in Vast-DB. Vast-tools is therefore able to identify a novel microexon, if the microexon and its connected exons are listed in Vast-DB. Augmented Transcriptome Mapping (ATMap) [15] is a different microexon prediction method. It first aligns the RNA-seq reads to the transcriptome and filters

them. All reads with an insertion of length 3 to 51 nucleotides, which overlaps an exon-exon boundary, are retained. The corresponding intron sequences are then scanned for the inserted sequences to identify unannotated microexons.

1.2.2 CLIP-seq

Cross-linking immunoprecipitation (CLIP) [44] sequencing is a protocol to determine the RNA sequences bound by a protein of interest. In particular, the binding motifs of RBPs can be determined from the set of CLIP sequences. The CLIP protocol was first published in 2003 and it was later combined with next generation sequencing in 2009 [45] to allow high-throughput analyses of genome wide RBP binding sites. CLIP-seq revolutionized the study of RBPs: Since the development of the original CLIP-seq protocol, many modifications have been published [46, 47]. Depending on the exact research objective, different protocols are beneficial and recommended.

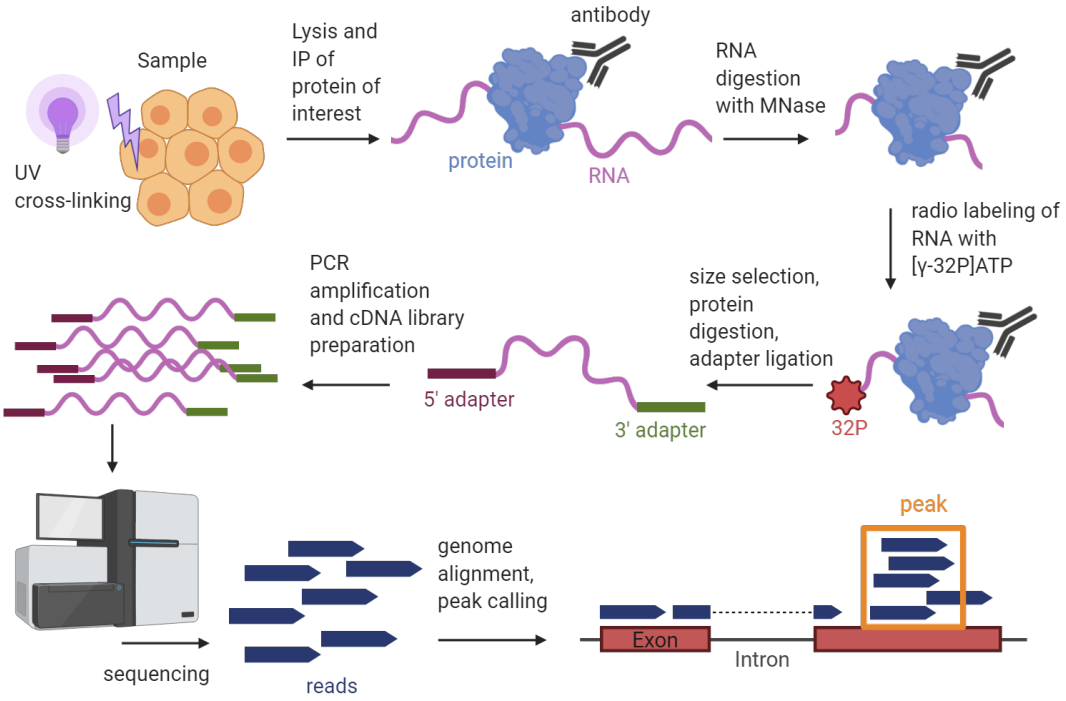


Figure 1.4: **CLIP sequencing protocol.** Schematic of a typical CLIP sequencing protocol and data processing. Created with BioRender.com

CLIP-seq protocol

An example of a CLIP-seq protocol is outlined in figure 1.4. RBP are covalently crosslinked to bound RNAs with UV light to stabilize the endogenous RBP and RNA interactions. The cells are lysed and the protein of interest is immunoprecipitated (IP) with a specific antibody. RNAs that are bound by the purified proteins are partially digested with an RNase to trim the ends that are not protected by the RBP and to prevent the co-purification of other RBPs that are bound to the same RNA molecule. The 5' end of RNAs are radio labelled with ^{32}P-γATP and protein-RNA complexes are size selected using sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS-page). The protein-RNA complexes are excised and the proteins are digested with Proteinase K to purify the bound RNAs. Adapters are ligated to the ends of RNA molecules and the molecules are reverse transcribed and amplified via real-time PCR. The cDNA library is subjected to high-throughput sequencing resulting in millions of reads. After quality control, reads are aligned to the reference genome followed by peak calling (identification of high quality binding sites) and, optionally, motif prediction.

The difference between the most common CLIP-seq protocols is mainly at which step, how and what type of adapters are ligated to the RNAs. A second difference is the final resolution of the identified protein-RNA binding sites: some methods have single-nucleotide resolution, i.e. they are able to identify the nucleotide that was initially crosslinked to the RBP, whereas other methods only identify transcript regions bound by the protein of interest. For example, photoactivable ribonucleoside-enhanced CLIP (PAR-CLIP) has better resolution than CLIP-seq. PAR-CLIP is characterized by T-C transitions in the cDNA at crosslink positions because cells are cultured with 4-thiouridin, a nucleotide analog that is misread by the reverse transcriptase [48]. Individual-nucleotide resolution CLIP (iCLIP) [49] even achieves single-nucleotide resolution because it does not discard truncated cDNAs where reverse transcription stopped at the crosslinked nucleotide. Instead, in the iCLIP protocol, 5' adapters are ligated to all cDNAs through a circularisation step. The iCLIP read starts then indicate the crosslink position, resulting in high resolution. The original CLIP-seq protocol is very time consuming and recent developments are targeted at speed and convenience [46].

CLIP-seq data analysis

CLIP-seq read preprocessing is identical to the RNA-seq analysis pipeline because reads are controlled for base calling quality and adapter contamination. After quality control, the CLIP-seq reads are mapped against the reference genome. If the RBP of interest binds introns, read alignment is performed with a splice-aware aligner. Otherwise no splice junctions are allowed.

CLIP-seq reads often contain diagnostic deletions or substitutions at the position where the RBP was crosslinked to the RNA molecule. These events are caused by the reverse transcriptase skipping or misreading the crosslinked nucleotide in the cDNA generation step. Consequently, CLIP-seq reads are usually allowed to have one or two mismatches in their final alignment.

Peak calling is the next step after read alignment. The goal is to identify regions with pileups of more reads than expected from random background noise. Regions with large number of reads - the peaks - correspond to putative binding sites of the protein of interest because these RNA sequences were enriched among the set of all bound RNAs. Depending on the RBP, CLIP-seq can identify a few peaks or thousands of peaks, if the RBP does not have a strong sequence preference. The list of peaks are then used in downstream analyses to identify binding motifs or to study the locations and types of RNAs bound by the protein. If available, identified peaks can be compared to known motifs to verify the specificity of the experiment.

There are many different peak calling algorithms available with distinct assumptions and strategies for peak identification depending on the used CLIP-seq protocol. The most important distinction is what part of a read is used to define a peak [50]. For standard CLIP-seq experiments, peak callers use whole reads. For PAR-CLIP experiments, T to C transitions are used as additional information and for iCLIP experiments, only the read starts are used.

A common principle that is followed by most algorithms is that peak calling has to be done per gene because it depends on gene expression. A strongly expressed gene with many transcripts in the cell is much more likely to be bound by the RBP than a lowly expressed gene with only a few or no transcripts. Some peak callers can take RNA-seq data as input or gene expression estimates. Control CLIP-seq samples, where the IP step was omitted can also be used to correct for relative RNA abundances and to filter out low affinity binding sites.

In general, peak callers try to distinguish potential binding sites from the background read coverage. Any region with more reads than expected (using a p-value threshold) are then considered a peak. The read distribution is modeled with a probability distributions, the negative binomial (or variations thereof) is the most common distribution. ASPeak [51] and Piranha [52] are example of such peak callers. An alternative approach to model the background read distribution is permutation: the reads aligned to a specific genomic region (intron, exon, gene) are randomly redistributed multiple times. A false discovery rate (FDR) can then be computed by comparing the observed to the random read distribution [50]. iCount [53] and PyicoCLIP [54] are examples of permutation based peak callers. CLIPper [55] use a combination of both approaches: it first uses permutation to define a read threshold per gene to differentiate between signal (potential peaks) and noise. Secondly, CLIPper removes all potential peaks with less observed reads than would be significant under a Poisson distribution. The most recent tools, such as omniCLIP [56], try to utilize information from replicate samples to model biological variation and to seamlessly integrate control CLIP-seq or RNA-seq samples to correct for gene expression. omniCLIP uses a non homogenous hidden markov model to model the observed read coverage and the observed diagnostic events (dependent on the used CLIP-seq protocol). The output from peak calling is usually a file with the genomic regions considered to be peaks and associated numbers such as the number of reads and p-value or score indicating how likely the region is a true peak.

Motif discovery

Most RBPs bind specific sequence motifs or they have at least binding preferences for specific nucleotides. The peak sequences identified with peak callers can be used to identify the RBP binding preferences with motif prediction tools. These tools use the peak sequences as positive set and compare them against a set of background sequences that are not bound by the protein. The positive set is searched for enriched patterns that are absent in the background. The output from motif prediction is a position weight matrix (PWM) which is a motif representation that specifies the probability of observing each of the four nucleotides at each of the motif positions. Examples of motif finding tools are HOMER (Hypergeometric Optimization of Motif EnRichment) [57], which uses a hypergeometric test to identify enriched motifs in the positive sequences or DREME (Discriminative Regular Expression Motif Elicitation) [58], which uses a Fisher's exact test instead.

1.3 Neurodegeneration

Neurodegeneration refers to the progressive loss of neuronal function ultimately leading to death of neurons. Neurodegeneration is caused by diseases, but aging is one of the main risk factors for the development of neurodegenerative disorders. Examples of well known neurodegenerative diseases include Alzheimer's, Parkinson's or amyotrophic lateral sclerosis (ALS). Clinical and pathological symptoms are often similar or shared between different neurodegenerative diseases, which poses a challenge for disease classification [59]. During my PhD, I focused my research on ALS.

1.3.1 Amyotrophic lateral sclerosis

ALS is a progressive neurodegenerative diseases, characterized by the progressive loss of upper and lower motor neurons ultimately leading to death. Patient experience increasing muscle weakness often starting in hands and feet. The mean age of ALS onset varies between studies but reported numbers range from 51 to 66 years [60]. Survival time from ALS diagnosis to death depends on age at onset and ranges between 24 and 50 months on average [60]. ALS incidence varies between populations and in Europe, it affects 2.1 to 3.8 in 100,000 persons per year [60].

10% of ALS cases are familial (fALS) with clear indication of genetic inheritance whereas 90% of ALS cases are sporadic (sALS) without known family history [61]. The disease causing mutation can be identified in the majority of familial cases, but only a subset of sporadic cases. Around 60% of fALS cases can be traced back to mutations in either of the two genes Superoxide dismutase 1 (SOD1) [62] and chromosome 9 open reading frame 72 (C9ORF72) [63, 64]. Another 5% of fALS cases are caused by mutations in TAR-DNA binding protein 43 (TDP-43) [65] or fused in sarcoma (FUS) [66, 67], two primarily nuclear RNA binding proteins. Most of the ALS-linked mutations are inherited in an autosomal dominant pattern Gros-Louis *et al.* [68].

The main molecular characteristic of ALS are cytoplasmic ubiquitinated protein inclusions in affected neurons. Misfolded, phosphorylated and ubiquitinated TDP-43 has been identified as the major component of these inclusions [69, 70], however, TDP-43 negative and FUS positive inclusions have also been reported in patients with FUS mutations [71]. The pathological description of the TDP-43/FUS positive inclusions is TDP-43/FUS proteinopathy.

ALS and another neurodegenerative disease called frontotemporal dementia (FTD) are often described as the two ends of one disease spectrum [72]. The reason is that FTD shares genetic, clinical and phenotypic markers with ALS. FTD leads to the degeneration of the frontal and temporal lobes of the brain and causes death within three to four years [73]. FTD patients present with progressive changes in personality and behaviour or they exhibit progressive language dysfunctions [74]. About 15% of ALS patients also have cognitive or behavioural impairments, similar to FTD patients [75]. Pathologically, the most frequent form of FTD is characterised by TDP-43 positive ubiquitinated inclusions indicating an overlap with ALS. Lastly, FTD is also genetically linked with ALS, because mutations in C9ORF72 explain about 25% of familial FTD cases [73]. On the other hand, about 15% of FTD patients develop motor neuron dysfunction similar to ALS suggesting that the two diseases share a common cause in a subset of patients [73].

1.3.2 TDP-43

TDP-43 is a primarily nuclear protein that binds RNA and DNA. It consists of an N-terminal domain that is important for dimerisation [76], two RNA recognition motifs (RRM) and a glycine rich region at the C-terminus where most ALS causing mutations are located. TDP-43 can shuttle between nucleus and cytoplasm where it incorporates into stress granules [77].

TDP-43 has important roles in transcription and alternative splicing regulation, RNA transport and microRNA processing. TDP-43 binds many different RNA targets [78] including its own transcript for autoregulation via a negative feedback mechanism [79]. A binding preference for GU rich regions was revealed by TDP-43 CLIP-seq and the NMR structure of the RNA binding domain [80, 81].

In patients with TDP-43 proteinopathy, TDP-43 is aggregated in inclusions leading to increased cytoplasmic TDP-43 concentration and nuclear clearance. In affected cells, cytoplasmic aggregation and nuclear loss of function lead to increased TDP-43 production, because the autoregulation mechanism is impaired: newly produced TDP-43 is immediately sequestered in inclusions and cannot shuttle back into the nucleus to regulate protein production leading to a vicious circle [82].

The loss of nuclear TDP-43 function presumably drives disease progression. Loss of TDP-43 has been associated with splicing defects [80] and Ling *et al.* [18] identified cryptic exons in TDP-43 knockdown cells, as well as patient brains. TDP-43 is binding

GU rich regions in the introns surrounding the cryptic exons and thereby represses the inclusion during alternative splicing. If TDP-43 is depleted from cells, the cryptic exons are not repressed anymore and can be detected by RNA-seq. Most cryptic exons cause frameshifts of premature termination codons leading to degradation via NMD and disrupted protein translation.

1.3.3 FUS

FUS is another RNA binding protein with structural similarity to TDP-43. It also contains a glycine rich region and an RNA binding domain consisting of an RRM, two glycine/arginie rich motifs and a zinc-finger (ZnF) domain [83]. Most known FUS mutations that are causative for ALS are located in the glycine rich region and the C-terminal nuclear localisation signal (NLS) [66, 67]. FUS and two other proteins EWS (Ewing's sarcoma protein) and TAF15 (TATA box binding protein associated factor) form the FET family of proteins with shared domain composition, function and cellular localisation [84]. FUS is a transcription and splicing factor [78] with additional roles in mRNA transport and stability, as well as DNA damage repair. FUS is mainly localised in the nucleus where it is actively imported by Transportin but the protein can leave the nucleus via passive diffusion [85]. FUS has been reported in neuronal processes, axons and at the synaptic site [86, 87]. Cytoplasmic roles of FUS include mRNA transport, RNA stabilisation and the formation of stress granules. It has been proposed that FUS might regulate synaptic translation via mRNA transport to the synaptic site. In a mouse model with ALS causing mutations in the FUS gene, axonal protein synthesis was reduced, suggesting a role of FUS in local translation [87]. FUS binds downstream of polyadenylation sites and enhances polyadenylation by recruiting CPSF160, a polyadenylation factor thereby regulating the stability of the mRNA [88].

Similar to TDP-43, FUS autoregulates its own expression via a negative feedback loop by binding to its own pre-mRNA. ALS causing FUS mutations in the nuclear localisation signal have been shown to compromise the autoregulation, leading to increased levels of cytoplasmic FUS [89, 90].

The nuclear RNA targets of FUS have been previously identified with different CLIP methods, including PAR-CLIP [91], iCLIP [92] and CLIP-seq [78, 88, 93, 94]. FUS binds mainly introns and 3'UTRs without a strong sequence specificity. There is some controversy about the sequence and structure preference of FUS binding sites: Some

studies report a binding to AU-rich stem-loops [92, 93], to single-stranded RNA [92] or to GU-rich motifs [78, 88, 92, 94] although with much lower affinity than other sequence specific RBPs. Three dimensional solution structures resolved by NMR of FUS domains bound to RNA confirmed that the FUS ZnF binds to GGU and the RRM binds RNA stem loops [95].

1.4 Goals of the thesis

My PhD is a joint project between a statistical bioinformatics group and an ALS research lab. For my PhD, we defined the following goals:

- Collaboration with Sonu Sahadevan to study the role of FUS at the synaptic site. I was responsible for all bioinformatic analyses:
  - CLIP-seq data analysis, including peak calling and differential comparison. This included data normalization.
  - Identification of synapse specific FUS RNA targets and comparison with nuclear RNA targets.
  - Motif analysis of synaptic FUS binding sites.
  - Differential gene expression analysis of RNA-seq data.
  - Comparison and integration of CLIP-seq and RNA-seq results.
  - Preparation of visualizations for the paper.
- Development of a pipeline for the prediction of cryptic exons in RNA-seq data. In particular:
  - Simulation of an RNA-seq data set with alternative splicing events and corresponding gene annotations.
  - Comparison of splice-aware genome alignment methods to find the best tool and parameter combination for the discovery of novel events.
  - Method development for the detection of unannotated exons from aligned RNA-seq reads.
  - Implementation of the prediction method in R.
  - Evaluation of the prediction method using the simulated data and comparison

to an already published method.

– Application and prediction of novel splicing events, especially cryptic exons, in already published ALS specific RNA-seq data sets.

1.5 Thesis outline

This thesis consists of four different papers. The first and the second are published and the third and fourth are manuscripts in preparation. My contribution to each of the four papers is summarized at the beginning of each chapter. The thesis concludes with a summary and outlook.

1.6 References

[1] E. Bianconi, A. Piovesan, F. Facchin *et al.* "An estimation of the number of cells in the human body". *Annals of Human Biology* **40**:6 (2013), pp. 463–471.

[2] J. D. Marth. "A unified vision of the building blocks of life". *Nature Cell Biology* **10**:9 (2008), p. 1015.

[3] J. E. Darnell, L. Philipson, R. Wall, and M. Adesnik. "Polyadenylic acid sequences: Role in conversion of nuclear RNA into messenger RNA". *Science* **174**:4008 (1971), pp. 507–510.

[4] Y. Wang, J. Liu, B. Huang *et al.* "Mechanism of alternative splicing and its regulation". *Biomedical Reports* **3**:2 (2015), pp. 152–158.

[5] L. V. Sharova, A. A. Sharov, T. Nedorezov *et al.* "Database for mRNA half-life of 19 977 genes obtained by DNA microarray analysis of pluripotent and differentiating mouse embryonic stem cells". *DNA Research* **16**:1 (2009), pp. 45–58.

[6] E. Yang, E. van Nimwegen, M. Zavolan *et al.* "Decay rates of human mRNAs: Correlation with functional characteristics and sequence attributes". *Genome Research* **13**:8 (2003), pp. 1863–1872.

[7] A. F. Palazzo and E. S. Lee. "Non-coding RNA: What is functional and what is junk?" *Frontiers in Genetics* **5**:JAN (2015), p. 2.

[8] H. Jung, D. Lee, J. Lee *et al.* "Intron retention is a widespread mechanism of tumor-suppressor inactivation". *Nature Genetics* **47**:11 (2015), pp. 1242–1248.

[9] J. J. Wong, W. Ritchie, O. A. Ebner *et al.* "Orchestrated intron retention regulates normal granulocyte differentiation". *Cell* (2013).

[10] F. Lejeune and L. E. Maquat. *Mechanistic links between nonsense-mediated mRNA decay and pre-mRNA splicing in mammalian cells*. 2005.

[11] A. M. Resch, A. Y. Ogurtsov, I. B. Rogozin, S. A. Shabalina, and E. V. Koonin. "Evolution of alternative and constitutive regions of mammalian 5' UTRs". *BMC Genomics* **10** (2009), pp. 1–14.

[12] L. W. Barrett, S. Fletcher, and S. D. Wilton. *Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements*. 2012.

[13] M. Irimia, R. J. Weatheritt, J. D. Ellis *et al.* "A highly conserved program of neuronal microexons is misregulated in autistic brains". *Cell* **159**:7 (2014), pp. 1511–1523.

[14] A. Torres-Méndez, S. Bonnal, Y. Marquez *et al.* "A novel protein domain in an ancestral splicing factor drove the evolution of neural microexons". *Nature Ecology and Evolution* **3**:4 (2019), pp. 691–701.

[15] Y. I. Li, L. Sanchez-Pulido, W. Haerty, and C. P. Ponting. "RBFOX and PTBP1 proteins regulate the alternative splicing of micro-exons in human brain transcripts". *Genome Research* **25**:1 (2015), pp. 1–13.

[16] C. R. Sibley, L. Blazquez, and J. Ule. "Lessons from non-canonical splicing". *Nature Reviews Genetics* (2016).

[17] J. P. Ling, R. Chhabra, J. D. Merran *et al.* "PTBP1 and PTBP2 Repress Nonconserved Cryptic Exons". *Cell Reports* **17**:1 (2016), pp. 104–113.

[18] J. P. Ling, O. Pletnikova, J. C. Troncoso, and P. C. Wong. "TDP-43 repression of nonconserved cryptic exons is compromised in ALS-FTD". *Science (New York, N.Y.)* **349**:6248 (2015), pp. 650–655.

[19] R. Lister, R. C. O'Malley, J. Tonti-Filippini *et al.* "Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis". *Cell* **133**:3 (2008), pp. 523–536.

[20] U. Nagalakshmi, Z. Wang, K. Waern *et al.* "The transcriptional landscape of the yeast genome de-

fined by RNA sequencing". *Science* **320**:5881 (2008), pp. 1344–1349.

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

[22] C. W. Fuller, L. R. Middendorf, S. A. Benner *et al.* "The challenges of sequencing by synthesis". *Nature Biotechnology* **27**:11 (2009), pp. 1013–1023.

[23] S. Andrews. *FastQC: a quality control tool for high throughput sequence data*. 2010. URL: `https://www.bioinformatics.babraham.ac.uk/projects/fastqc/`.

[24] M. Martin. "Cutadapt removes adapter sequences from high-throughput sequencing reads". *EMBnet.journal* **17**:1 (2011), pp. 10–12.

[25] M. Dodt, J. T. Roehr, R. Ahmed, and C. Dieterich. "FLEXBAR-Flexible Barcode and Adapter Processing for Next-Generation Sequencing Platforms." *Biology* **1**:3 (2012), pp. 895–905.

[26] A. Dobin, C. A. Davis, F. Schlesinger *et al.* "STAR: Ultrafast universal RNA-seq aligner". *Bioinformatics* **29**:1 (2013), pp. 15–21.

[27] D. Kim, J. M. Paggi, C. Park, C. Bennett, and S. L. Salzberg. "Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype". *Nature Biotechnology* **37**:8 (2019), pp. 907–915.

[28] Y. Liao, G. K. Smyth, and W. Shi. "FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features". *Bioinformatics* **30**:7 (2014), pp. 923–930.

[29] S. Anders, P. T. Pyl, and W. Huber. "HTSeq-A Python framework to work with high-throughput sequencing data". *Bioinformatics* **31**:2 (2015), pp. 166–169.

[30] R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford. "Salmon provides fast and bias-aware quantification of transcript expression". *Nature Methods* **14**:4 (2017), pp. 417–419.

[31] N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter. "Near-optimal RNA-Seq quantification". *aRxiv* October 2015 (2015), pp. 4–7.

[32] D. C. Wu, J. Yao, K. S. Ho, A. M. Lambowitz, and C. O. Wilke. "Limitations of alignment-free tools in total RNA-seq quantification". *BMC Genomics* **19**:1 (2018), pp. 1–14.

[33] C. Zhang, B. Zhang, L. L. Lin, and S. Zhao. "Evaluation and comparison of computational tools for RNA-seq isoform quantification". *BMC Genomics* **18**:1 (2017), pp. 1–11.

[34] M. D. Robinson, D. J. McCarthy, and G. K. Smyth. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data". *Bioinformatics* **26**:1 (2010), pp. 139–140.

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

[36] D. J. McCarthy, Y. Chen, and G. K. Smyth. "Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation". *Nucleic Acids Research* **40**:10 (2012), pp. 4288–4297.

[37] E. T. Wang, R. Sandberg, S. Luo *et al.* "Alternative isoform regulation in human tissue transcriptomes". *Nature* **456**:7221 (2008), pp. 470–476.

[38] J. P. Venables, R. Klinck, A. Bramard *et al.* "Identification of alternative splicing markers for breast cancer". *Cancer Research* **68**:22 (2008), pp. 9525–9531.

[39] S. Heber, M. Alekseyev, S. H. Sze, H. Tang, and P. A. Pevzner. "Splicing graphs and EST assembly problem". *Bioinformatics* **18** (2002), S181–S188.

[40] L. H. Legault and C. N. Dewey. "Inference of alternative splicing from RNA-Seq data with probabilistic splice graphs". *Bioinformatics* **29**:18 (2013), pp. 2300–2310.

[41] C. Trapnell, B. A. Williams, G. Pertea *et al.* "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation". *Nature Biotechnology* **28**:5 (2010), pp. 511–515.

[42] Q. Tan, H. K. Yalamanchili, J. Park *et al.* "Extensive cryptic splicing upon loss of RBM17 and TDP43 in neurodegeneration models". *Human Molecular Genetics* **25**:23 (2016), pp. 5083–5093.

[43] J. Humphrey, W. Emmett, P. Fratta, A. M. Isaacs, and V. Plagnol. "Quantitative analysis of cryptic splicing associated with TDP-43 depletion". *BMC Medical Genomics* **10**:1 (2017), pp. 1–21.

[44] J. Ule, K. B. Jensen, M. Ruggiu *et al.* "CLIP Identifies Nova-Regulated RNA Networks in the Brain". *Science* **302**:5648 (2003), pp. 1212–1215.

[45] G. W. Yeo, N. G. Coufal, T. Y. Liang *et al.* "An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells". *Nature Structural and Molecular Biology* **16**:2 (2009), pp. 130–137.

[46] F. C. Y. Lee and J. Ule. "Technology Review Advances in CLIP Technologies for Studies of Protein-RNA Interactions". *Molecular Cell* **69**:3 (2018), pp. 354–369.

[47] M. Ramanathan, D. F. Porter, and P. A. Khavari. "Methods to study RNA – protein interactions". *Nature Methods* **16**:March (2019).

[48] M. Hafner, M. Landthaler, L. Burger *et al.* "Transcriptome-wide Identification of RNA-Binding Protein and MicroRNA Target Sites by PAR-CLIP". *Cell* **141**:1 (2010), pp. 129–141.

[49] J. König, K. Zarnack, G. Rot *et al.* "iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution". *Nature structural & ...* **17**:7 (2010), pp. 909–915.

[50] A. M. Chakrabarti, N. Haberman, A. Praznik, N. M. Luscombe, and J. Ule. "Data Science Issues in Studying Protein – RNA Interactions with CLIP Technologies". April (2018), pp. 235–261.

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

[52] P. J. Uren, E. Bahrami-Samani, S. C. Burns *et al.* "Site identification in high-throughput RNA-protein interaction data". *Bioinformatics* **28**:23 (2012), pp. 3013–3020.

[53] T. Curk, G. Rot, Č. Gorup *et al. iCount: protein-RNA interaction iCLIP data analysis*. URL: `https://github.com/tomazc/iCount`.

[54] S. Althammer, J. González-Vallinas, C. Ballaré *et al.* "Pyicos: a versatile toolkit for the analysis of high-throughput sequencing data". **27**:24 (2011), pp. 3333–3340.

[55] M. T. Lovci, D. Ghanem, H. Marr *et al.* "Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges". *Nature Publishing Group* **20**:12 (2013), pp. 1434–1442.

[56] P. Drewe-boss. "omniCLIP: probabilistic identification of protein-RNA interactions from CLIP-seq data" (2018), pp. 1–4.

[57] S. Heinz, C. Benner, N. Spann *et al.* "Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities". *Molecular Cell* **38**:4 (2010), pp. 576–589.

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

[59] S. Przedborski, M. Vila, and V. Jackson-Lewis. "Neurodegeneration: What is it and where are we?" *Journal of Clinical Investigation* **111**:1 (2003), pp. 3–10.

[60] E. Longinetti and F. Fang. "Epidemiology of amyotrophic lateral sclerosis". *Current Opinion in Neurology* **32**:5 (2019), pp. 771–776.

[61] L. P. Rowland and N. A. Shneider. "Amyotrophic Lateral Sclerosis". *English Journal* **344**:22 (2001), pp. 1688–1700.

[62] D. R. Rosen, T. Siddique, D. Patterson *et al.* "Mutations in Cu/Zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis". *Nature* **362**:6415 (1993), pp. 59–62.

[63] M. DeJesus-Hernandez, I. R. Mackenzie, B. F. Boeve *et al.* "Expanded GGGGCC Hexanucleotide Repeat in Noncoding Region of C9ORF72 Causes Chromosome 9p-Linked FTD and ALS". *Neuron* **72**:2 (2011), pp. 245–256.

[64] A. E. Renton, E. Majounie, A. Waite *et al.* "A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD". *Neuron* **72**:2 (2011), pp. 257–268.

[65] E. Kabashi, P. N. Valdmanis, P. Dion *et al.* "TARDBP mutations in individuals with sporadic and familial amyotrophic lateral sclerosis". *Nature Genetics* **40**:5 (2008), pp. 572–574.

[66] T. J. Kwiatkowski, D. A. Bosco, A. L. LeClerc *et al.* "Mutations in the FUS/TLS gene on chromosome 16 cause familial amyotrophic lateral sclerosis". *Science* **323**:5918 (2009), pp. 1205–1208.

[67] C. Vance, B. Rogelj, T. Hortobágyi *et al.* "Mutations in FUS, an RNA processing protein, cause familial amyotrophic lateral sclerosis type 6". *Science* **323**:5918 (2009), pp. 1208–1211.

[68] F. Gros-Louis, C. Gaspar, and G. A. Rouleau. *Genetics of familial and sporadic amyotrophic lateral sclerosis*. 2006.

[69] T. Arai, M. Hasegawa, H. Akiyama *et al.* "TDP-43 is a component of ubiquitin-positive tau-negative inclusions in frontotemporal lobar degeneration and amyotrophic lateral sclerosis". *Biochemical and Biophysical Research Communications* **351**:3 (2006), pp. 602–611.

[70] M. Neumann, D. M. Sampathu, L. K. Kwong *et al.* "Ubiquitinated TDP-43 in frontotemporal lobar degeneration and amyotrophic lateral sclerosis". *Science* **314**:5796 (2006), pp. 130–133.

[71] R. Rademakers, H. Stewart, M. Dejesus-Hernandez *et al.* "FUS gene mutations in familial and sporadic amyotrophic lateral sclerosis". *Muscle and Nerve* **42**:2 (2010), pp. 170–176.

[72] S. C. Ling, M. Polymenidou, and D. W. Cleveland. *Converging mechanisms in als and FTD: Disrupted RNA and protein homeostasis*. 2013.

[73] A. S. Ng, R. Rademakers, and B. L. Miller. "Frontotemporal dementia: A bridge between dementia and neuromuscular disease". *Annals of the New York Academy of Sciences* **1338**:1 (2015), pp. 71–93.

[74] G. M. McKhann. "Clinical and Pathological Diagnosis of Frontotemporal Dementia". *Archives of Neurology* **58**:11 (2001), p. 1803.

[75] G. M. Ringholz, ; S. H. Appel, ; M. Bradshaw *et al.* *Prevalence and patterns of cognitive impairment in sporadic ALS*. Tech. rep. 2005.

[76] T. Afroz, E. M. Hock, P. Ernst *et al.* "Functional and dynamic polymerization of the ALS-linked protein TDP-43 antagonizes its pathologic aggregation". *Nature Communications* **8**:1 (2017), pp. 1–14.

[77] C. Colombrita, E. Zennaro, C. Fallini *et al.* "TDP-43 is recruited to stress granules in conditions of oxidative insult". *Journal of Neurochemistry* **111**:4 (2009), pp. 1051–1061.

[78] C. Lagier-Tourenne, M. Polymenidou, K. R. Hutt *et al.* "Divergent roles of ALS-linked proteins FUS/TLS and TDP-43 intersect in processing long pre-mRNAs". *Nature Neuroscience* **15**:11 (2012), pp. 1488–97.

[79] Y. M. Ayala, L. De Conti, S. E. Avendaño-Vázquez *et al.* "TDP-43 regulates its mRNA levels through a negative feedback loop". *The EMBO Journal* **30**:2 (2011), pp. 277–288.

[80] M. Polymenidou, C. Lagier-tourenne, K. R. Hutt *et al.* "Long pre-mRNA depletion and RNA missplicing contribute to neuronal vulnerability from loss of TDP-43". *Nature Neuroscience* **14**:4 (2011), pp. 459–468.

[81] P. J. Lukavsky, D. Daujotyte, J. R. Tollervey *et al.* "Molecular basis of UG-rich RNA recognition by the human splicing factor TDP-43". *Nature Structural and Molecular Biology* **20**:12 (2013), pp. 1443–1449.

[82] M. Budini, E. Buratti, M. Budini, and E. Buratti. "TDP-43 Autoregulation: Implications for Disease". *J Mol Neurosci* **45** (2011), pp. 473–479.

[83] C. Lagier-Tourenne, M. Polymenidou, and D. W. Cleveland. "TDP-43 and FUS/TLS: emerging roles in RNA processing and neurodegeneration". *Human Molecular Genetics* **19**:R1 (2010), R46–R64.

[84] A. Y. Tan and J. L. Manley. "The TET Family of Proteins: Functions and Roles in Disease". *Journal of Molecular Cell Biology* **1**:2 (2009), pp. 82–92.

[85] E. M. Hock, Z. Maniecka, M. Hruska-Plochan *et al.* "Hypertonic Stress Causes Cytoplasmic Transloca-

tion of Neuronal, but Not Astrocytic, FUS due to Impaired Transportin Function". *Cell Reports* **24**:4 (2018), pp. 987–1000.

[86] M. Schoen, J. M. Reichel, M. Demestre *et al.* "Super-resolution microscopy reveals presynaptic localization of the ALS/FTD related protein FUS in hippocampal neurons". *Frontiers in Cellular Neuroscience* **9**:JAN2016 (2016), pp. 1–16.

[87] J. López-Erauskin, T. Tadokoro, M. W. Baughn *et al.* "ALS/FTD-Linked Mutation in FUS Suppresses Intra-axonal Protein Synthesis and Drives Disease Without Nuclear Loss-of-Function of FUS". *Neuron* **100**:4 (2018), pp. 816–830.

[88] A. Masuda, J. I. Takeda, T. Okuno *et al.* "Position-specific binding of FUS to nascent RNA regulates mRNA length". *Genes and Development* **29**:10 (2015), pp. 1045–1057.

[89] Y. Zhou, S. Liu, G. Liu, A. Öztürk, and G. G. Hicks. "ALS-Associated FUS Mutations Result in Compromised FUS Alternative Splicing and Autoregulation". *PLoS Genetics* **9**:10 (2013).

[90] J. Humphrey, N. Birsa, C. Milioto *et al.* "FUS ALS-causative mutations impact FUS autoregulation and the processing of RNA-binding proteins through intron retention" (2019).

[91] J. I. Hoell, E. Larsson, S. Runge *et al.* "RNA targets of wild-type and mutant FET family proteins". *Nature Structural and Molecular Biology* **18**:12 (2011), pp. 1428–1431.

[92] B. Rogelj, L. E. Easton, G. K. Bogu *et al.* "Widespread binding of FUS along nascent RNA regulates alternative splicing in the brain". *Scientific Reports* **2** (2012), pp. 1–10.

[93] S. Ishigaki, A. Masuda, Y. Fujioka *et al.* "Position-dependent FUS-RNA interactions regulate alternative splicing events and transcriptions". *Scientific Reports* **2** (2012), pp. 1–9.

[94] T. Nakaya, P. Alexiou, M. Maragkakis, A. Chang, and Z. Mourelatos. "FUS regulates genes coding for RNA-binding proteins in neurons by binding to their highly conserved introns". *Rna* **19**:4 (2013), pp. 498–509.

[95] F. E. Loughlin, P. J. Lukavsky, T. Kazeeva *et al.* "The Solution Structure of FUS Bound to RNA Reveals a Bipartite Mode of RNA Recognition with Both Sequence and Shape Specificity". *Molecular Cell* **73**:3 (2019), pp. 490–504.

Part II.

Scientific Contributions

2 RNA Sequencing Data: Hitchhiker's Guide to Expression Analysis[1]

This review article introduces state-of-the-art RNA sequencing methods and emerging technologies. We explain recent developments in RNA-seq data analysis, including read alignment and transcript/gene expression quantification. We also introduce the theory behind preprocessing and normalisation for differential gene expression analysis, as well as the statistical inference.

I wrote the introduction chapter together with Mark Robinson. I designed and created figure 1 and analysed public data to generate the ridge plots in figure 2. All authors edited the whole manuscript.

[1]Originally published in Van den Berge, K., Hembach, K. M., Soneson, C., Tiberi, S., Clement, L., Love, M. I., Patro, R., & Robinson, M. D. "RNA Sequencing Data: Hitchhiker's Guide to Expression Analysis" *Annual Review of Biomedical Data Science*. **2**:1 (2019), pp. 139–173.

Annual Review of Biomedical Data Science

RNA Sequencing Data: Hitchhiker's Guide to Expression Analysis

Koen Van den Berge,[1,*] Katharina M. Hembach,[2,*] Charlotte Soneson,[2,3,*] Simone Tiberi,[2,*] Lieven Clement,[1,†] Michael I. Love,[4,†] Rob Patro,[5,†] and Mark D. Robinson[2,†]

[1]Bioinformatics Institute Ghent and Department of Applied Mathematics, Computer Science and Statistics, Ghent University, 9000 Ghent, Belgium

[2]Institute of Molecular Life Sciences and SIB Swiss Institute of Bioinformatics, University of Zurich, 8057 Zurich, Switzerland; email: mark.robinson@imls.uzh.ch

[3]Current Affiliation: Friedrich Miescher Institute for Biomedical Research and SIB Swiss Institute of Bioinformatics, 4058 Basel, Switzerland

[4]Department of Biostatistics and Department of Genetics, University of North Carolina, Chapel Hill, North Carolina 27514, USA

[5]Department of Computer Science, Stony Brook University, Stony Brook, New York 11794, USA

Annu. Rev. Biomed. Data Sci. 2019. 2:139–73

First published as a Review in Advance on April 30, 2019

The *Annual Review of Biomedical Data Science* is online at biodatasci.annualreviews.org

https://doi.org/10.1146/annurev-biodatasci-072018-021255

*These authors contributed equally to this article

†These authors contributed equally to this article

Keywords

RNA sequencing, gene expression, high-dimensional data, differential expression analysis, expression quantification

Abstract

Gene expression is the fundamental level at which the results of various genetic and regulatory programs are observable. The measurement of transcriptome-wide gene expression has convincingly switched from microarrays to sequencing in a matter of years. RNA sequencing (RNA-seq) provides a quantitative and open system for profiling transcriptional outcomes on a large scale and therefore facilitates a large diversity of applications, including basic science studies, but also agricultural or clinical situations. In the past 10 years or so, much has been learned about the characteristics of the RNA-seq data sets, as well as the performance of the myriad of methods developed. In this review, we give an overview of the developments in RNA-seq data analysis, including experimental design, with an explicit focus on the quantification of gene expression and statistical approaches

for differential expression. We also highlight emerging data types, such as single-cell RNA-seq and gene expression profiling using long-read technologies.

INTRODUCTION: OVERVIEW OF THE RNA SEQUENCING ASSAY

> After that it gets a bit complicated, and there's all sort of stuff going on in dimensions thirteen to twenty-two that you really wouldn't want to know about. All you really need to know for the moment is that the universe is a lot more complicated than you might think, even if you start from a position of thinking it's pretty damn complicated in the first place.
>
> —*Mostly Harmless* by Douglas Adams

Molecular biologists use gene expression studies to get a snapshot of the RNA molecules present in a biological system, which dictates what cells are doing or are capable of. The original RNA sequencing (RNA-seq) protocols, published over 10 years ago (1–5), described the sequencing of complementary DNA (cDNA) fragments on a large scale from a population of cells. Since then, the system has been optimized for different types and qualities of starting material, as well as different research questions, and many mature protocols are available.

A basic overview of the main steps in a standard RNA-seq experiment is given in **Figure 1**. The first step is the extraction and purification of RNA from a sample, followed by an enrichment of target RNAs. Most commonly used is poly(A) capture, to select for polyadenylated RNAs, or ribosomal depletion, to deplete ribosomal and transfer RNAs that are highly abundant in a cell (approximately 95% of total RNA) (6) and are usually not of primary interest (7). The selected

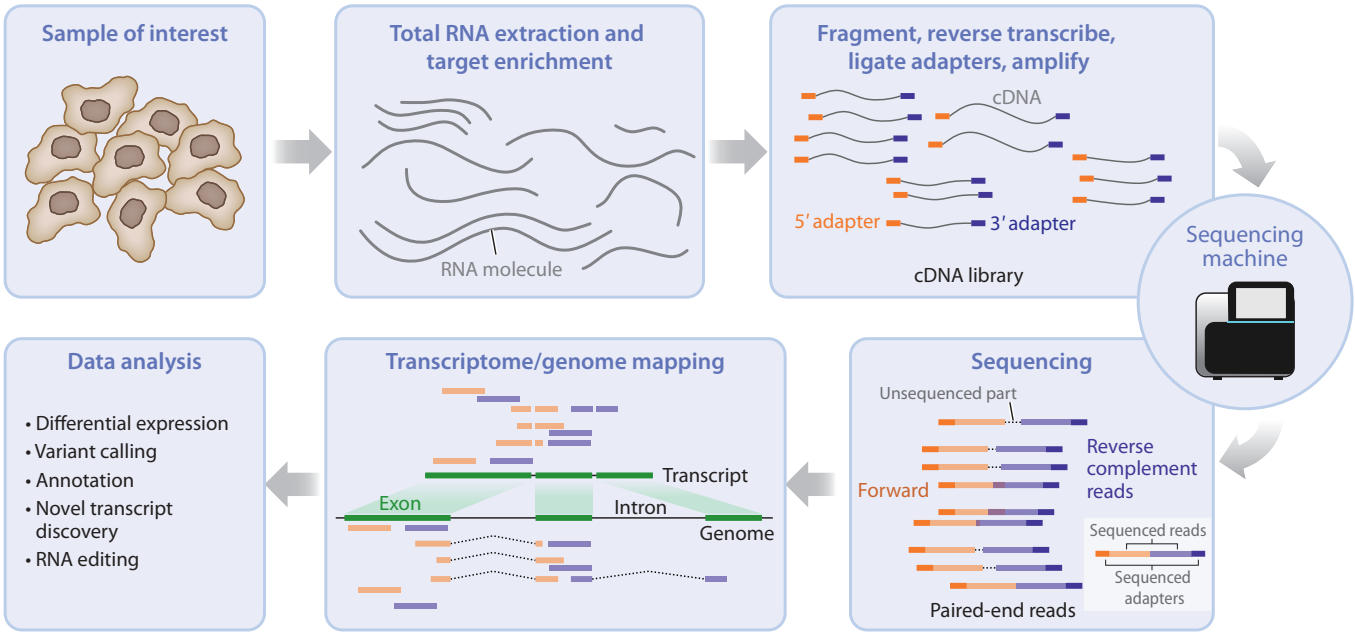


Figure 1

Overview of the experimental steps in an RNA sequencing (RNA-seq) protocol. The complementary DNA (cDNA) library is generated from isolated RNA targets and then sequenced, and the reads are mapped against a reference genome or transcriptome. Downstream data analysis depends on the goal of the experiment and can include, among other things, assessing differential expression, variant calling, or genome annotation.

RNAs are then chemically or enzymatically fragmented to molecules of appropriate size (e.g., 300–500 bp for Illumina's TruSeq). Current dominant systems (e.g., Illumina) only sequence DNA; single-stranded target RNAs are thus reverse-transcribed to cDNA (first strand), the RNA is then degraded, and the first-strand cDNA is complemented to a double strand. Adapter sequences are either ligated to the 3′ and 5′ end of the double-stranded cDNA or used as primers in the reverse transcription reaction. The final cDNA library consists of cDNA inserts flanked by an adapter sequence on each end. In the last step, the cDNA library is amplified by polymerase chain reaction (PCR) using parts of the adapter sequences as primers.

For Illumina sequencing, the library is loaded onto a flow cell where the cDNAs bind to short oligonucleotides complementary to the adapter sequence. Bridge amplification creates dense clonal clusters of each cDNA loaded (8). The sequence of each cluster is determined by a process called sequencing by synthesis (9): Single-stranded templates are read as the complementary strand is generated. A single fluorescently labeled deoxynucleoside triphosphate (dNTP) is added in each step. The label acts as a terminator and prevents the incorporation of more than one dNTP at the same time. After the fluorescent label has been imaged, it is enzymatically cleaved and the next dNTP can bind to the chain. Base calls are inferred directly from the measured fluorescent signal intensity.

cDNA libraries can be sequenced in one of two modes: single-end or paired-end. In single-end mode, only one end of the cDNA insert is sequenced, whereas in paired-end mode, both ends are sequenced, yielding two reads in opposite orientation, one from each end.

There are protocols for unstranded and stranded RNA-seq (10, 11), where the latter preserves information about the coding strand of each fragment, which is useful in compact genomes or with expressed RNAs that originate from opposite strands of the same genomic locus. One possibility to construct a stranded library is to use deoxyuridine triphosphates (dUTPs) in the generation of the second strand cDNA and to degrade the dUTP-labeled cDNA before PCR amplification (12). Other protocols use alternative adapters to distinguish between 5′ and 3′ ends of the RNA (13).

RNA-seq has greatly evolved over time, with early experiments having 35-bp reads and modern (Illumina-based) experiments typically employing 50-bp (single-end) or 100-bp (paired-end) reads (**Figure 2*a***). Most RNA-seq experiments comprise between 10 and 100 million reads, with a trend toward deeper sequencing over time (**Figure 2*b***). The number of samples per project has remained constant over the years, with a median of around eight samples (**Figure 2*c***). Rapid enhancements in sequencing technology have enabled not only longer read lengths (e.g., 250–300 bp for Illumina's MiSeq) and much higher throughput for the same cost, but also much lower amounts of required starting material. Meanwhile, third-generation technologies, such as Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT), allow the sequencing of single molecules and have now been used for sequencing full-length transcripts on a transcriptome-wide scale (14). Further developments are summarized below in the section titled Long-Read Transcriptome Sequencing. In addition, single-cell RNA-seq (scRNA-seq) is a rapidly emerging technique that can be used to sequence the sparse transcriptome of individual cells. Some of the early developments in this area are captured in the section titled Single-Cell Transcriptome Sequencing.

Design Aspects of RNA-seq

The basics of scientific experimental design apply equally for RNA-seq experiments (e.g., see Reference 16). For example, whether the desired experiment is a simple two-group design or a full factorial design, one should consider randomizing experimental units to treatments to avoid confounding factors (e.g., via blocking over batches). If the experiment is run in multiple batches

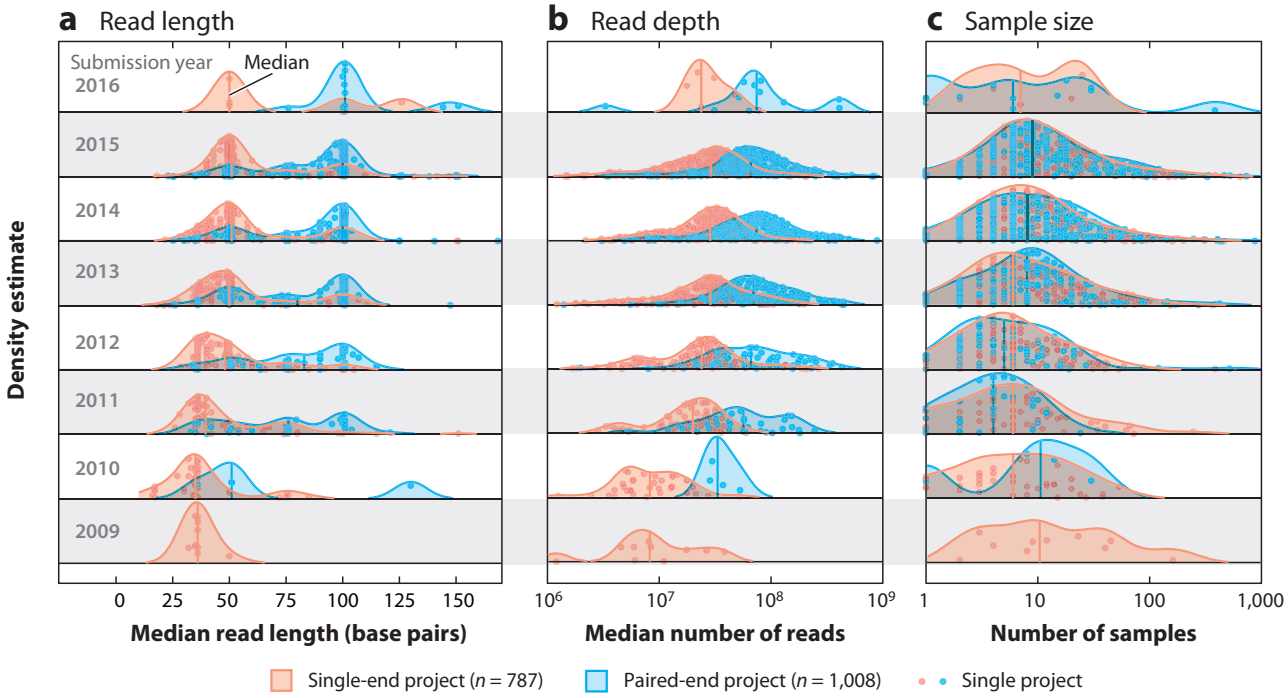


Figure 2

Ridge plots showing the progression of read length, depth, and sample size in Sequence Read Archive (SRA) projects using the `recount` package (15). The projects are separated by the submission year of the biosample. (*a*) Median read length of all samples per project and year. (*b*) Median number of reads across all samples per project and year. (*c*) Number of samples in each project. Each point represents one project.

(e.g., a limited number of samples per run), it is critical to represent every experimental condition in each batch, so that, when comparing conditions, differences within a batch can be averaged over in the statistical modeling.

Specific aspects to be considered while designing an RNA-seq experiment include the number of replicates and the depth of sequencing. Ultimately, in modern genomic experiments where resources (e.g., material from subjects) are scarce and the RNA-seq experiment is itself a hypothesis-generating tool, the first driver of sample size is budget. Many RNA-seq studies use as few as three replicates per condition (**Figure 2*c***), near the minimum required to do any statistical analysis.

Sample size calculators can compute the required number of samples to achieve a user-defined power for detecting differential expression (DE) (17–20). However, the user must define many parameters, such as the expected alignment rate, the desired power, the significance level, and the log-fold change (LFC) of DE genes. A recent study concluded that the recommended sample sizes vary across tools, even when estimates from pilot data are available (21). Another issue with sample size calculators is how to precisely define the outcome: Do we want to find as many DE genes as possible? Do we want a certain power for the lowly expressed genes or the highly expressed ones? In many cases, RNA-seq experiments are exploratory and thus a means to further experimentation.

Nonetheless, there is a trade-off between the number of samples and the sequencing depth in terms of discovery performance. Increasing the number of reads might seem always beneficial, but a large proportion of the reads originate from a small pool of highly expressed genes, and there is effectively no signal saturation. **Figure 3** highlights that more than 80% of reads are attributed to the 10% most expressed genes, acknowledging that transcript length also plays a role (22). An increased number of reads only marginally increases the coverage of lowly expressed genes,

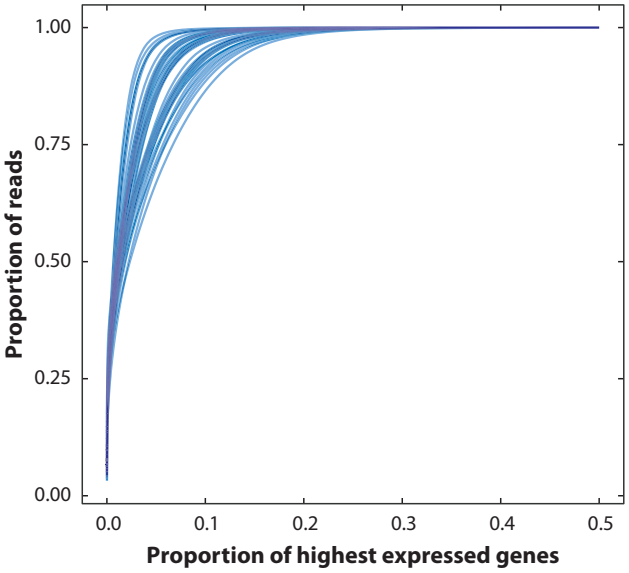


Figure 3

Cumulative proportion of reads among the top expressed genes. The *x*-axis orders genes according to the total number of reads they attract, and the *y*-axis displays the cumulative fraction of total reads. Each line represents a single sample. Counts were downloaded from `recount` (15), and 50 samples were randomly selected from accession number SRP060416.

and therefore, the statistical power to detect DE does not improve considerably, especially if the experiment already comprises ~10 million reads per sample (23). In most cases, the budget is better spent on replicates. For example, Schurch et al. (24) showed that a higher number of replicates is required to identify DE genes with low-fold change, and that ideally at least six replicates per condition should be used.

There are options for additional capture of genes with low expression, but these require additional labor and cost. In targeted RNA-seq (RNA CaptureSeq), specific regions are first captured by probes that are complementary to the region of interest and these selected regions are prepared and sequenced (25, 26). After capture, the quantitative nature of the assay is maintained (25); such capture is especially useful in degraded samples (e.g., patient material stored in paraffin blocks) where the poly(A) tails may not be present.

RNA-seq Applications

Clearly, the popularity of RNA-seq is driven by its large number of applications. One obvious application area is genome annotation. Even the well-studied transcriptomes of humans or model organisms such as mice, zebrafish, or fruit flies are not complete. Thus, transcriptomics is used to annotate novel transcriptional events, such as exon skipping, alternative 3′ acceptor or 5′ donor sites, or intron retention, and to understand their usage in normal, developmental, or pathological conditions. Transcriptomic studies identified previously unknown phenomena, such as microexons (27), cryptic exons (28), so-called skiptic exons (29), circular RNAs (30), enhancer RNAs (31), fusion genes (32), and so-called epitranscriptomics involving RNA base modifications (33).

One of the main application areas is gene regulation. RNA-seq enables the comparison of gene/transcript/exon expression between different tissues, cell types, genotypes, stimulation conditions, time points, disease states, growth conditions, and so on. Ultimately, the goal of such comparisons

is to identify the genes that change in expression to understand the molecular pathways that are used or altered or the regulatory components that are utilized.

Gene expression has been used for the molecular subclassification of cancer since the early days of microarrays (34). RNA-seq offers this same capacity but at higher resolution and can include, for example, categorization by splicing (35). There is considerable interest in using RNA-seq in clinical applications to augment or corroborate the information given by genome sequencing (36, 37). Other applications include spatial transcriptomics, where cellular positional information is maintained in the preparation of cDNA fragments (38); host–pathogen interactions via dual RNA-seq, where the transcriptomes of both host and pathogen are simultaneously assayed (39); the analysis of genetic variation among expressed genes (40); RNA editing events (41); the characterization of long noncoding RNAs (42); and metatranscriptomics (43).

Despite the many use cases for bulk RNA-seq, there are applications where single-cell resolution is desired, especially when studying heterogeneous tissues that consist of more than one cell type. While bulk RNA-seq can be computationally deconvoluted to estimate the composition of cells present (44), it is not possible to discover new cell types or perform cell-type-specific analyses with bulk RNA-seq, and thus scRNA-seq opens the door to new applications.

Outline

This review focuses on data analysis aspects, the computational steps involved (focusing on DE), various statistical and computational challenges, and the approaches that have been proposed to address them. We focus on Illumina-based RNA-seq data on model organisms, as that is the dominant application area. There are already excellent reviews for major application or computational areas, such as de novo (or reference-based) transcriptome assembly (45), allele-specific expression analyses (46), expression quantitative trait loci mapping (47), splicing (48), analysis of gene regulatory networks (49), and pathway analyses (50, 51). In most applications, the overarching goal is to identify DE, at either the gene, transcript, or exon level. The set of DE entities provides a snapshot into the molecular underpinnings of a stimulus, a disease condition, a genetic mutation, or any other perturbation being interrogated. In most cases, DE is only an intermediate (although critical) step to understanding the biological system under study.

The review is organized as follows. First, we discuss alignment and quantification, where RNA-seq reads are placed in the context of the genome or annotation catalogs and the relative expression level of each target is assessed. Following quantification, we discuss the basics of DE, to lay the foundation for the current frameworks, and variants of DE, to highlight the diverse conceptual tools available to run the discovery process. Finally, we discuss two rapidly evolving research areas that have experienced considerable activity in recent years, single-cell transcriptome sequencing and long-read transcriptome sequencing (LRTS).

ALIGNMENT AND QUANTIFICATION

After an experiment has been conducted, the analyst is presented with files containing up to billions of short cDNA fragments. Following sufficient quality control of the sequencing reactions, alignment to a reference genome or (de novo assembled) transcriptome is one of the critical steps in translating the raw data into something quantitative.

Because the sequenced fragments are derived from cDNA corresponding to fully (or partially) spliced transcripts, reads often span the boundaries of splice junctions (SJs), resulting in so-called junction-spanning reads (**Figure 4**). This results in contiguous read sequences whose constituent subsequences may be separated by tens of thousands of nucleotides. This poses a considerable

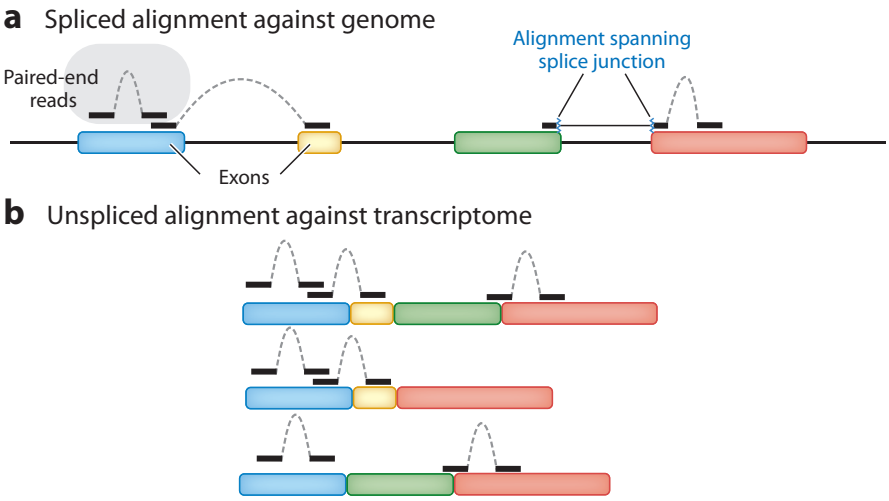


Figure 4

An illustration of spliced alignment of RNA sequencing (RNA-seq) fragments to a genome (*a*) and direct alignment to a transcriptome (*b*). Reads are designated by thick solid lines, while dashed arcs represent the pairing relationship between paired-end reads. This illustration depicts alignment to a single four-exon gene consisting of three distinct transcripts. In the spliced alignment (*a*), the left read of the rightmost pair is a junction-spanning alignment to the red–green exon boundary. In the direct alignment to the transcriptome (*b*), one observes how the same alignment (e.g., the alignment to the blue exon) is repeated for each transcript.

computational challenge, as the position of SJs in spanning reads needs to be accurately identified for a read to be properly aligned. There are two main approaches for handling spliced reads, each with its own challenges and benefits: a spliced alignment against a reference genome or an unspliced alignment against a reference transcriptome (a database of all isoforms). A main challenge in spliced alignment against a reference genome is the proper alignment of reads that span an SJ, especially when these junctions are not annotated a priori. Meanwhile, the main challenge in unspliced alignment to a transcriptome is the redundant sequence among related isoforms, which often leads to a high multimapping rate.

Spliced Alignment to a Reference Genome

A popular solution for handling RNA-seq alignments is to use a splice-aware aligner. Early RNA-seq aligners, e.g., TopHat (52), made use of DNA-seq aligners, such as Bowtie (53), by first building a catalog of putative SJs to which the reads can be directly aligned.

More recent splice-aware alignment tools (54–64) account for read splicing directly. They also can utilize the locations of known SJs and discover previously unannotated SJs. When a read partially aligns, the annotated SJ database is consulted to check if the alignment ends prematurely as the result of the read spanning a known splice site. In this case, compatible downstream splice sites can be considered as candidate loci to align the remaining portion of the read. Even if no annotated splice site exists at the point where the alignment ends, the tool can interrogate the terminal nucleotides in the partial alignment to see if they are compatible with known canonical (or user-provided) donor or acceptor sites, providing evidence that the partial alignment stops as the result of a splicing event.

One of the primary difficulties in aligning reads across SJs is that only a small portion of the read spans into one of the exons. Splice-aware aligners including STAR (55), HISAT(2) (56),

Subread (57), and GMAP (58) attempt to deal with such cases by using evidence from reads that confidently align across SJs. In such strategies, new SJs are added to the index when they display high-confidence evidence, i.e., when multiple reads with a sufficient anchoring sequence span the SJ. Trusted SJs are then used to help align reads that start or end near junction boundaries.

Unspliced Alignment to a Reference Transcriptome

In organisms where transcriptomes are well characterized, an alternative to splice-aware genome alignment is direct transcriptome alignment, which consists of aligning against a set of known transcripts. Since the transcript sequences are already spliced, reads should align contiguously, and many of the computationally expensive steps and heuristics can be avoided. Moreover, when no reasonable quality reference genome is available for reference-based transcript assembly (e.g., when a transcriptome has been assembled de novo), alignment directly to the assembled transcripts is the only available option. However, transcriptome alignment induces a high degree of multimapping, and dealing with this becomes a primary computational challenge. For example, if a gene has three distinct isoforms, a constitutive exon of this gene will appear three times in the transcriptome reference (e.g., **Figure 4*b***). Additionally, mapping only to annotated transcripts does not allow one to find novel splicing or expression patterns (e.g., novel exons), and it becomes difficult to assess retained introns or partial splicing; of course, it is possible to augment the transcriptome with unspliced variants. The choice of genome versus transcriptome alignment is largely driven by the desired target application and the constraints of downstream analyses.

Gene- and Transcript-Level Quantification From RNA-seq Data

One of the main uses of RNA-seq is to assess gene- and transcript-level abundances. Accurate abundance estimation is crucial to common downstream applications, including assessing all the notions of DE. Most commonly, abundances are estimated at the level of genes, but recently transcript-level abundances have become more widely used, and there are trade-offs in choosing between the two levels of resolution.

Gene-level quantification consists of assigning fragments (reads or read pairs) to genes, where the gene is often taken to represent the amalgamation of all transcripts produced from a specific strand at a specific locus (65), which typically share some exons or parts of exons. The total expression of a gene is the sum of the expression of its isoforms. Any fragment arising from any isoform of a gene is assigned to the underlying gene. There are typically two paths that can be taken to obtain gene-level quantifications: direct fragment overlap counting of gene features, and transcript-level quantification followed by aggregation to the gene level.

Direct fragment counting of gene features is done by first mapping RNA-seq reads to the genome with a splice-aware aligner, and then using a tool like featureCounts (66), HTSeq (67), or the built-in capability of STAR (55) to assess how many fragments overlap each gene; the same approach can be used to quantify other disjoint genomic features, such as nonoverlapping exonic segments. Even in this basic pipeline, there are many ways certain conditions can be handled. For example, should a fragment reside completely within a feature to be counted? If a fragment maps to multiple features, should it be discarded, counted toward each feature, or somehow partially allocated? Of course, direct fragment counting approaches exhibit desirable features: They are conceptually simple and typically quite fast. Conversely, they suffer from various disadvantages: They have no principled way of handling multimapping reads (e.g., arising from paralogous genes), and they are oblivious to potentially important compositional changes not reflected directly in gene-level read counts (e.g., isoform switching). Additionally, since such methods assess the frequency

of reads overlapping a gene, they must grapple with the definition of a gene. For example, should a gene be the union or the intersection of exons of all transcripts of the gene? Should intronic reads be included? Although the concept of a gene is a useful abstraction, transcripts are assayed in RNA-seq and so present a conceptually cleaner target for quantification.

Transcript-level quantification consists of the assignment of fragments to specific transcripts, which is more challenging but has a number of advantages. It admits a clear interpretation, since transcripts are what the cell expresses; it allows for improved biological resolution and decoding of potentially important biological changes, such as isoform switching; it is the most appropriate level to model and correct for technical biases (68–71); and it provides a proper model for handling reads that multimap, as failing to do so can lead to systematically poor quantification for genes in gene families (72). Solving the transcript-level abundance estimation problem requires a principled solution to aggregating to gene-level estimates (73–75). Conversely, transcript-level quantification is not without disadvantages. Alternative splicing implies that many fragments are ambiguous in their origin, and they must be assigned probabilistically, necessitating the adoption of a model, which may fail to adequately capture reality; this read ambiguity translates to additional uncertainty in the estimated transcript abundances.

Transcript Quantification

Methods for transcript quantification are based primarily on defining a generative model of RNA-seq reads and then trying to perform inference on this model to obtain the relevant quantities (i.e., transcript abundances); see **Figure 5**. There has been a tremendous amount of research on quantifying transcript-level abundance from high-throughput sequencing data; here we describe a few major highlights.

Initial probabilistic frameworks for transcript identification and abundance estimation using EST (expressed sequence tags) data were already being developed before Illumina-based sequencing (76), but Jiang & Wong (77) were among the first to attempt isoform-level abundance estimation using RNA-seq data. They defined counts over exons and exon junctions as arising according to a Poisson model and viewed transcripts as vectors of inclusion and exclusion of these exons and junctions. By expressing the likelihood of the model parameters given the observed data, they posed a statistical model that admits efficient inference, for which they obtained the point estimate by gradient ascent and provided estimates of the posterior distributions of the parameters via importance sampling. This work represents one of the first proper statistical formulations of the problem. However, the approach does not account for fragments that map to multiple genes, and it requires annotations of transcripts in terms of the gene–transcript relationship, as well as the exon and junction read inclusion matrix.

Li et al. (73, 78) proposed one of the most widely adopted generative models for transcript quantification, RSEM. They defined a fragment-level model of RNA-seq experiments in terms of sampling molecules from an underlying population, proportional to the product of their abundance and length, and then generated fragments from the sampled molecules. Primary quantities of interest are estimated, including the nucleotide fractions (the fraction of all sequenced nucleotides deriving from each transcript) and the transcript fractions (the fraction of all transcripts in the initial population comprising each transcript species); these quantities can be directly converted into popular abundance units, such as transcripts per million (TPM) or estimated counts. Notably, they proposed computing the maximum likelihood (ML) estimates using an expectation-maximization (EM) algorithm (see **Figure 5**) and introduced a modified Gibbs sampling procedure to allow estimating credible intervals for the abundance estimates (73). The model is quite general: It works at the fragment level, and it can account for numerous protocol-related aspects,

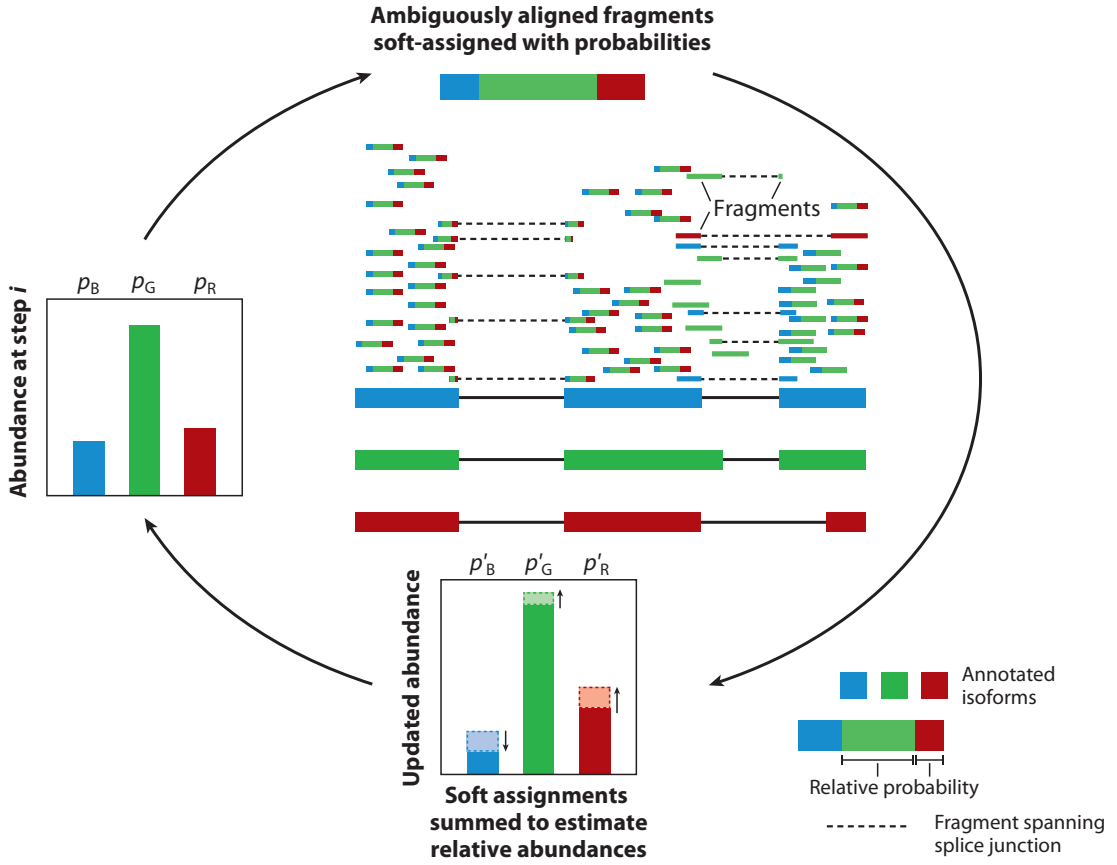


Figure 5

An illustration of the alignment of various reads to a gene with three isoforms: blue (B), green (G), and red (R). In this example, we wish to estimate the abundances of these isoforms, but most reads have ambiguous origins and need to be probabilistically assigned to the transcripts (relative probabilities for each read are shown by the magnitudes of the three colors). Some reads are consistent only with the B and G transcripts, and a few reads uniquely align to a single transcript (single color). In the expectation-maximization (or related) algorithm, given the current abundance estimates, fragments are probabilistically assigned to transcripts, and then estimated abundances are updated by summarizing the (proportional) allocations over all fragments; transcript abundance estimates are determined by iterating the procedure until convergence.

including single-end and paired-end sequencing, directional versus unstranded protocols, various coverage biases, etc. Further, the model relies only on the transcript sequences and not on the relationships to genes or annotations of exons and SJs. Thus, it can be easily applied to both well-characterized and newly assembled transcriptomes. One drawback of a fragment-level model, however, is that each EM iteration scales with the total number of alignments, which is indeed large in most RNA-seq experiments.

Instead of modeling each fragment individually, MMSeq models sufficient statistics (79, 80). Reads are categorized into equivalence classes, where two reads are equivalent if they align to the same set of transcripts. The approach works both within and across genes and does not require the shared regions that give rise to the equivalence classes to correspond to any known annotation (e.g., exon or SJ). MMSeq uses an EM method that works directly over these equivalence classes, allowing efficient inference of transcript-level abundance in this model. In addition to this ML approach, a Gibbs sampling procedure was introduced that can estimate transcript abundances using summary statistics from samples of the estimated posterior, which also allows one to assess uncertainty in the transcript-level abundance estimation and groups of transcripts with correlated

posterior estimates. The underlying likelihood function of the equivalence class–based model is not equivalent to that of the fragment-level model in RSEM, although subsequent work explored other factorizations of the full fragment-level likelihood that either preserved equality with the RSEM model while speeding up inference (81) or sacrificed equality to balance efficiency and fidelity (82, 83). eXpress demonstrated how fragment-level inference could be made much more efficient by modifying the inferential algorithm itself (i.e., online EM), rather than the factorization of the underlying likelihood function (84, 85).

Cufflinks is widely known as both a reference-guided transcript assembly algorithm and a quantification tool (86). Quantification either is restricted to a reference annotation or allows new transcripts to be identified via alignments; transcript abundances are estimated via an EM algorithm to determine the ML estimates given the observed data. While we do not focus on assembly methods here, given the close relationship between transcript identification (assembly) and quantification, numerous approaches attempt to solve both problems together, either stagewise or jointly (82, 87–94).

BitSeq introduced a model similar to RSEM that jointly performs quantification and DE, together with fully Bayesian inference (95). BitSeq focused on sampling from the posterior distributions of transcript abundances, given the fragment alignments, giving accurate estimates (96) and useful information about posterior uncertainty and posterior correlation, which is used in the DE step (95). To combat the heavy computational requirements, Hensman et al. (97) introduced a variational Bayesian (VB) approximation that can be efficiently optimized. TIGAR introduced a VB approach to the transcript abundance estimation problem (98), and the VB EM algorithm was shown to outperform the standard EM algorithm. However, Hensman et al. (97) introduced a novel optimization procedure called VBNG (VB natural gradient), which is a gradient ascent algorithm that considers the information geometry (99) of the underlying problem. Hensman et al. also suggested that EM-based methods tend to find solutions near the boundary of the parameter space, and that their quantifications are less robust than either fully Bayesian or VB estimates (97).

Many of these approaches, among others, simplify the model or improve the efficiency of the inferential procedure, but all rely on full alignments of each read, which can be computationally intensive and time consuming. Recently, several new methods bypass the alignment step and instead adopt lightweight models for quantification. Sailfish defines the transcript abundance likelihood in terms of the constituent *k*-mers of the underlying transcriptome and their abundance in the read data (100). Since the *k*-mers are completely known in advance, the relevant equivalence classes can be precomputed, which reduces the inferential problem to one of simply counting *k*-mers and performing inference via an EM algorithm such as the SQUAREM algorithm (101). This approach increases the speed of abundance estimation by over an order of magnitude compared to full alignment approaches. Building on the idea of *k*-mer-based abundance estimation, RNA-Skim takes the approach of Sailfish even further, identifying sets of distinctive *k*-mers, termed sigmers (102). Transcripts are clustered into groups, and sigmers are identified as *k*-mers that are unique to (and indicative of) each cluster. Quantification is then performed by counting the sigmers in the read data, instead of all *k*-mers, and the EM algorithm is used to estimate transcript abundances from sigmer equivalence class counts. While very fast, these *k*-mer-based approaches do not retain the coherence of the *k*-mers along a read, which can reduce specificity, and they cannot easily estimate certain aspects of the generative model, like the fragment length distribution. Addressing these shortcomings, kallisto relies on the use of pseudoalignments to directly compute the sufficient statistics of the equivalence class–based model of transcript abundance estimation (103). This approach uses *k*-mers to identify the transcripts with which fragments are compatible, but does not treat the *k*-mers independently. The pseudoalignments can be computed in such a way that

equivalence class counts are generated without considering or computing individual fragment-to-transcript alignments, and this can often be achieved by querying only a small number of the *k*-mers present in a fragment, allowing for efficient and accurate estimation in the equivalence class–based model using an EM algorithm. Salmon is another lightweight quantification approach that avoids full alignments, although they can still be used as input (104). It uses a two-phase algorithm for transcript abundance estimation: an online phase using a stochastic collapsed VB inference algorithm (105), where abundances and auxiliary parameters are estimated (e.g., GC bias parameters, sequence-specific bias parameters, fragment length distribution), and an update using mini batches of mappings. Salmon uses a lightweight mapping algorithm to compute the likely transcripts, positions, and orientations of origin of each fragment and adopts a fragment-level GC bias modeling approach (71), which reduces misidentification of expressed isoforms when read coverage is not uniform along the transcripts due to GC content. In the offline phase, a factorized likelihood function is optimized until parameter convergence. The granularity of the likelihood factorization used by Salmon can be adjusted (83) in a way that allows one to trade off between the fragment model of RSEM and the count-based model of MMSeq. In the offline phase, the factorized likelihood is optimized using a VB EM algorithm (98) or a traditional EM algorithm. Combining the efficient determination of fragment–transcript compatibility with RNA-Skim's sigmer concept, Fleximer uses a new matching algorithm that uses sets of sigmers to determine the likely loci of origin of reads, instead of treating each sigmer independently (106). A generalized suffix tree is used to organize the reference sequences, and a segment graph that demonstrates how segments of sequence are shared among reference transcripts is used to select an informative and robust set of sigmers for quantification. Reads are mapped against the reference by matching them to sigmers using a precomputed automaton. This process produces a set of transcript equivalence classes, along with a corresponding count for each sometimes termed the transcript compatibility count; this is used with an EM algorithm to estimate transcript abundances.

Due to their vastly improved speed, ease of use, and reduced computational requirements, alignment-free approaches have become popular for assessing transcript- and gene-level abundance using RNA-seq data. Recent benchmarks (107–110) suggest that, in addition to being fast, such methods can produce accurate abundance estimates—at least to the extent that simulation-based studies, which sometimes adopt the assumed generative models of the quantification approaches, can be relied upon to assess such accuracy. However, there remain opportunities for improving transcript-level quantification methods. For example, the underlying models can likely be further enhanced to account for complexities in the fragmentation patterns of molecules prior to sequencing (111), to better balance robustness and sample-specific accuracy (112), and to address uncharacterized biases. Additionally, most of these approaches (lightweight and otherwise) assume that the annotation of transcripts to be quantified is complete. The accuracy of quantification can suffer when this is not the case, although it is possible to computationally flag transcripts whose estimates are unreliable (113).

BASICS OF DIFFERENTIAL EXPRESSION

Following alignment and quantification, the next challenge often is assessing DE from the estimated feature abundances. We first present a general context and describe the statistical frameworks and overall workflow. The starting point is a count table with rows representing features (e.g., genes) and columns representing samples (i.e., experimental units). The goal of DE is to formulate and test a statistical hypothesis for each feature. Depending on the experimental design, the context, and the research question, more complex analyses are often required. We elaborate on further variations of the overall workflow in the section titled Variants of Differential Expression.

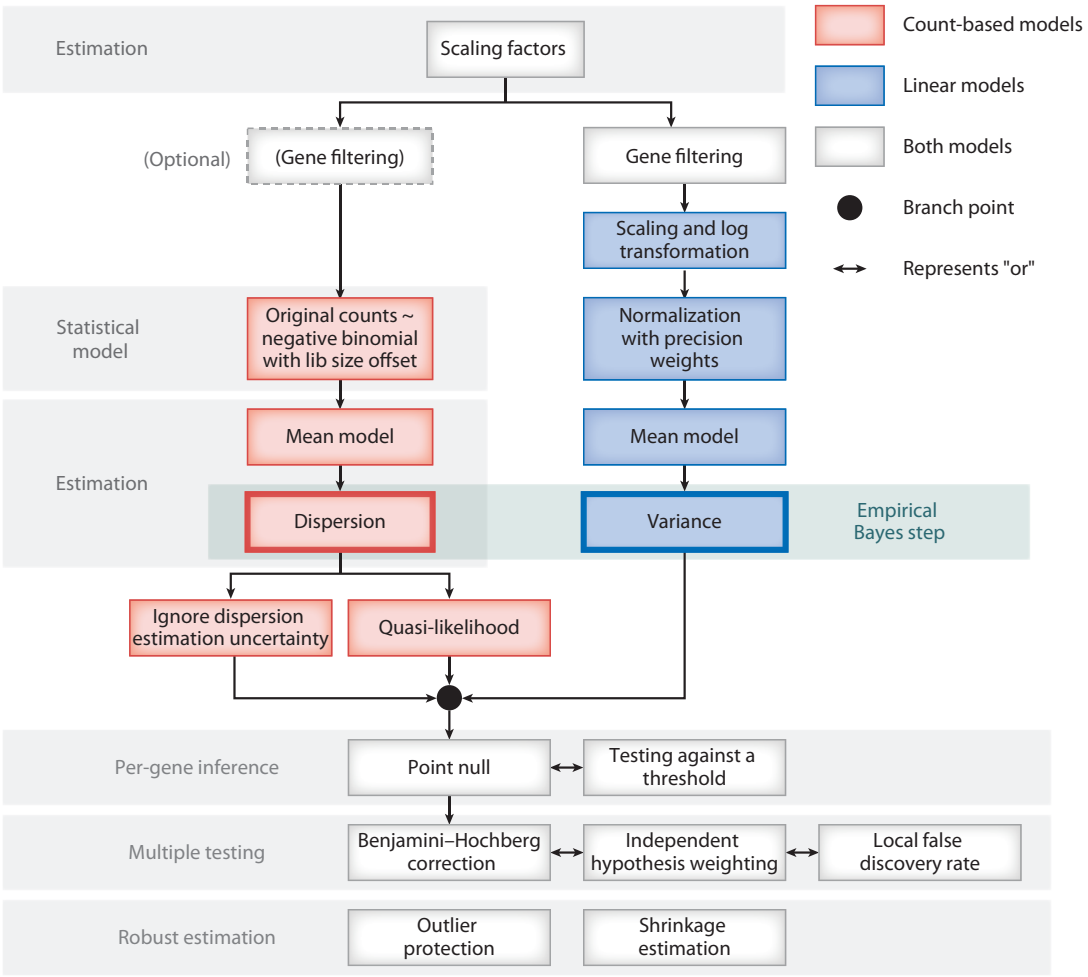


Figure 6

Schematic overview of a DE analysis for RNA sequencing data. Red boxes correspond to pipelines for count-based models (e.g., edgeR, DESeq2), while blue boxes correspond to a linear-model-based pipeline (e.g., limma-voom).

The general workflow involves the following steps (see **Figure 6**): filtering and normalization (preprocessing), specification of the statistical model and estimation of model parameters, statistical inference on the relevant parameters, and adjustment for multiple testing. We introduce this general workflow from the perspective of classical models for count regression. We then discuss various notable deviations, including alternative estimation and inference frameworks and additional strategies to ensure robustness.

Typically, only a limited number of replicates are available (e.g., three to five replicates per condition). The achievable statistical power from such small sample sizes can be low, even for a single feature, with the real interest lying in inference on thousands of features simultaneously. This parallel inference challenge is common to various genome-scale experiments, and the statistical community has contributed strategies to improve the overall performance, from which a few themes have emerged. For example, in estimating parameters for a given feature, one should consider the information coming from the other features in the data set (114). In general, genomics data are ripe for using empirical Bayes methods to moderate estimates, where priors for a feature are derived from a suitable set of other features measured in the data set. In addition,

moderating variance parameters is critical, and indeed, much of the success of earlier parallel inference frameworks (e.g., for microarrays) can be attributed to variance moderation, whether in an ad hoc strategy (115) or in hierarchical models (116). Other tricks such as regularization of regression parameters or considerations for robustness provide additional performance benefits. Taken together, the challenges associated with vast parallel inference can be greatly eased by adopting one or more of these strategies.

Preprocessing: Filtering and Normalization

The vast number of features in a typical RNA-seq experiment leads to a large multiple testing burden. However, many features are largely uninformative; for example, features with low expression provide little evidence for DE. Therefore, filtering strategies are employed that predominantly remove uninformative features and reduce the multiple testing burden. Bourgon et al. (117) showed that filtering is valid if it is independent of the DE test statistic; thus, filtering on residual variance is invalid, while filtering on expression strength, as is commonly done, is valid.

The observed counts of the features cannot be directly compared across samples since there are differences in sequencing depth across libraries. Several methods have been developed to normalize counts to facilitate cross-sample comparisons, although in most count-based models, the counts themselves are not modified and instead scaling factors accompany the analysis. Initial attempts focused on a simple correction for sequencing depth, using the total sum of counts for each sample (i.e., the library size) as a scaling factor (3, 118). However, variation in library preparation or RNA composition between samples also contributes to cross-sample variability and should be accounted for (119). In addition, a few highly expressed genes can largely drive the sampling of fragments, thus leading to inaccurate scaling of the counts. A popular approach is to calculate a size factor (119, 120) for each sample. This can be considered a robust global fold change between the current sample and a (pseudo) reference sample derived from all samples. DESeq's median of ratios method and edgeR's trimmed mean of *M*-values (TMM) method (where *M*-values denote empirical fold changes between two samples) are the most popular scaling approaches (121). Both procedures assume that most genes are not DE and adopt robust summarization methods to calculate the size factors (effective library sizes) to reduce the impact of DE genes (TMM uses a trimmed weighted mean; DESeq uses the median of the log-expression ratios). More advanced normalization methods have since emerged to address other technical artifacts such as GC content and transcript length effects and to accommodate within- and between-lane normalization, e.g., CQN (122) and EDASEQ (123). Moreover, methods based on external spike-in features have been introduced to address normalization for applications where many features are DE or where the basic assumptions of conventional normalization methods are violated (124–126). Recently, a normalization technique has been proposed for RNA-seq data with large differences between conditions that assumes similar distributions in biological replicates, while accommodating for differences between conditions (127).

The normalization size factors are built into the DE analysis workflow as offsets in the statistical models (see below). Notably, size factors are treated as fixed and known, although they are actually random variables estimated from the data (128), and it is unclear how ignoring their associated uncertainty affects the downstream DE analysis.

Modeling and Estimation

Because of the typically small sample size, DE tools mainly implement parametric methods (120, 129–132). Initially, count data were log-transformed and linear models were used for DE analysis

(4). However, log-transformed counts suffer from heteroscedasticity (a systematic mean-variance trend) intrinsic to count data, rendering the standard linear model, which assumes homoscedasticity, suboptimal. In addition, fitting continuous models to (transformed) count data introduces a further approximation. Therefore, discrete count distributions have gained more traction in the initial frameworks.

Gene expression variability across technical replicates (i.e., resequencing the same sample), so-called shot noise, has been shown to approximately follow a Poisson distribution (118), for which the variance is equal to the mean. Biological replication introduces additional cross-sample variability, and analysis frameworks therefore have resorted to one of the natural extensions, the gamma-Poisson or the negative binomial (NB) distribution, which has an additional dispersion parameter and a quadratic mean-variance relationship,

$$Y_{fi} \sim \mathrm{NB}(\mu_{fi},\ \varphi_f),$$
$$\mathrm{Var}(Y_{fi}) = \mu_{fi} + \varphi_f \mu_{fi}^2,$$

where Y_{fi} denotes the read count of feature f in sample i, φ_f is the dispersion for feature f, and $\mu_{fi} = s_i \theta_{fi}$ represents the average expression, which is driven by the true (relative) mRNA concentration in the sample, θ_{fi}, multiplied by a normalization scaling factor, s_i; there also exists a characteristic dispersion-mean trend in RNA-seq data sets (**Figure 7*a***). Initial implementations focused on two-group comparisons (120, 133) and were later extended to the generalized linear model (GLM) framework, an extension of classical linear models to non-Gaussian responses (134). GLMs allow for the inclusion of multiple treatments or covariates, thus broadening the

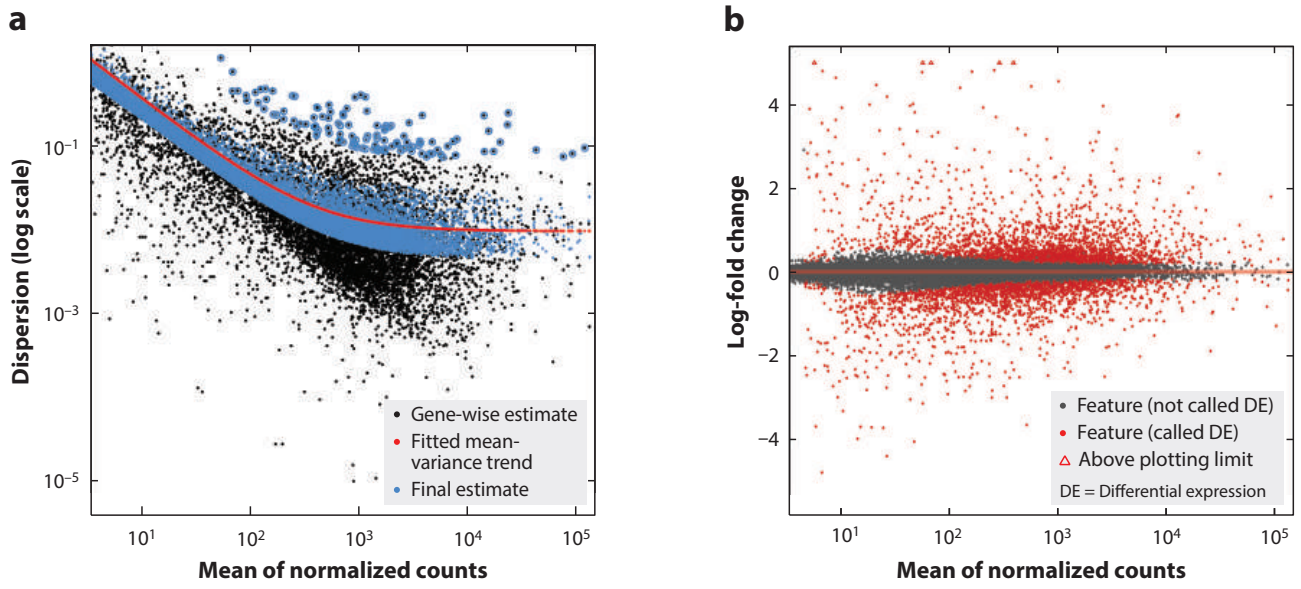


Figure 7

(*a*) A dispersion-mean plot of the RNA sequencing (RNA-seq) experiment from Reference 135, as processed in Reference 136. The dispersion trend smoothly decreases for genes with higher expression and eventually reaches an asymptote, which can be considered the biological variability present in the data set for a typical gene. (*b*) MA (log ratio over mean) plot of the same RNA-seq experiment. The *y*-axis shows the moderated log-fold change and the *x*-axis shows the mean of normalized counts. Red points denote differential expression detection according to a nominal false discovery rate threshold of 0.05.

applicability. The NB GLM model can be formulated as

$$Y_{fi} \sim \mathrm{NB}(\mu_{fi},\ \varphi_f),$$

$$\log \mu_{fi} = \eta_{fi},$$

$$\eta_{fi} = X_i \beta_f + \log s_i,$$

where η_{fi} is the linear predictor, X_i denotes the design matrix, β_f represents the regression parameters, and $\log s_i$ are scaling (normalization) offsets. Regardless of the model, the parameters θ_{fi} or, equivalently, (a linear contrast of) β_f would represent the parameter(s) of interest for inference.

Reliable estimation of the dispersion parameter φ_f is nontrivial due to limited sample sizes. Traditional ML estimators for the dispersion are negatively biased (137) since they do not account for the fact that the mean is also estimated from the data. Early implementations estimated a single common dispersion parameter for all features (137), with the rationale to obtain a stable estimate by borrowing strength over all genes. However, the common dispersion assumption is unrealistic and relaxed estimation schemes were proposed, such as moderation toward a common dispersion (133) or estimation in strata of similar expression strength (120). For example, DESeq adopts a method of moments (MM) estimator and assumes the dispersion to be a smooth function of the mean. To avoid too liberal inference, one then sets the dispersion as the maximum between the smooth fit and the gene-wise MM estimate; however, while robust to outliers, this method tends to overestimate the variance and is therefore conservative (138, 139). Later approaches resorted to an approximate conditional inference scheme, the Cox–Reid adjusted profile likelihood (APL) (140), to correct for the bias in the ML estimator (134). Again, stable estimation is provided by leveraging information across genes (**Figure 8**). In particular, edgeR uses a maximized weighted APL to trade off between gene-specific and shared dispersion estimators upon estimating the dispersion-mean trend across all genes (similar to DESeq). The weighted likelihood,

$$\mathrm{APL}_f(\varphi_f) + G_0 \mathrm{APL}_{sf}(\varphi_f),$$

consists of the APL for a specific feature f (first component) and a shared likelihood (second component), which can be interpreted as a prior from a Bayesian perspective, thus representing an approximate empirical Bayes solution (133). The weight given to the prior likelihood, G_0, can also be estimated from the data (141). Analogously, DSS (Dispersion Shrinkage for Sequencing) and DESeq2 model the log φ_f as a Gaussian random variable, and Bayes' formula is applied to generate a posterior mode for each gene (129, 142). Hyperparameters for the (Gaussian) prior are inferred from the data using either the MM or the Cox–Reid estimator across all genes. Once dispersion estimates are available, the parameters of the mean model, β_f, can be estimated using standard algorithms for GLMs.

Statistical Inference

After fitting a GLM to each feature, the statistical inference typically involves testing the null hypothesis H_0 that there is no DE between conditions, i.e., that the LFC is zero, against the alternative H_1 that the LFC differs from zero. In the GLM framework, the null hypothesis can be represented as either a single regression parameter or a linear combination of parameters (contrasts), which is defined by a vector or matrix L such that H_0 is the hypothesis that $L\beta_f = 0$. Indeed,

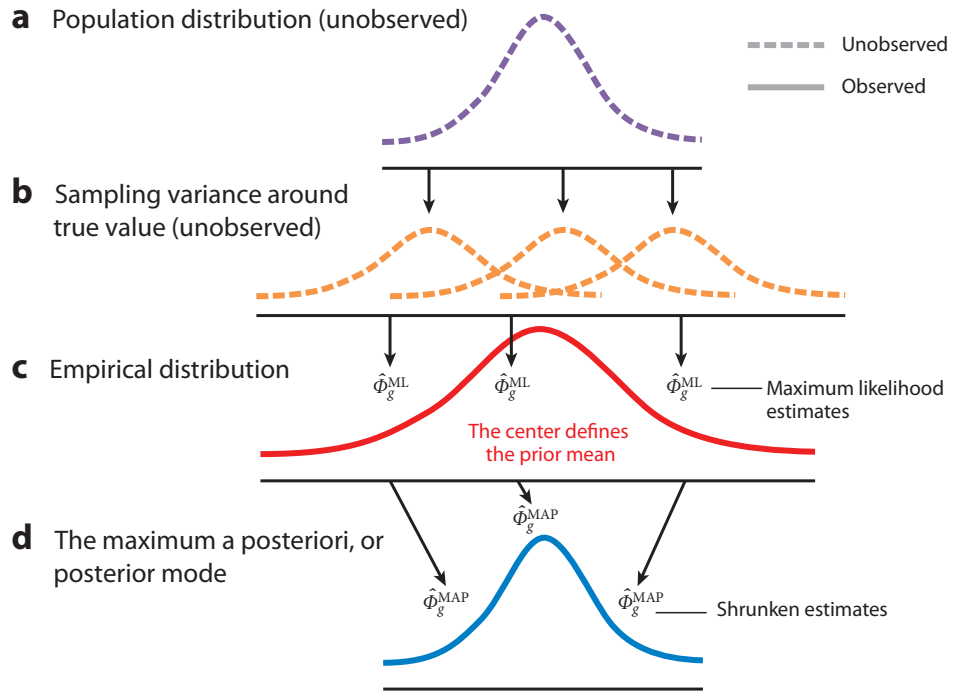


Figure 8

Steps in an empirical Bayes model. In an RNA sequencing experiment, one assesses the observed differences in gene expression across groups of samples with respect to within-group variance. (*a*) The unobserved population distribution for the true within-group variance of each gene. (*b*) Variances are estimated from limited sample size experiments, and so there is sampling variance in our estimate of the variance. A maximum likelihood estimate (MLE) or a bias-corrected estimator for expression variance can be used. (*c*) Thousands of genes are typically observed and estimates are made for each, providing an empirical distribution of MLEs across all genes. This empirical distribution of MLEs can be used to determine a prior distribution for empirical Bayes analysis; the posterior distribution for the variance of each gene is calculated using Bayes' formula. (*d*) Distribution of the maximum a posteriori (MAP), or posterior mode, estimates of variance over all genes. The posterior modes represent shrunken estimates, where the amount of shrinkage is determined by the shape of the likelihood and the width of the prior distribution.

a regression parameter in an NB GLM with a canonical link function can be interpreted as an LFC between groups and thus provides a measure of effect size.

There are multiple hypothesis tests available for GLMs with known (asymptotic) distribution under the null hypothesis. Likelihood ratio tests (LRTs) compare the likelihood of a full model, upon estimating all parameters without constraints, with the likelihood of a reduced model, where one or some of the parameters are constrained according to H_0. LRT statistics are asymptotically χ^2-distributed under H_0, and this type of test is implemented in both edgeR and DESeq2. By default, however, DESeq2 adopts a Wald test. Wald tests are attractive from a computational point of view since they only require fitting the full model and calculating the variance–covariance matrix of the regression coefficients. The Wald test statistic for a single model parameter or a single contrast, $W = \widehat{LFC}\ /\text{se}(LFC)$, asymptotically follows a standard normal distribution under H_0, where se(LFC) is the standard error of LFC and $\widehat{LFC}$ is the ML estimate of LFC. From ML theory, it is known that LRTs have better properties (e.g., invariance to transformation) than Wald tests in GLMs (143); however, RNA-seq tools moderate dispersion estimates and do not re-estimate them under H_0, so it is unclear whether these benefits carry over to RNA-seq data analysis in practice.

Multiple Testing

The p-values obtained from the statistical inference must be corrected for multiple testing to avoid excess false positives. While it is possible to control the probability of returning at least one false positive in the list of detections by adopting familywise error rate corrections, this stringent form of correction is overly conservative. Indeed, when screening many thousands of features, one is typically willing to tolerate a certain proportion of false positives to obtain a larger number of true positives. The false discovery rate (FDR), which gained significant popularity, controls the expected fraction of false positives in the detected set of features, i.e., $FDR = E[V/\max(R, 1)]$, where V is the number of false positive rejections and R is the total number of detections. The FDR was introduced by Benjamini & Hochberg (144) and has become common practice in high-dimensional data analyses because of its simplicity and solid theoretical justification. Indeed, it can be shown that the FDR is justified under a range of dependency structures between genes (145) and can be approached from both frequentist and Bayesian perspectives.

Variations to the General Workflow

There is a large and growing number of alternatives to the basic framework mentioned above: different inferences based on the same models, alternative models, more robust approaches, different testing regimes, variations on multiple testing corrections, and so on. In this section, we summarize some of these developments.

Alternative models (inference frameworks). NB count models, which underpin many DE tools, assume a quadratic mean-variance relationship. Inference, however, may benefit from a more flexible variance structure, and for this, other models have been proposed. One strategy uses quasi-likelihood (QL), which requires only that mean and variance are specified to be able to make inference on the mean model parameters (146). The QL method adopts the same mean model structure as the NB but introduces an additional overdispersion parameter such that $\mathrm{Var}(Y_{fi}) = \psi_f(\mu_{fi} + \varphi_f \mu_{fi}^2)$, where ψ_f is estimated using a moderated MM estimator. QL naturally allows (asymptotic) hypothesis tests based on t- and F-statistics, thus accommodating the uncertainty in the estimation of the additional QL dispersion parameter. Another variation is the use of a more flexible distribution, such as the NB power distribution, which adds an additional parameter (147) to the NB. Within the NB framework itself, Bayesian methods have also been developed. A fully Bayesian approach has the benefit that various aspects of the posterior can be reported (e.g., credible intervals), and the degree of parameter shrinkage naturally depends on the amount of information available for that gene (a trade-off between expression magnitude, dispersion, and residual degrees of freedom). One of the early methods was ShrinkBayes, a fully Bayesian approach that included multiple mixture priors (e.g., Gaussian) (148, 149) and where fitting was accomplished using integrated nested Laplace approximations (150), which avoids the Markov chain Monte Carlo sampling. Another alternative is to remain within computationally and inferentially efficient Gaussian linear models, after suitably transforming the (normalized) count data. For example, limma-voom models log-transform normalized counts using a linear model while adjusting for heteroskedasticity via weighted regression, where the observation weights are computed from the observed mean-variance relationship (151). In this case, moderated t- and F-statistics are used for inference. Finally, nonparametric methods have been developed, which are more robust to outliers and do not require distributional assumptions. For example, SAMSeq (152) adopts the Wilcoxon test to assess DE between groups and uses resampling procedures to adjust for differences in sequencing depth.

Robust log-fold change estimation. The standard NB workflow typically makes use of APL NB likelihood for parameter estimation, combined with empirical Bayes procedures to borrow strength across features when estimating the dispersion parameter. There are two related challenges: (*a*) Ratios of smaller counts result in more variable LFCs (**Figure 7*b***), and (*b*) the estimation of LFC can be sensitive to outliers. This makes it difficult to rank genes according to LFC since lowly expressed or outlier-affected genes are likely to dominate the top list. To derive more robust LFC estimates, researchers have adopted several approaches. First, prior counts have been used in the numerator and denominator of the LFC; effective shrinkage is accomplished by augmenting each count with a carefully chosen value, although the optimal value may vary across data sets. Second, edgeR-robust (139), for instance, adopts an M-estimation approach by iteratively downweighting outlying observations within the GLM fitting procedure, dampening the effect of outliers on both mean and variance estimates. Alternatively, outliers can be identified and removed and/or imputed by taking advantage of the remaining data for a feature (129). Lastly, priors can be imposed on the LFC parameters. For example, DESeq2 includes a zero-centered Gaussian prior in the NB GLM and provides the posterior mode of LFC as output (129). The width of the prior is set conservatively, using a weighted upper quantile of the observed LFCs. New alternative shrinkage estimators in DESeq2 incorporate priors with heavier tails that introduce less bias, using either a mixture of normal distributions (153) or a Cauchy distribution (154).

Accounting for unobserved effects. As mentioned above, (G)LMs can adjust for known confounders. However, genomic data can also be affected by unknown, and hence unobserved, confounders. This problem is widespread in publicly available data, which typically do not contain sufficient metadata on potential batch effects caused by lab, protocol, date, etc. Batch correction methods can leverage the parallel structure of high-throughput transcriptomic data to identify unknown and unobserved systematic effects. SVA (surrogate variable analysis) (155, 156) and RUV (remove unwanted variation) (125) methods, for instance, estimate surrogate variables through singular value decomposition on control features or on a matrix of model residuals so that the phenotypic effect of interest is not captured by the surrogates. RUV also has the option to exploit information in replicate samples. The estimated surrogate variables can subsequently be included as predictors in the statistical model to adjust for the batch effects.

Statistical inference by testing against a threshold. The standard approach for detecting DE in RNA-seq involves a simple null hypothesis H_0 that the LFC is zero. However, statistical significance does not guarantee that the fold changes are large enough to be biologically relevant. Analysts often produce candidate gene lists by applying a threshold on the magnitude of the LFC, but the statistical properties of this approach are unclear. The FDR is a set property and has no interpretation when the set, post-FDR calculation is altered (157). To address these practical and theoretical concerns, researchers have adopted several tests relative to an LFC threshold, a procedure initially proposed for microarray data (158). This results in a composite null hypothesis H_0, such as $|LFC| < a$. Implementations differ: DESeq2 replaces the composite null with a simple null hypothesis at the boundary of the parameter space (129), whereas edgeR uses a modified likelihood ratio test or a quasi-likelihood F-test against a threshold (159).

Small-sample inference. The null distributions for Wald or LRT statistics for count models are only valid asymptotically, and the number of replicates is often too low for these approximations to be fully effective, which may lead to an inflated FDR. Initial implementations provided exact tests (137), but these can only be applied in simple designs. Another strategy is small-sample

asymptotics, which makes use of higher-order approximations that are still compatible with the GLM framework (160).

Multiple testing. While the FDR achieves a more reasonable sensitivity–specificity trade-off than familywise error rate correction approaches, other developments beyond simple filtering aim to further reduce the multiple testing burden. Storey's q-value, for instance, estimates the proportion of true null hypotheses from the data to increase power (161), while others adopt a data-driven weighting of the p-values in the FDR correction (162). Although the FDR is deeply rooted in statistical theory, it is not guaranteed that methods will control error rates at the nominal level in real applications. NB methods, for instance, rely on the asymptotic theory, which might not hold for applications with low sample sizes. A study has suggested that coregulation of genes induces intergene correlations, which can alter the null distribution of the statistical test (163); local FDR approaches were introduced that empirically estimate the null distribution (164). Other developments address issues in testing many hypotheses for every gene (e.g., multifactorial designs). The conventional approach is to control the FDR on each hypothesis, but this does not allow for straightforward prioritization since genes typically have a different ranking for each hypothesis. Stagewise testing procedures can be interpreted as generalizations of analyses of variance with post hoc tests for high-throughput contexts (165, 166), thus allowing a natural ordering of the genes according to an omnibus test (all effects of interest) while providing FDR control at the gene level.

VARIANTS OF DIFFERENTIAL EXPRESSION

The previous section introduced count-based DE in general terms: Each row of a count matrix is submitted to a statistical model (often by first estimating moderated variance parameters over the whole data set) and hypothesis tests of interest are conducted, with an adjustment for multiple testing. In this section, we discuss additional approaches to interrogate RNA-seq data in terms of DE.

Although DE is of obvious interest, this can manifest or be defined in multiple ways (see **Figure 9**). One may want to cast inferences to the gene level, but measurements are made at the fragment level. We use the term "differential gene expression" (DGE) to refer to hypothesis testing related to the total outcome of an annotated gene, by comparing either accumulated TPM estimates or raw counts while including an adjustment for average transcript length via offsets (74). If the expression of transcripts is the feature of interest (independent of other transcripts), differential transcript expression (DTE) analyses can be conducted. Alternatively, one could be interested in whether at least one transcript from a gene is DE. This requires statistical testing at the transcript level and then aggregation to the gene level. Yet another strategy is to consider whether the relative abundance (i.e., proportions) of transcripts for a specific genomic locus changes between conditions, which is commonly termed differential transcript usage (DTU) or, more generally, differential splicing (DS). A surrogate for DTU, differential exon usage, is conducted on exon-level quantifications; in this case, the goal is to identify exons that deviate from proportional expression to separate differential usage from DE. Yet another alternative is to quantify and test differences at the event level, where reads supporting (or not supporting) an event (e.g., inclusion of a cassette exon) are summarized and compared (167).

There is certainly a question of which analysis path to choose. Conceptually, pure DTE points to all kinds of DE, and while casting a wide net of potentially interesting genes might seem appealing, there are some considerations to be made. For example, if a given transcript is DE, often the question becomes, What happens to the expression of the other transcripts for this gene? Are all transcripts changing in the same direction? If so, it may be better in terms of sensitivity to detect

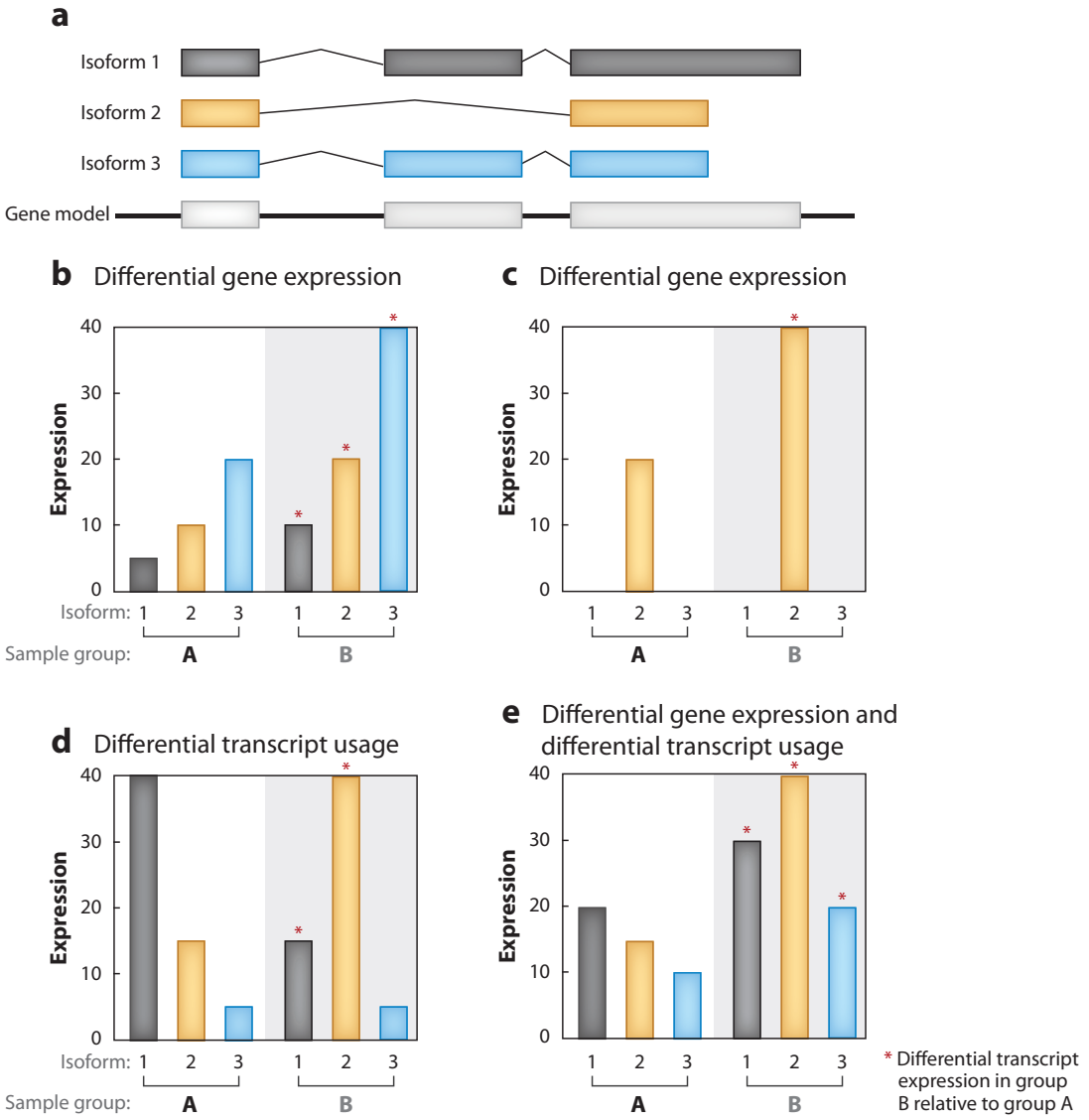


Figure 9

Schematic illustration of some examples of differential gene expression, differential transcript expression, and differential transcript usage for a gene with three isoforms (1, 2, and 3) in a two-group comparison (A versus B).

an aggregated output (i.e., DGE). Or, transcript-level expression can be represented as a genewise multivariate outcome and isoform switches can be considered collectively, i.e., by assessing DTU, which is not affected in either direction by DGE. DTU implies DTE while the opposite is not necessarily true. We generally favor two clear but orthogonal analyses (DGE and DTU) over a catchall DTE analysis (74), but this will ultimately be application dependent, and scientists should clearly define their question of interest in advance.

Differential Transcript Expression

Modeling transcript-level count data for DE presents some additional challenges due to increased variability and resolution compared to gene-level analyses. For example, transcript-level

abundance estimates are considerably more variable than gene-level counts due to ambiguous assignment of fragments to isoforms (74). Thus, transcript quantifications inferred by popular tools such as RSEM, Salmon, or kallisto carry more uncertainty, which should be accounted for in the downstream DE analysis.

Transcript quantifications still have many of the properties of count data (e.g., mean-variance relationships) and thus could be used as inputs to the frameworks mentioned above. However, quantifications are estimates that may obscure inference when plugging them into count-based RNA-seq tools. Cufflinks was one of the first methods to use estimated abundances and their corresponding standard errors to perform DTE (and DGE) analyses; the method quantifies transcript abundances via a likelihood model and an EM algorithm, and tests of DE are performed by applying the delta method on the abundance parameters (75, 86). Bayesian approaches for identifying DTE based on estimated counts, e.g., ranking via Bayes factors, include EBSeq (132), which uses an empirical Bayesian hierarchical model, and MMSeq (79, 168), which fits a linear mixed model to data via Markov chain Monte Carlo techniques. Similarly, BitSeq (and later cjBitSeq) introduced a generative model that couples both quantification and DE using fully Bayesian inference (95, 169). Most recently, with the advent of ultrafast transcript quantification algorithms, sleuth uses bootstrap samples of each sample of reads to determine the so-called inferential variance and integrates this into the DE calculation through a variance components model on the log-transformed scale (170).

Differential Transcript Usage (Differential Splicing)

One of the first statistical models for DTU, cuffdiff, calculates the square root of the Jensen–Shannon divergence on estimated transcript proportions and uses the delta method to estimate the variance of this metric under the null hypothesis of no change in proportions (86). Another conceptually distinct approach formulates a Poisson mixed effects model on exon- and junction-level quantifications and searches for exon–condition interactions that represent differential usage (171). Such departure-from-parallelism modeling was introduced in earlier analyses of probe-level microarray data for DTU (172); on RNA-seq data, this approach was further formalized with DEXSeq (173), which uses an NB model on exon-level counts. Exon by exon, DEXSeq tests whether an improvement in fit is achieved by adding a single exon–condition interaction, which represents the differential usage of that exon across conditions. A comparison study showed that DEXSeq has a good performance in well-annotated transcriptomes and that filtering of lowly expressed transcripts improves error control (174); in addition, DEXSeq also works well with transcript quantifications as input (175).

In a similar vein, DRIMSeq (176) and LeafCutter (177) employ the Dirichlet-multinomial (DM) distribution to perform the same inference task but treat the output of a gene's expression as a multivariate response; Bayesian inference for the DM model has also been considered in BayesDRIMSeq (178). Several tools neglect the uncertainty in estimated transcript-level counts, and this is perhaps the reason for inflated FDRs (175). To address this, RATs (relative abundance of transcripts) uses bootstrapped (transcript-level) quantifications to infer DTU via a *G*-test of independence, based on the multinomial distribution, on the two groups' isoform counts (179). Instead of considering estimated counts and their uncertainties, Bayesian methods such as cjBitSeq (169) focus on the group of transcripts that each read is compatible with (i.e., equivalence classes). In this way, quantification is not required because the DS tools treat the transcript allocation of reads as an unknown latent variable.

Event-Level Analyses Based on Percent Spliced-In

Some methods perform differential analyses based on percent spliced-in values (PSIs). PSIs can be computed either for specific events (retained intron, cassette exon, etc.) or at the transcript level and indicate the fraction of RNA-seq reads supporting the event, obtained as the ratio between the number of reads including the event and the total number of reads including and excluding the event. The difference of the PSIs between conditions is then used to assess DS, performed separately for each event (or transcript). Some of the main DS tools based on PSIs include rMATS (180), which uses an LRT, and SUPPA2 (181), whose test is based on comparing the observed difference in PSIs across conditions to the empirical cumulative density function of the within-replicates differences of PSIs of SJs from similarly expressed transcripts.

Event-level analysis, similar to DEXSeq's exon-level approach, separately focuses on each splicing event, and results could be aggregated to the gene level by considering the most significant event- or transcript-level test, appropriately adjusted for multiple testing (173, 181, 182).

Multistage Testing

As mentioned, DS analyses can be approached at the gene-, transcript- or event-specific level. While gene-level tests often have higher sensitivity, testing each individual transcript provides increased resolution. However, neither gene- nor transcript-level tests guarantee FDR control on the full set. Stagewise testing procedures (165, 166) instead first screen for significant genes and only consider significant transcripts from those genes. This procedure gives gene-level FDR control and allows researchers to leverage the power from gene-level tests while interpreting results at the transcript level (175). The same procedure can be applied by replacing transcript-level tests with exon- or event-specific tests.

SINGLE-CELL TRANSCRIPTOME SEQUENCING

One of the emerging data types in transcriptomics is scRNA-seq, whereby the expressed content of individual cells is prepared and sequenced. In this case, experimental design is again of critical importance to avoid confounding the data (183). Experimentally, capture and reverse transcription efficiency become important, given that the number of mRNA transcripts per mammalian cell is estimated to be between 50,000 and 300,000 (184).

Two main experimental approaches are used: plate-based, where cells are sorted into individual wells for lysis and library preparation, and droplet-based, where each cell is absorbed (together with reagents) and processed within an oil droplet (185). Several variations of these protocols are now available, increasing the number of cells assayed, but ultimately only a small fraction of the expressed RNAs (cDNAs), often the most highly expressed transcripts, are captured. The features that distinguish scRNA-seq from bulk RNA-seq data include (*a*) generally low depth of sampling for each cell (due to cost, but also due to lower diversity of cDNA fragments); (*b*) so-called dropout, where a cell expresses a transcript but it is unobserved; and (*c*) higher levels of biological (since no averaging) and technical (e.g., more amplification) variation.

Nonetheless, researchers can distinguish cell identities, where identity represents the combined effects of cell type (permanent features) and cell state (transient features) (186). The Human Cell Atlas, among other projects, opens the door for exploring spatial context (187), developmental patterns (188), immune responses (189), response to therapy (190), and an increasing range of basic science and clinical investigations (191–193).

Although many computational aspects of scRNA-seq data are beyond the scope of this review (e.g., dimensionality reduction techniques, ordering cells into lineages), one connected application

area that has already received considerable attention is DE analysis. In the simplest setting, cells are first partitioned into different classes (e.g., assumed to correspond to different cell types) via clustering, with the subsequent aim of finding markers for each cluster (e.g., to annotate cell types). To perform this task, a statistical model uses cells as experimental units, as opposed to samples in bulk analyses; thus, it is worth considering the population to which the conclusions extrapolate.

To date, several methods have been developed to decipher DE between cell types, many of which have been comparatively assessed in recent benchmarks (194, 195). Many of these single cell–specific methods are extensions or variations of existing bulk approaches. For example, SCDE formulated the RPM (reads per million) data for a given gene across cells as a mixture of Poisson and NB components; using a Bayesian approach, probabilities of observing a given fold change are converted into empirical *p*-values (196). MAST uses a hurdle model on $\log(TPM + 1)$ data, where a logistic regression is used to model whether a gene is expressed, and a Gaussian linear model is used conditional on expression. Inferences for the two sets of regression parameters are done in a Bayesian framework that also provides regularization (197). Again, extending existing approaches, Van den Berge et al. (198) proposed a zero-inflated NB (ZINB) model; model fitting is done within the ZINB-WaVE (wanted variation extraction) framework (199), estimating cell- and gene-specific posterior probabilities for counts to belong to the NB count component of the ZINB mixture model. These probabilities are used as observation weights in the downstream estimation of regression parameters in the classical NB framework.

Nonetheless, many DE methods focus on assessing changes in the mean parameters. However, since cell subsets are being compared, we may not have simple shifts in the mean. Instead, it may be informative to detect and understand changes in expression variability across conditions (200). Alternatively, full distributions (instead of means or variances) can be compared, as was proposed in a Bayesian framework in scDD (201), highlighting not only DE but also differential proportions (change in the relative usage of low and high expression), differential modality (change in the number and place of the mode of expression), or some combination thereof.

In many applications of single-cell DE analysis, the sample sizes (numbers of cells) are generally larger than those commonly used within the optimized frameworks built for bulk RNA-seq data, and thus it seems that the distributional assumptions play less of a role for effective inference. Indeed, a recent comparison highlighted decent performance of *t*-tests and Wilcoxon rank sum nonparametric tests in comparing single-cell subsets (194).

Beyond comparing cell types, which may involve multiple experimental units (e.g., patients), it will be of increasing interest to compare expression levels of genes across biological replicates and conditions. For example, it may be of interest to understand cell-type-specific immune responses following a stimulus. A recent study compared multiple patients across stimulated and unstimulated conditions by first computationally separating immune cell types (189); to do this, the researchers aggregated cells from a given cell type into a pseudobulk RNA-seq data set and performed DE using standard tools.

LONG-READ TRANSCRIPTOME SEQUENCING

The short read length of Illumina-based RNA-seq complicates the unambiguous placement of reads to the genome, especially in repeat regions (202), and adds difficulties to the assembly, identification, and quantification of expressed isoforms (203–205). In contrast, so-called third-generation, or long-read, sequencing technologies, led by PacBio (206) and ONT, can generate much longer reads. By sequencing single molecules, they can also forego PCR amplification, hence reducing coverage biases (207, 208). Currently, long-read technologies incur a higher average cost

and a higher error rate than short-read sequencing (203, 209). However, this is a rapidly developing field, and improvements in error rates and throughput are to be expected.

The strategies employed by PacBio and ONT to generate long sequencing reads of single molecules differ in many ways. PacBio, with its RSII and Sequel instruments, uses SMRT (single molecule real time) sequencing (210), where the reactions take place inside so-called zero-mode waveguides (ZMWs) (211). At the bottom of each ZMW, there is a single DNA polymerase molecule. As the polymerase processes a DNA fragment, the incorporation of each nucleotide leads to a fluorescent signal, which is detected by the ZMW and converted to a base call. A specific characteristic of the PacBio library preparation system is the creation of SMRTbell templates (212), which are obtained by ligating SMRTbell hairpin adapters. The result is a circular construct, where the two strands of the template are separated by adapters with known sequences. As the construct is processed by the polymerase in the ZMW, the original template can be passed multiple times. Since the sequencing errors are largely random (213), the base-level error rate can be considerably reduced by forming a consensus over these passes.

ONT, in contrast, uses a different sequencing strategy based on protein nanopores placed in a polymer membrane (214) for its MinION and PromethION sequencers. A current is passed through the nanopores, and as the template molecule is passed through the pore by a motor protein, each combination of bases induces a change in the current. Analyzing the exact nature of this change allows for the identification of the template sequence. By adding a hairpin sequence to the end of the double-stranded cDNA fragment before denaturing it into a single-stranded molecule and passing it through the nanopore, both the template sequence and its complement are included in a single read and can be combined at the base calling step to generate a higher-quality, so-called 2D, read (215). In contrast to PacBio, ONT also offers direct sequencing of RNA (216). Advantages of this include that the reverse transcription step is avoided, which may reduce biases, and that RNA modifications can be directly observed, since they also change the current passing through the nanopore in characteristic ways (217). However, at present, the required amount of starting material is considerably higher than for cDNA protocols.

Applications to cDNA (RNA) include both transcriptome-wide sequencing and characterization of specific genes via targeted sequencing (14, 203, 218–222), as well as performance evaluations based on synthetic transcript catalogs [ERCC (External RNA Controls Consortium) with 92 sequences or SIRV (spike-in RNA variant) with 68]. With LRTS, every read potentially represents a full-length transcript. If this were indeed true, de novo (reference-free) identification of the full set of observed isoforms would be straightforward and would only require grouping reads expected to differ only by sequencing errors (which, depending on the error rate and isoform similarity, may not be trivial). However, this is not currently the case, due to fragmentation and degradation of template molecules during library preparation and early termination of the sequencing, which leads to ambiguities in transcript identification (223). This means that it is not easy to determine whether truncated variants are present.

Transcript identification from LRTS can be either reference-based or reference-free. The latter typically involves clustering reads based on similarity, followed by polishing the consensus sequence within each cluster (14, 224–227). Since LRTS is still a young field, methods and tools for reference-based alignment are still emerging but so far include a mix of established tools, such as GMAP (58), and new innovations, such as minimap2 (228). A recent study comparing PacBio, ONT, and Illumina data (203) showed that the long-read technologies were indeed much better at correctly identifying expressed SIRV transcripts than de novo assemblies of short reads.

The rapid technological developments in LRTS also mean that the read generation process (e.g., biases affecting the observation of a given read) is still largely unknown. In addition, read lengths are extremely variable, error rates are relatively high, and throughput is still relatively low.

For the PacBio RSII instrument, the selection of transcript molecules is biased toward short sequences (223). Thus, samples are typically size-fractionated before sequencing, which distorts the abundance estimates. Taken together, these and other aspects make accurate transcript quantification from LRTS more difficult, and new models and tools will be needed. Encouragingly, a recent study showed that by combining LRTS and Illumina data, more accurate quantifications for the artificial SIRV transcripts could be achieved (203).

Since abundance estimation for long reads returns values in the form of read (or transcript) counts, it is plausible that the DE machinery developed for short-read data can be applied in a similar way. The quality of the DE calls will be directly dependent on the accuracy of the abundance estimates. However, the current low depth of sequencing compared to short-read data sets will ultimately lead to low power to detect DE features.

SUMMARY

> I'm a scientist and I know what constitutes proof. But the reason I call myself by my childhood name is to remind myself that a scientist must also be absolutely like a child. If [they] see a thing, [they] must say that [they] see it, whether it was what [they] thought [they] were going to see or not. See first, think later, then test. But always see first. Otherwise you will only see what you were expecting. Most scientists forget that.
>
> –adapted from *The Ultimate Hitchhiker's Guide to the Galaxy* by Douglas Adams

In this review, we gave an overview of the data science of gene expression analysis, with a focus on methods to estimate transcript-level abundance and statistical tools for assessing DE. Notably, RNA-seq data are often an intermediate discovery step where the detected molecular changes represent candidates for further follow-up. Nonetheless, the analysis of RNA-seq data for gene expression is already very mature, due to a deep understanding of the biases present, to the implementation of efficient data structures and algorithms for processing the data into (estimated) count tables, and to a refined understanding of how well tools perform via the many benchmarks available.

Ultimately, the success of RNA-seq lies in its wide range of applications, and it is likely that Illumina-based short-fragment RNA-seq will continue to be the workhorse for the field for many years. With the increasing fidelity of single-cell protocols, many tools are emerging to deal with the additional complexities of single-cell measurements, and these will be further refined in the coming years. Similarly, with the decreasing costs and lower error rates of long-read technologies, it may be possible to characterize alternative transcription quantitatively with full-length transcript sequencing, thus considerably reducing read-to-transcript ambiguity; however, much still needs to be learned about the biases present.

DISCLOSURE STATEMENT

R.P. is a cofounder of Ocean Genomics.

LITERATURE CITED

1. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, et al. 2008. Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* 133:523–36
2. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, et al. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 320:1344–49
3. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5:621–28

4. Cloonan N, Forrest ARR, Kolle G, Gardiner BBA, Faulkner GJ, et al. 2008. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods* 5:613–19
5. Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, et al. 2008. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* 453:1239–43
6. Palazzo AF, Lee ES. 2015. Non-coding RNA: What is functional and what is junk? *Front. Genet.* 6:2
7. Zhao W, He X, Hoadley KA, Parker JS, Hayes DN, Perou CM. 2014. Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling. *BMC Genom.* 15:419
8. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456:53–59
9. Ju J, Kim DH, Bi L, Meng Q, Bai X, et al. 2006. Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. *PNAS* 103:19635–40
10. Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, et al. 2010. Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat. Methods* 7:709–15
11. Zhao S, Zhang Y, Gordon W, Quan J, Xi H, et al. 2015. Comparison of stranded and non-stranded RNA-seq transcriptome profiling and investigation of gene overlap. *BMC Genom.* 16:675
12. Parkhomchuk D, Borodina T, Amstislavskiy V, Banaru M, Hallen L, et al. 2009. Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Res.* 37:e123
13. Mamanova L, Turner DJ. 2011. Low-bias, strand-specific transcriptome Illumina sequencing by on-flowcell reverse transcription (FRT-seq). *Nat. Protoc.* 6:1736–47
14. Wang B, Tseng E, Regulski M, Clark TA, Hon T, et al. 2016. Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. *Nat. Commun.* 7:11708
15. Collado-Torres L, Nellore A, Kammers K, Ellis SE, Taub MA, et al. 2017. Reproducible RNA-seq analysis using *recount2*. *Nat. Biotechnol.* 35:319–21
16. Lazic SE. 2017. *Experimental Design for Laboratory Biologists: Maximising Information and Improving Reproducibility*. Cambridge, UK: Cambridge Univ. Press. 1st ed.
17. Hart SN, Therneau TM, Zhang Y, Poland GA, Kocher J-P. 2013. Calculating sample size estimates for RNA sequencing data. *J. Comput. Biol.* 20:970–78
18. Guo Y, Zhao S, Li CI, Sheng Q, Shyr Y. 2014. RNAseqPS: a web tool for estimating sample size and power for RNAseq experiment. *Cancer Inform.* 13:1–5
19. Zhao S, Li C-I, Guo Y, Sheng Q, Shyr Y. 2018. RnaSeqSampleSize: real data based sample size estimation for RNA sequencing. *BMC Bioinform.* 19:191
20. Busby MA, Stewart C, Miller CA, Grzeda KR, Marth GT. 2013. Scotty: a web tool for designing RNA-Seq experiments to measure differential gene expression. *Bioinformatics* 29:656–57
21. Poplawski A, Binder H. 2018. Feasibility of sample size calculation for RNA-seq studies. *Brief. Bioinform.* 19:713–20
22. Oshlack A, Wakefield MJ. 2009. Transcript length bias in RNA-seq data confounds systems biology. *Biol. Direct.* 4:14
23. Liu Y, Zhou J, White KP. 2014. RNA-seq differential expression studies: More sequence or more replication? *Bioinformatics* 30:301–4
24. Schurch NJ, Schofield P, Gierliński M, Cole C, Sherstnev A, et al. 2016. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA* 22:839–51
25. Mercer TR, Clark MB, Crawford J, Brunck ME, Gerhardt DJ, et al. 2014. Targeted sequencing for gene discovery and quantification using RNA CaptureSeq. *Nat. Protoc.* 9:989–1009
26. Cabanski CR, Magrini V, Griffith M, Griffith OL, McGrath S, et al. 2014. cDNA hybrid capture improves transcriptome analysis on low-input and archived samples. *J. Mol. Diagn.* 16:440–51
27. Irimia M, Weatheritt RJ, Ellis JD, Parikshak NN, Gonatopoulos-Pournatzis T, et al. 2014. A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell* 159:1511–23
28. Eom T, Zhang C, Wang H, Lay K, Fak J, et al. 2013. NOVA-dependent regulation of cryptic NMD exons controls synaptic protein levels after seizure. *eLife* 2:e00178

29. Fratta P, Sivakumar P, Humphrey J, Lo K, Ricketts T, et al. 2018. Mice with endogenous TDP-43 mutations exhibit gain of splicing function and characteristics of amyotrophic lateral sclerosis. *EMBO J.* 37:e98684
30. Salzman J, Gawad C, Wang PL, Lacayo N, Brown PO. 2012. Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types. *PLOS ONE* 7:e30733
31. Kim T-K, Hemberg M, Gray JM. 2015. Enhancer RNAs: a class of long noncoding RNAs synthesized at enhancers. *Cold Spring Harb. Perspect. Biol.* 7:a018622
32. Parker BC, Zhang W. 2013. Fusion genes in solid tumors: an emerging target for cancer diagnosis and treatment. *Chin. J. Cancer* 32:594–603
33. Frye M, Jaffrey SR, Pan T, Rechavi G, Suzuki T. 2016. RNA modifications: What have we learned and where are we headed? *Nat. Rev. Genet.* 17:365–72
34. Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, et al. 2000. Molecular portraits of human breast tumours. *Nature* 406:747–52
35. Climente-González H, Porta-Pardo E, Godzik A, Eyras E. 2017. The functional impact of alternative splicing in cancer. *Cell Rep.* 20:2215–26
36. Cieślik M, Chinnaiyan AM. 2017. Cancer transcriptome profiling at the juncture of clinical translation. *Nat. Rev. Genet.* 19:93–109
37. Pedersen G, Kanigan T. 2016. Clinical RNA sequencing in oncology: Where are we? *Per Med.* 13:209–13
38. Ståhl PL, Salmén F, Vickovic S, Lundmark A, Navarro JF, et al. 2016. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* 353:78–82
39. Westermann AJ, Gorski SA, Vogel J. 2012. Dual RNA-seq of pathogen and host. *Nat. Rev. Microbiol.* 10:618–30
40. Piskol R, Ramaswami G, Li JB. 2013. Reliable identification of genomic variants from RNA-seq data. *Am. J. Hum. Genet.* 93:641–51
41. Park E, Williams B, Wold BJ, Mortazavi A. 2012. RNA editing in the human ENCODE RNA-seq data. *Genome Res.* 22:1626–33
42. Uszczynska-Ratajczak B, Lagarde J, Frankish A, Guigó R, Johnson R. 2018. Towards a complete map of the human long non-coding RNA transcriptome. *Nat. Rev. Genet.* 19:535–48
43. Bashiardes S, Zilberman-Schapira G, Elinav E. 2016. Use of metatranscriptomics in microbiome research. *Bioinform. Biol. Insights* 10:19–25
44. Racle J, de Jonge K, Baumgaertner P, Speiser DE, Gfeller D. 2017. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. *eLife* 6:e26476
45. Martin JA, Wang Z. 2011. Next-generation transcriptome assembly. *Nat. Rev. Genet.* 12:671–82
46. Castel SE, Levy-Moonshine A, Mohammadi P, Banks E, Lappalainen T. 2015. Tools and best practices for data processing in allelic expression analysis. *Genome Biol.* 16:195
47. Sun W, Hu Y. 2013. eQTL mapping using RNA-seq data. *Stat. Biosci.* 5:198–219
48. Alamancos GP, Agirre E, Eyras E. 2014. Methods to study splicing from high-throughput RNA sequencing data. *Methods Mol. Biol.* 1126:357–97
49. van Dam S, Võsa U, van der Graaf A, Franke L, de Magalhães JP. 2018. Gene co-expression analysis for functional classification and gene-disease predictions. *Brief. Bioinform.* 19:575–92
50. Khatri P, Sirota M, Butte AJ. 2012. Ten years of pathway analysis: current approaches and outstanding challenges. *PLOS Comput. Biol.* 8:e1002375
51. de Leeuw CA, Neale BM, Heskes T, Posthuma D. 2016. The statistical properties of gene-set analysis. *Nat. Rev. Genet.* 17:353–64
52. Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25:1105–11
53. Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10:R25
54. Engström PG, Steijger T, Sipos B, Grant GR, Kahles A, et al. 2013. Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat. Methods* 10:1185–91
55. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, et al. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29:15–21

56. Kim D, Langmead B, Salzberg SL. 2015. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* 12:357–60
57. Liao Y, Smyth GK, Shi W. 2013. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res.* 41:e108
58. Wu TD, Watanabe CK. 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* 21:1859–75
59. Lin H-N, Hsu W-L. 2017. DART: a fast and accurate RNA-seq mapper with a partitioning strategy. *Bioinformatics* 34:190–97
60. Sedlazeck FJ, Rescheneder P, von Haeseler A. 2013. NextGenMap: fast and accurate read mapping in highly polymorphic genomes. *Bioinformatics* 29:2790–91
61. Medina I, Tárraga J, Martínez H, Barrachina S, Castillo MI, et al. 2016. Highly sensitive and ultrafast read mapping for RNA-seq analysis. *DNA Res.* 23:93–100
62. Baruzzo G, Hayer KE, Kim EJ, Di Camillo B, FitzGerald GA, Grant GR. 2017. Simulation-based comprehensive benchmarking of RNA-seq aligners. *Nat. Methods* 14:135–39
63. Bonfert T, Kirner E, Csaba G, Zimmer R, Friedel CC. 2015. ContextMap 2: fast and accurate context-based RNA-seq mapping. *BMC Bioinform.* 16:122
64. Wu TD, Nacu S. 2010. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* 26:873–81
65. Aken BL, Ayling S, Barrell D, Clarke L, Curwen V, et al. 2016. The Ensembl gene annotation system. *Database* 2016:baw093
66. Liao Y, Smyth GK, Shi W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30:923–30
67. Anders S, Pyl PT, Huber W. 2015. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31:166–69
68. Roberts A, Trapnell C, Donaghey J, Rinn JL, Pachter L. 2011. Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biol.* 12:R22
69. Hansen KD, Brenner SE, Dudoit S. 2010. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res.* 38:e131
70. Liu X, Shi X, Chen C, Zhang L. 2015. Improving RNA-Seq expression estimation by modeling isoform- and exon-specific read sequencing rate. *BMC Bioinform.* 16:332
71. Love MI, Hogenesch JB, Irizarry RA. 2016. Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation. *Nat. Biotechnol.* 34:1287–91
72. Robert C, Watson M. 2015. Errors in RNA-Seq quantification affect genes of relevance to human disease. *Genome Biol.* 16:177
73. Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinform.* 12:323
74. Soneson C, Love MI, Robinson MD. 2015. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research* 4:1521
75. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, et al. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* 7:562–78
76. Xing Y, Yu T, Wu YN, Roy M, Kim J, Lee C. 2006. An expectation-maximization algorithm for probabilistic reconstructions of full-length isoforms from splice graphs. *Nucleic Acids Res.* 34:3150–60
77. Jiang H, Wong WH. 2009. Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics* 25:1026–32
78. Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN. 2010. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics* 26:493–500
79. Turro E, Su S-Y, Gonçalves Â, Coin LJM, Richardson S, Lewin A. 2011. Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. *Genome Biol.* 12:R13
80. Richard H, Schulz MH, Sultan M, Nürnberger A, Schrinner S, et al. 2010. Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments. *Nucleic Acids Res.* 38:e112
81. Nicolae M, Mangul S, Măndoiu II, Zelikovsky A. 2011. Estimation of alternative splicing isoform frequencies from RNA-Seq data. *Algorithms Mol. Biol.* 6:9

82. Mezlini AM, Smith EJM, Fiume M, Buske O, Savich GL, et al. 2013. iReckon: simultaneous isoform discovery and abundance estimation from RNA-seq data. *Genome Res.* 23:519–29
83. Zakeri M, Srivastava A, Almodaresi F, Patro R. 2017. Improved data-driven likelihood factorizations for transcript abundance estimation. *Bioinformatics* 33:i142–51
84. Roberts A, Pachter L. 2013. Streaming fragment assignment for real-time analysis of sequencing experiments. *Nat. Methods* 10:71–73
85. Cappé O, Moulines E. 2009. On-line expectation–maximization algorithm for latent data models. *J. R. Stat. Soc. B* 71:593–613
86. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, et al. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28:511–15
87. Li W, Feng J, Jiang T. 2011. IsoLasso: a LASSO regression approach to RNA-Seq based transcriptome assembly. *J. Comput. Biol.* 18:1693–707
88. Canzar S, Andreotti S, Weese D, Reinert K, Klau GW. 2016. CIDANE: comprehensive isoform discovery and abundance estimation. *Genome Biol.* 17:16
89. Maretty L, Sibbesen JA, Krogh A. 2014. Bayesian transcriptome assembly. *Genome Biol.* 15:501
90. Shi X, Wang X, Wang T-L, Hilakivi-Clarke L, Clarke R, Xuan J. 2018. SparseIso: a novel Bayesian approach to identify alternatively spliced isoforms from RNA-seq data. *Bioinformatics* 34:56–63
91. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* 33:290–95
92. Tomescu AI, Kuosmanen A, Rizzi R, Mäkinen V. 2013. A novel min-cost flow method for estimating transcript expression with RNA-Seq. *BMC Bioinform.* 14(Suppl. 5):S15
93. Bernard E, Jacob L, Mairal J, Vert J-P. 2014. Efficient RNA isoform identification and quantification from RNA-Seq data with network flows. *Bioinformatics* 30:2447–55
94. Shao M, Kingsford C. 2017. Accurate assembly of transcripts through phase-preserving graph decomposition. *Nat. Biotechnol.* 35:1167–69
95. Glaus P, Honkela A, Rattray M. 2012. Identifying differentially expressed transcripts from RNA-seq data with biological variation. *Bioinformatics* 28:1721–28
96. SEQC/MAQC-III Consort. 2014. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat. Biotechnol.* 32:903–14
97. Hensman J, Papastamoulis P, Glaus P, Honkela A, Rattray M. 2015. Fast and accurate approximate inference of transcript expression from RNA-seq data. *Bioinformatics* 31:3881–89
98. Nariai N, Hirose O, Kojima K, Nagasaki M. 2013. TIGAR: transcript isoform abundance estimation method with gapped alignment of RNA-Seq data by variational Bayesian inference. *Bioinformatics* 29:2292–99
99. Amari S-I, Nagaoka H. 2000. *Methods of Information Geometry*, transl. D Harada. Transl. Math. Monogr. 191. Oxford: Am. Math. Soc.
100. Patro R, Mount SM, Kingsford C. 2014. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat. Biotechnol.* 32:462–64
101. Varadhan R, Roland C. 2004. *Squared extrapolation methods (SQUAREM): a new class of simple and efficient numerical schemes for accelerating the convergence of the EM algorithm*. Work. Pap. 63, Johns Hopkins Univ. Dep. Biostat., Baltimore, MD. **https://biostats.bepress.com/jhubiostat/paper63/**
102. Zhang Z, Wang W. 2014. RNA-Skim: a rapid method for RNA-Seq quantification at transcript level. *Bioinformatics* 30:i283–92
103. Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* 34:525–27
104. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. 2017. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* 14:417–19
105. Foulds J, Boyles L, DuBois C, Smyth P, Welling M. 2013. Stochastic collapsed variational Bayesian inference for latent Dirichlet allocation. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ed. R Ghani, TE Senator, P Bradley, R Parekh, J He, pp. 446–54. New York: Assoc. Comput. Mach.

106. Ju CJ-T, Li R, Wu Z, Jiang J-Y, Yang Z, Wang W. 2017. Fleximer: accurate quantification of RNA-Seq via variable-length k-mers. In *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pp. 263–72. New York: Assoc. Comput. Mach.
107. Kanitz A, Gypas F, Gruber AJ, Gruber AR, Martin G, Zavolan M. 2015. Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data. *Genome Biol*. 16:150
108. Germain P-L, Vitriolo A, Adamo A, Laise P, Das V, Testa G. 2016. RNAontheBENCH: computational and empirical resources for benchmarking RNAseq quantification and differential expression methods. *Nucleic Acids Res*. 44:5054–67
109. Teng M, Love MI, Davis CA, Djebali S, Dobin A, et al. 2016. A benchmark for RNA-seq quantification pipelines. *Genome Biol*. 17:74
110. Zhang C, Zhang B, Lin L-L, Zhao S. 2017. Evaluation and comparison of computational tools for RNA-seq isoform quantification. *BMC Genom*. 18:583
111. Prakash C, Haeseler AV. 2017. An enumerative combinatorics model for fragmentation patterns in RNA sequencing provides insights into nonuniformity of the expected fragment starting-point and coverage profile. *J. Comput. Biol.* 24:200–12
112. Jones DC, Kuppusamy KT, Palpant NJ, Peng X, Murry CE, et al. 2016. Isolator: accurate and stable analysis of isoform-level expression in RNA-Seq experiments. bioRxiv 088765. **https://doi.org/10.1101/088765**
113. Soneson C, Love MI, Patro R, Hussain S, Malhotra D, Robinson MD. 2018. A junction coverage compatibility score to quantify the reliability of transcript abundance estimates and annotation catalogs. *Life Sci. Alliance* 2:e201800175
114. Efron B, Hastie T. 2016. *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. New York: Cambridge Univ. Press. 1st ed.
115. Tusher VG, Tibshirani R, Chu G. 2001. Significance analysis of microarrays applied to the ionizing radiation response. *PNAS* 98:5116–21
116. Smyth GK. 2004. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* 3:3
117. Bourgon R, Gentleman R, Huber W. 2010. Independent filtering increases detection power for high-throughput experiments. *PNAS* 107:9546–51
118. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. 2008. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* 18:1509–17
119. Robinson MD, Oshlack A. 2010. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*. 11:R25
120. Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biol*. 11:R106
121. Dillies M-A, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, et al. 2013. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief. Bioinform.* 14:671–83
122. Hansen KD, Irizarry RA, Wu Z. 2012. Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics* 13:204–16
123. Risso D, Schwartz K, Sherlock G, Dudoit S. 2011. GC-content normalization for RNA-Seq data. *BMC Bioinform*. 12:480
124. Lovén J, Orlando DA, Sigova AA, Lin CY, Rahl PB, et al. 2012. Revisiting global gene expression analysis. *Cell* 151:476–82
125. Risso D, Ngai J, Speed TP, Dudoit S. 2014. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.* 32:896–902
126. Taruttis F, Feist M, Schwarzfischer P, Gronwald W, Kube D, et al. 2017. External calibration with *Drosophila* whole-cell spike-ins delivers absolute mRNA fold changes from human RNA-Seq and qPCR data. *Biotechniques* 62:53–61
127. Hicks SC, Okrah K, Paulson JN, Quackenbush J, Irizarry RA, Bravo HC. 2018. Smooth quantile normalization. *Biostatistics* 19:185–98

128. Vallejos CA, Risso D, Scialdone A, Dudoit S, Marioni JC. 2017. Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nat. Methods* 14:565–71
129. Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15:550
130. Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26:139–40
131. Hardcastle TJ, Kelly KA. 2010. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinform.* 11:422
132. Leng N, Dawson JA, Thomson JA, Ruotti V, Rissman AI, et al. 2013. EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics* 29:1035–43
133. Robinson MD, Smyth GK. 2007. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* 23:2881–87
134. McCarthy DJ, Chen Y, Smyth GK. 2012. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* 40:4288–97
135. Himes BE, Jiang X, Wagner P, Hu R, Wang Q, et al. 2014. RNA-Seq transcriptome profiling identifies CRISPLD2 as a glucocorticoid responsive gene that modulates cytokine function in airway smooth muscle cells. *PLOS ONE* 9:e99625
136. Love MI, Anders S, Kim V, Huber W. 2016. RNA-Seq workflow: gene-level exploratory analysis and differential expression. *F1000Research* 4:1070
137. Robinson MD, Smyth GK. 2008. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics* 9:321–32
138. Soneson C, Delorenzi M. 2013. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinform.* 14:91
139. Zhou X, Lindsay H, Robinson MD. 2014. Robustly detecting differential expression in RNA sequencing data using observation weights. *Nucleic Acids Res.* 42:e91
140. Cox DR, Reid N. 1987. Parameter orthogonality and approximate conditional inference. *J. R. Stat. Soc. B* 49:1–39
141. Chen Y, Lun ATL, Smyth GK. 2014. Differential expression analysis of complex RNA-seq experiments using edgeR. In *Statistical Analysis of Next Generation Sequencing Data*, ed. S Datta, D Nettleton, pp. 51–74. Cham, Switz.: Springer Int.
142. Wu H, Wang C, Wu Z. 2013. A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. *Biostatistics* 14:232–43
143. Nelder JA, Wedderburn RWM. 1972. Generalized linear models. *J. R. Stat. Soc. A* 135:370–84
144. Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* 57:289–300
145. Benjamini Y, Yekutieli D. 2001. The control of the false discovery rate in multiple testing under depencency. *Ann. Stat.* 29:1165–88
146. Lund SP, Nettleton D, McCarthy DJ, Smyth GK. 2012. Detecting differential expression in RNA-sequence data using quasi-likelihood with shrunken dispersion estimates. *Stat. Appl. Genet. Mol. Biol.* 11:5
147. Di Y, Schafer DW, Cumbie JS, Chang JH. 2011. The NBP negative binomial model for assessing differential gene expression from RNA-Seq. *Stat. Appl. Genet. Mol. Biol.* 10:24
148. van de Wiel MA, Leday GGR, Pardo L, Rue H, van der Vaart AW, van Wieringen WN. 2013. Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors. *Biostatistics* 14:113–28
149. van de Wiel MA, Neerincx M, Buffart TE, Sie D, Verheul HMW. 2014. ShrinkBayes: a versatile R-package for analysis of count-based sequencing data in complex study designs. *BMC Bioinform.* 15:116
150. Rue H, Martino S, Chopin N. 2009. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. B* 71:319–92
151. Law CW, Chen Y, Shi W, Smyth GK. 2014. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* 15:R29
152. Li J, Tibshirani R. 2013. Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. *Stat. Methods Med. Res.* 22:519–36

153. Stephens M. 2016. False discovery rates: a new deal. *Biostatistics* 18:275–94
154. Zhu A, Ibrahim JG, Love MI. 2018. Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. *Bioinformatics*. In press. **https://doi.org/10.1093/bioinformatics/bty895**
155. Leek JT, Storey JD. 2007. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLOS Genet*. 3:12
156. Leek JT. 2014. svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Res*. 42:e161
157. Finner H, Roters M. 2001. On the false discovery rate and expected type I errors. *Biomet. J.* 43:985–1005
158. McCarthy DJ, Smyth GK. 2009. Testing significance relative to a fold-change threshold is a TREAT. *Bioinformatics* 25:765–71
159. Chen Y, Lun ATL, Smyth GK. 2016. From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. *F1000Research* 5:1438
160. Di Y, Emerson SC, Schafer DW, Kimbrel JA, Chang JH. 2013. Higher order asymptotics for negative binomial regression inferences from RNA-sequencing data. *Stat. Appl. Genet. Mol. Biol.* 12:49–70
161. Storey JD. 2003. The positive false discovery rate: a Bayesian interpretation and the *q*-value. *Ann. Stat.* 31:2013–35
162. Ignatiadis N, Klaus B, Zaugg JB, Huber W. 2016. Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nat. Methods* 13:577–80
163. Efron B. 2004. Large-scale simultaneous hypothesis testing. *J. Am. Stat. Assoc.* 99:96–104
164. Efron B. 2007. Size, power and false discovery rates. *Ann. Stat*. 35:1351–77
165. Van den Berge K, Soneson C, Robinson MD, Clement L. 2017. stageR: a general stage-wise method for controlling the gene-level false discovery rate in differential expression and differential transcript usage. *Genome Biol*. 18:151
166. Heller R, Manduchi E, Grant GR, Ewens WJ. 2009. A flexible two-stage procedure for identifying gene sets that are differentially expressed. *Bioinformatics* 25:1019–25
167. Kakaradov B, Xiong HY, Lee LJ, Jojic N, Frey BJ. 2012. Challenges in estimating percent inclusion of alternatively spliced junctions from RNA-seq data. *BMC Bioinform*. 13(Suppl. 6):S11
168. Turro E, Astle WJ, Tavaré S. 2014. Flexible analysis of RNA-seq data using mixed effects models. *Bioinformatics* 30:180–88
169. Papastamoulis P, Rattray M. 2018. A Bayesian model selection approach for identifying differentially expressed transcripts from RNA sequencing data. *J. R. Stat. Soc. C* 67:3–23
170. Pimentel H, Bray NL, Puente S, Melsted P, Pachter L. 2017. Differential analysis of RNA-seq incorporating quantification uncertainty. *Nat. Methods* 14:687–89
171. Blekhman R, Marioni JC, Zumbo P, Stephens M, Gilad Y. 2010. Sex-specific and lineage-specific alternative splicing in primates. *Genome Res*. 20:180–89
172. Purdom E, Simpson KM, Robinson MD, Conboy JG, Lapuk AV, Speed TP. 2008. FIRMA: a method for detection of alternative splicing from exon array data. *Bioinformatics* 24:1707–14
173. Anders S, Reyes A, Huber W. 2012. Detecting differential usage of exons from RNA-seq data. *Genome Res*. 22:2008–17
174. Soneson C, Matthes KL, Nowicka M, Law CW, Robinson MD. 2016. Isoform prefiltering improves performance of count-based methods for analysis of differential transcript usage. *Genome Biol*. 17:12
175. Love MI, Soneson C, Patro R. 2018. Swimming downstream: statistical analysis of differential transcript usage following Salmon quantification. *F1000Research* 7:952
176. Nowicka M, Robinson MD. 2016. DRIMSeq: a Dirichlet-multinomial framework for multivariate count outcomes in genomics. *F1000Research* 5:1356
177. Li YI, Knowles DA, Humphrey J, Barbeira AN, Dickinson SP, et al. 2017. Annotation-free quantification of RNA splicing using LeafCutter. *Nat. Genet.* 50:151–58
178. Papastamoulis P, Rattray M. 2017. Bayesian estimation of differential transcript usage from RNA-seq data. *Stat. Appl. Genet. Mol. Biol.* 16:367–86

179. Froussios K, Mourão K, Simpson GG, Barton GJ, Schurch NJ. 2017. Identifying differential isoform abundance with RATs: a universal tool and a warning. bioRxiv 132761. **https://doi.org/10.1101/132761**
180. Shen S, Park JW, Lu Z-X, Lin L, Henry MD, et al. 2014. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *PNAS* 111:E5593–601
181. Trincado JL, Entizne JC, Hysenaj G, Singh B, Skalic M, et al. 2018. SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biol.* 19:40
182. Yi L, Pimentel H, Bray NL, Pachter L. 2018. Gene-level differential analysis at transcript-level resolution. *Genome Biol.* 19:53
183. Hicks SC, Townes FW, Teng M, Irizarry RA. 2017. Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics* 19:562–78
184. Marinov GK, Williams BA, McCue K, Schroth GP, Gertz J, et al. 2014. From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. *Genome Res.* 24:496–510
185. Svensson V, Vento-Tormo R, Teichmann SA. 2018. Exponential scaling of single-cell RNA-seq in the past decade. *Nat. Protoc.* 13:599–604
186. Wagner A, Regev A, Yosef N. 2016. Revealing the vectors of cellular identity with single-cell genomics. *Nat. Biotechnol.* 34:1145–60
187. Moor AE, Itzkovitz S. 2017. Spatial transcriptomics: paving the way for tissue-level systems biology. *Curr. Opin. Biotechnol.* 46:126–33
188. Kumar P, Tan Y, Cahan P. 2017. Understanding development and stem cells using single cell-based analyses of gene expression. *Development* 144:17–32
189. Kang HM, Subramaniam M, Targ S, Nguyen M, Maliskova L, et al. 2018. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* 36:89–94
190. Paulson KG, Voillet V, McAfee MS, Hunter DS, Wagener FD, et al. 2018. Acquired cancer resistance to combination immunotherapy from transcriptional loss of class I HLA. *Nat. Commun.* 9:3868
191. Giladi A, Amit I. 2018. Single-cell genomics: a stepping stone for future immunology discoveries. *Cell* 172:14–21
192. Trapnell C. 2015. Defining cell types and states with single-cell genomics. *Genome Res.* 25:1491–98
193. Sandberg R. 2014. Entering the era of single-cell transcriptomics in biology and medicine. *Nat. Methods* 11:22–24
194. Soneson C, Robinson MD. 2018. Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods* 15:255–61
195. Jaakkola MK, Seyednasrollah F, Mehmood A, Elo LL. 2017. Comparison of methods to detect differentially expressed genes between single-cell populations. *Brief. Bioinform.* 18:735–43
196. Kharchenko PV, Silberstein L, Scadden DT. 2014. Bayesian approach to single-cell differential expression analysis. *Nat. Methods* 11:740–42
197. Finak G, McDavid A, Yajima M, Deng J, Gersuk V, et al. 2015. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA-seq data. *Genome Biol.* 16:278
198. Van den Berge K, Perraudeau F, Soneson C, Love MI, Risso D, et al. 2018. Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications. *Genome Biol.* 19:24
199. Risso D, Perraudeau F, Gribkova S, Dudoit S, Vert J-P. 2018. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat. Commun.* 9:284
200. Eling N, Richard AC, Richardson S, Marioni JC, Vallejos CA. 2018. Correcting the mean-variance dependency for differential variability testing using single-cell RNA sequencing data. *Cell Syst.* 7:284–94.e12
201. Korthauer KD, Chu L-F, Newton MA, Li Y, Thomson J, et al. 2016. A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome Biol.* 17:222
202. Treangen TJ, Salzberg SL. 2012. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.* 13:36–46
203. Weirather JL, de Cesare M, Wang Y, Piazza P, Sebastiano V, et al. 2017. Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Research* 6:100

204. Steijger T, Abril JF, Engström PG, Kokocinski F, RGASP Consort., et al. 2013. Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods* 10:1177–84
205. Tilgner H, Grubert F, Sharon D, Snyder MP. 2014. Defining a personal, allele-specific, and single-molecule long-read transcriptome. *PNAS* 111:9869–74
206. Gonzalez-Garay ML. 2016. Introduction to isoform sequencing using Pacific Biosciences Technology (Iso-Seq). In *Transcriptomics and Gene Regulation*, ed. J Wu, pp. 141–60. Dordrecht, Neth.: Springer Neth.
207. Laver T, Harrison J, O'Neill PA, Moore K, Farbos A, et al. 2015. Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomol. Detect. Quantif.* 3:1–8
208. Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, et al. 2013. Characterizing and measuring bias in sequence data. *Genome Biol.* 14:R51
209. Teng JLL, Yeung ML, Chan E, Jia L, Lin CH, et al. 2017. PacBio but not Illumina technology can achieve fast, accurate and complete closure of the high GC, complex *Burkholderia pseudomallei* two-chromosome genome. *Front. Microbiol.* 8:1448
210. Eid J, Fehr A, Gray J, Luong K, Lyle J, et al. 2009. Real-time DNA sequencing from single polymerase molecules. *Science* 323:133–38
211. Levene MJ, Korlach J, Turner SW, Foquet M, Craighead HG, Webb WW. 2003. Zero-mode waveguides for single-molecule analysis at high concentrations. *Science* 299:682–86
212. Travers KJ, Chin C-S, Rank DR, Eid JS, Turner SW. 2010. A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Res.* 38:e159
213. Carneiro MO, Russ C, Ross MG, Gabriel SB, Nusbaum C, DePristo MA. 2012. Pacific Biosciences sequencing technology for genotyping and variation discovery in human data. *BMC Genom.* 13:375
214. Wang Y, Yang Q, Wang Z. 2014. The evolution of nanopore sequencing. *Front. Genet.* 5:449
215. Quick J, Quinlan AR, Loman NJ. 2014. A reference bacterial genome dataset generated on the MinION^{TM} portable single-molecule nanopore sequencer. *Gigascience* 3:22
216. Garalde DR, Snell EA, Jachimowicz D, Sipos B, Lloyd JH, et al. 2018. Highly parallel direct RNA sequencing on an array of nanopores. *Nat. Methods* 15:201–6
217. Smith AM, Jain M, Mulroney L, Garalde DR, Akeson M. 2017. Reading canonical and modified nucleotides in 16S ribosomal RNA using nanopore direct RNA sequencing. bioRxiv 132274. **http://doi.org/10.1101/132274**
218. Sharon D, Tilgner H, Grubert F, Snyder M. 2013. A single-molecule long-read survey of the human transcriptome. *Nat. Biotechnol.* 31:1009–14
219. Oikonomopoulos S, Wang YC, Djambazian H, Badescu D, Ragoussis J. 2016. Benchmarking of the Oxford Nanopore MinION sequencing for quantitative and qualitative assessment of cDNA populations. *Sci. Rep.* 6:31602
220. Abdel-Ghany SE, Hamilton M, Jacobi JL, Ngam P, Devitt N, et al. 2016. A survey of the sorghum transcriptome using single-molecule long reads. *Nat. Commun.* 7:11706
221. Au KF, Sebastiano V, Afshar PT, Durruthy JD, Lee L, et al. 2013. Characterization of the human ESC transcriptome by hybrid sequencing. *PNAS* 110:E4821–30
222. Treutlein B, Gokce O, Quake SR, Südhof TC. 2014. Cartography of neurexin alternative splicing mapped by single-molecule long-read mRNA sequencing. *PNAS* 111:E1291–99
223. Rhoads A, Au KF. 2015. PacBio sequencing and its applications. *Genom. Proteom. Bioinform.* 13:278–89
224. Marchet C, Lecompte L, Da Silva C, Cruaud C, Aury JM, et al. 2017. *De novo* clustering of long reads by gene from transcriptomics data. *Nucleic Acids Res.* 47:e2
225. Workman RE, Myrka AM, Wong GW, Tseng E, Welch KC Jr., Timp W. 2018. Single-molecule, full-length transcript sequencing provides insight into the extreme metabolism of the ruby-throated hummingbird *Archilochus colubris*. *Gigascience* 7:1–12
226. An D, Cao HX, Li C, Humbeck K, Wang W. 2018. Isoform sequencing and state-of-art applications for unravelling complexity of plant transcriptomes. *Genes* 9:43
227. Gordon SP, Tseng E, Salamov A, Zhang J, Meng X, et al. 2015. Widespread polycistronic transcripts in fungi revealed by single-molecule mRNA sequencing. *PLOS ONE* 10:e0132628
228. Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34:3094–100

Annual Review of Biomedical Data Science

Volume 2, 2019

Contents

Errata

An online log of corrections to *Annual Review of Biomedical Data Science* articles may be found at http://www.annualreviews.org/errata/biodatasci

3 ARMOR: An Automated Reproducible MOdular Workflow for Preprocessing and Differential Analysis of RNA-seq Data[1]

In this paper, we present ARMOR, a snakemake workflow for the analysis of RNA-seq data. ARMOR enables the automated and reproducible differential analysis of RNA-seq experiments. This project was a collaboration and all authors designed and implemented the workflow and wrote the documentation and manuscript.

[1]Originally published in Orjuela, S., Huang, R., Hembach, K. M., Robinson, M. D., & Soneson, C. "ARMOR: an Automated Reproducible MOdular workflow for preprocessing and differential analysis of RNA-seq data" *G3: Genes, Genomes, Genetics.* **9**:7 (2019).

G3 Genes | Genomes | Genetics

SOFTWARE AND DATA RESOURCES

ARMOR: An Automated Reproducible MOdular Workflow for Preprocessing and Differential Analysis of RNA-seq Data

Stephany Orjuela,*,†,‡,1,2 Ruizhu Huang,*,†,1,2 Katharina M. Hembach,*,†,§,1,2 Mark D. Robinson,*,†,3 and Charlotte Soneson*,†,3,4

*SIB Swiss Institute of Bioinformatics, Zurich, Switzerland, †Institute of Molecular Life Sciences, ‡Institute of Molecular Cancer Research, and §Department of Quantitative Biomedicine, University of Zurich, Zurich, Switzerland

ORCID IDs: 0000-0002-1508-461X (S.O.); 0000-0003-3285-1945 (R.H.); 0000-0001-5041-6205 (K.M.H.); 0000-0002-3048-5518 (M.D.R.); 0000-0003-3833-2169 (C.S.)

ABSTRACT The extensive generation of RNA sequencing (RNA-seq) data in the last decade has resulted in a myriad of specialized software for its analysis. Each software module typically targets a specific step within the analysis pipeline, making it necessary to join several of them to get a single cohesive workflow. Multiple software programs automating this procedure have been proposed, but often lack modularity, transparency or flexibility. We present ARMOR, which performs an end-to-end RNA-seq data analysis, from raw read files, via quality checks, alignment and quantification, to differential expression testing, geneset analysis and browser-based exploration of the data. ARMOR is implemented using the Snakemake workflow management system and leverages conda environments; Bioconductor objects are generated to facilitate downstream analysis, ensuring seamless integration with many R packages. The workflow is easily implemented by cloning the GitHub repository, replacing the supplied input and reference files and editing a configuration file. Although we have selected the tools currently included in ARMOR, the setup is modular and alternative tools can be easily integrated.

KEYWORDS
RNA sequencing
Differential expression
Exploratory data analysis
Quality control

Since the first high-throughput RNA-seq experiments about a decade ago, there has been a tremendous development in the understanding of the characteristic features of the collected data, as well as the methods used for the analysis. Today there are vetted, well-established algorithms and tools available for many aspects of RNA-seq data analysis (Conesa *et al.* 2016; Van Den Berge *et al.* 2018). In this study, we capitalize on this knowledge and present a modular, light-weight RNA-seq workflow covering the most common parts of a typical end-to-end RNA-seq data analysis with focus on differential expression. The application is implemented using the Snakemake workflow management system (Köster and Rahmann 2012), and allows the user to easily perform quality assessment, adapter trimming, genome alignment, transcript and gene abundance quantification, differential expression analysis and geneset analyses with a simple command, after specifying the required reference files and information about the experimental design in a configuration file. Reproducibility is ensured via the use of conda environments, and all relevant log files are retained for transparency. The output is provided in state-of-the-art R/Bioconductor objects, ensuring interoperability with a broad range of Bioconductor packages. In particular, we provide a template to facilitate browser-based interactive visualization of the quantified abundances and the results of the statistical analyses with iSEE (Rue-Albrecht *et al.* 2018).

Among already existing pipelines for automated reference-based RNA-seq analysis, several focus either on the preprocessing and quality control steps (He *et al.* 2018; Ewels *et al.* 2018; Tsyganov *et al.* 2018), or on the downstream analysis and visualization of differentially expressed

doi: https://doi.org/10.1534/g3.119.400185
Manuscript received March 12, 2019; accepted for publication May 13, 2019; published Early Online May 14, 2019.

Supplemental material available at FigShare: https://doi.org/10.25387/g3.8053280.
[1]These authors contributed equally to this work.
[2]The order of the shared first authors was determined randomly, using the sample() function in R v3.5.2, with the random seed 1552397284.
[3]Corresponding authors: E-mail: (mark.robinson@imls.uzh.ch) and (charlotte.soneson@fmi.ch)
[4]Present address: Friedrich Miescher Institute for Biomedical Research and SIB Swiss Institute of Bioinformatics, Basel, Switzerland

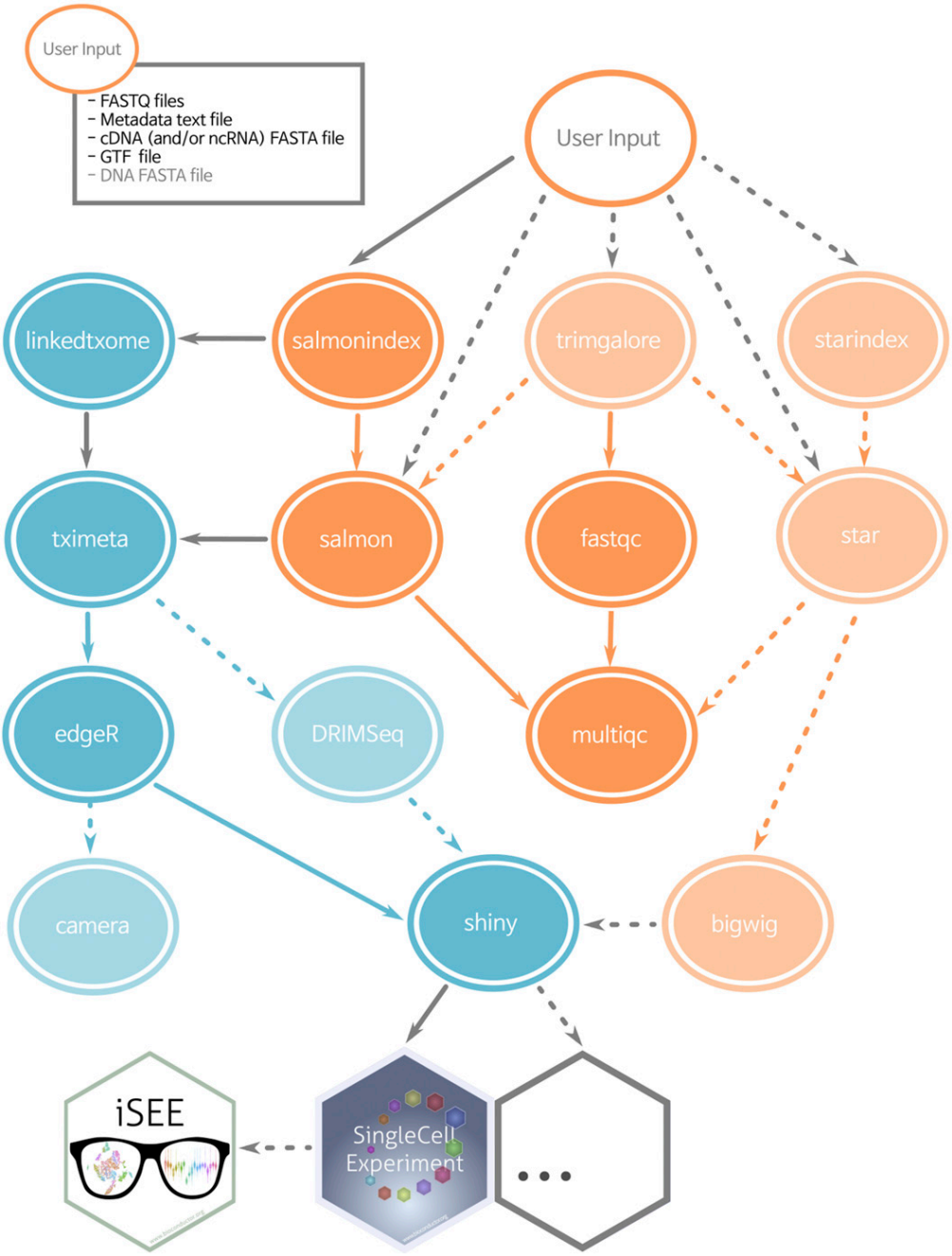


Figure 1 Simplified directed acyclic graph (DAG) of the ARMOR workflow. Blue ellipses are rules run in R, orange ellipses from software called as shell commands. Dashed lines and light-colored ellipses are optional rules, controlled in the configuration file. By default all rules are executed.

genes (Marini 2018; Monier *et al.* 2019; Powell 2018), or do not provide a single framework for the preprocessing and downstream analysis (Steinbaugh *et al.* 2018). Some workflows are based on predefined reference files and can only quantify abundances for human or mouse (Torre *et al.* 2018; Cornwell *et al.* 2018; Wang 2018). Additionally, workflows that conduct differential gene expression analysis often do not allow comparisons between more than two groups, or more complex experimental designs (Girke 2018; Cornwell *et al.* 2018). Some existing pipelines only provide a graphical user interface to design and execute fully automated analyses (Hung *et al.* 2018; Afgan *et al.* 2018). In addition to reference-based tools, there are also pipelines that perform *de novo* transcriptome assembly before downstream analysis (*e.g.*, https://github.com/dib-lab/elvers).

ARMOR performs both preprocessing and downstream statistical analysis of the RNA-seq data, building on standard statistical analysis methods and commonly used data containers. It distinguishes itself from existing workflows in several ways: (i) Its modularity, reflected in its fully and easily customizable framework. (ii) The transparency of the output and analysis, meaning that all code is accessible and can be modified by the user. (iii) The seamless integration with downstream analysis and visualization packages, especially those within Bioconductor (Huber *et al.* 2015; Amezquita *et al.* 2019). (iv) The ability to specify any fixed-effect experimental design and any number of contrasts, in a standardized format. (v) The inclusion of a test for differential transcript usage in addition to differential gene expression analysis. While high-performance computing environments and cloud computing are

not specifically targeted, Snakemake enables the usage of a cluster without the need to modify the workflow itself.

In general, we do not advocate fully automated analysis. All rigorous data analyses need exploratory steps and spot checks at various steps throughout the process, to ensure that data are of sufficient quality and to spot potential errors (*e.g.*, sample mislabelings). ARMOR handles the automation of "bookkeeping" tasks, such as running the correct sequence of software for all samples, and compiling the data and reports in standardized formats. If errors are identified, the workflow can re-run only the parts that need to be updated.

ARMOR is available from https://github.com/csoneson/ARMOR.

MATERIALS AND METHODS

Overview

The ARMOR workflow is designed to perform an end-to-end analysis of bulk RNA-seq data, starting from FASTQ files with raw sequencing reads (Figure 1). Reads first undergo quality control with FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and (optionally) adapter trimming using TrimGalore! (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), before being mapped to a transcriptome index using Salmon (Patro *et al.* 2017) and (optionally) aligned to the genome using STAR (Dobin *et al.* 2013). Estimated transcript abundances from Salmon are imported into R using the tximeta package (Soneson *et al.* 2015; Love *et al.* 2019) and analyzed for differential gene expression and (optionally) differential transcript usage with edgeR (Robinson *et al.* 2010) and DRIMSeq (Nowicka and Robinson 2016). The quantifications, provided metadata, and results from the statistical analyses are exported as SingleCellExperiment objects (Lun and Risso 2019) ensuring interoperability with a large part of the Bioconductor ecosystem (Huber *et al.* 2015; Amezquita *et al.* 2019). Quantification and quality control results are summarized in a MultiQC report (Ewels *et al.* 2016). Other tools can be easily exchanged for those listed above by modifying the Snakefile and/or the template analysis code.

Input file specification

ARMOR can be used to analyze RNA-seq data from any organism for which a reference transcriptome and (optionally) an annotated reference genome are available from either Ensembl (Zerbino *et al.* 2018) or GENCODE (Frankish *et al.* 2019). Paths to the reference files, as well as the FASTQ files with the sequencing reads, are specified by the user in a configuration file. In addition, the user prepares a metadata file – a tab-delimited text file listing the name of the samples, the library type (single- or paired-end) and any other covariates that will be used for the statistical analysis. The checkinputs rule in the Snakefile can be executed to make sure all the input files and the parameters in the configuration file have been correctly specified.

Workflow execution

ARMOR is implemented as a modular Snakemake (Köster and Rahmann 2012) workflow, and the execution of the individual steps is controlled by the provided Snakefile. Snakemake will automatically keep track of the dependencies between the different parts of the workflow; rerunning the workflow will thus only regenerate results that are out of date or missing given these dependencies. Via a set of variables specified in the configuration file, the user can easily decide to include or exclude the optional parts of the workflow (shaded ellipses in Figure 1). By adding or modifying targets in the Snakefile, users can include any additional or specialized types of analyses that are not covered by the original workflow.

```
ARMOR
├── config.yaml
├── envs
│   ├── environment_R.yaml
│   └── environment.yaml
├── example_data
│   ├── FASTQ
│   │   ├── SRR1039508_R1.fastq.gz
│   │   ├── SRR1039508_R2.fastq.gz
│   │   ├── SRR1039509_R1.fastq.gz
│   │   ├── SRR1039509_R2.fastq.gz
│   │   ├── SRR1039512_R1.fastq.gz
│   │   ├── SRR1039512_R2.fastq.gz
│   │   ├── SRR1039513_R1.fastq.gz
│   │   └── SRR1039513_R2.fastq.gz
│   ├── metadata.txt
│   ├── README.md
│   └── reference
│       ├── Ensembl.GRCh38.93
│       │   ├── Homo_sapiens.GRCh38.93.1.1.10M.gtf
│       │   ├── Homo_sapiens.GRCh38.cdna.all.1.1.10M.fa.gz
│       │   ├── Homo_sapiens.GRCh38.dna.chromosome.1.1.10M.fa
│       │   └── Homo_sapiens.GRCh38.ncrna.1.1.10M.fa.gz
│       └── Gencode28
│           ├── gencode.v28.annotation.1.1.10M.gtf
│           ├── gencode.v28.transcripts.1.1.10M.fa.gz
│           └── GRCh38.primary_assembly.genome.1.1.10M.fa
├── img
│   ├── ...
├── LICENSE
├── README.md
├── scripts
│   ├── custom_iSEE_panels.R
│   ├── DRIMSeq_dtu.Rmd
│   ├── edgeR_dge.Rmd
│   ├── generate_linkedTxome.R
│   ├── generate_report.R
│   ├── install_pkgs.R
│   ├── list_packages.R
│   ├── prepare_shiny.Rmd
│   ├── run_render.R
│   └── run_tximeta.R
└── Snakefile
```

Figure 2 The files and directory structure that make up the ARMOR workflow.

By default, all software packages that are needed for the analysis will be installed in an auto-generated conda environment, which will be automatically activated by Snakemake before the execution of each rule. The desired software versions can be specified in the provided environment file. If the user prefers, local installations of (all or a subset of) the required software can also be used (as described in **Software management**).

Software management

First, the user needs to have a recent version of Snakemake and conda installed. There is a range of possibilities to manage the software for the ARMOR workflow. The recommended option is to allow conda and the workflow itself to manage everything, including the installation of the needed R packages. The workflow is executed this way with the command

```
snakemake --use-conda
```

The first time the workflow is run, the conda environments will be generated and all necessary software will be installed. Any subsequent invocations of the workflow from this directory will use these generated environments. An alternative option is to use ARMOR's envs/environment.yaml file to create a conda environment that can be manually activated, by running the command

```
conda env create --name ARMOR \
                 --file envs/environment.yaml
conda activate ARMOR
```

The second command activates the environment. Once the environment is activated, ARMOR can be run by simply typing

Table 1 Metadata table for the Wnt signaling data

Names	type	condition
Q10-Chir-1_R1	SE	d4Tcf__chir
Q10-Chir-2_R1	SE	d4Tcf__chir
Q10-Chir-3_R1	SE	d4Tcf__chir
b-cat-KO-Chir-1_R1	SE	dBcat__chir
b-cat-KO-Chir-2_R1	SE	dBcat__chir
b-cat-KO-Chir-3_R1	SE	dBcat__chir
WT-Chir-1_R1	SE	WT__chir
WT-Chir-2_R1	SE	WT__chir
WT-Chir-3_R1	SE	WT__chir
Q10-unstim-1_R1	SE	d4Tcf__unstim
Q10-unstim-2_R1	SE	d4Tcf__unstim
Q10-unstim-3_R1	SE	d4Tcf__unstim
b-cat-KO-ustim-1_R1	SE	dBcat__unstim
b-cat-KO-ustim-2_R1	SE	dBcat__unstim
b-cat-KO-ustim-3_R1	SE	dBcat__unstim
WT-unstim-1_R1	SE	WT__unstim
WT-unstim-2_R1	SE	WT__unstim
WT-unstim-3_R1	SE	WT__unstim

`snakemake`

Additionally, the user can circumvent the use of conda, and make sure that all software is already available and in the user's PATH. For this, Snakemake and the tools listed in envs/environment.yaml need to be manually installed, in addition to a recent version of R and the R packages listed in scripts/install_pkgs.R.

For either of the options mentioned above, the `useCondaR` flag in the configuration file controls whether a local R installation, or a conda-installed R, will be used. If `useCondaR` is set to `False`, the path to a local R installation (*e.g.*, `Rbin:<path>`) must be specified in the configuration file, along with the path to the R package library (*e.g.*, `R_LIBS_USER="<path>"`) in the .Renviron file. If the specified R library does not contain the required packages, Snakemake will try to install them (*i.e.*, write permissions would be needed). ARMOR has been tested on macOS and Linux systems.

Statistical analysis

ARMOR uses the quasi-likelihood framework of edgeR (Robinson *et al.* 2010; Lun *et al.* 2016) to perform tests for differential gene expression, camera (Wu and Smyth 2012) to perform associated geneset analysis, and DRIMSeq (Nowicka and Robinson 2016) to test for differential transcript usage between conditions. All code to perform the statistical analyses is provided in Rmarkdown templates (Allaire *et al.* 2018; Xie *et al.* 2018), which are executed at runtime. This setup gives the user flexibility to use any experimental design supported by these tools, and to test any contrast(s) of interest, by specifying these in the configuration file using standard R syntax, *e.g.*,

```
design:"~ 0 + group"
contrast:groupA-groupB
```

Arbitrarily complex designs and multiple contrasts are supported. In addition, by editing the template code, users can easily configure the analysis, add additional plots, or even replace the statistical test if desired. After compilation, all code used for the statistical analysis, together with the results and version information for all packages used, is retained in a standalone html report, ensuring transparency and reproducibility and facilitating communication of the results.

Output files

The output files from all steps in the ARMOR workflow are stored in a user-specified output directory, together with log files for each step,

```
E-MTAB-7029
├── download_files.R
├── FASTQ
│   ├── b-cat-KO-Chir-1_R1.fastq.gz
│   ├── b-cat-KO-Chir-2_R1.fastq.gz
│   ├── b-cat-KO-Chir-3_R1.fastq.gz
│   ├── b-cat-KO-ustim-1_R1.fastq.gz
│   ├── b-cat-KO-ustim-2_R1.fastq.gz
│   ├── b-cat-KO-ustim-3_R1.fastq.gz
│   ├── Q10-Chir-1_R1.fastq.gz
│   ├── Q10-Chir-2_R1.fastq.gz
│   ├── Q10-Chir-3_R1.fastq.gz
│   ├── Q10-unstim-1_R1.fastq.gz
│   ├── Q10-unstim-2_R1.fastq.gz
│   ├── Q10-unstim-3_R1.fastq.gz
│   ├── WT-Chir-1_R1.fastq.gz
│   ├── WT-Chir-2_R1.fastq.gz
│   ├── WT-Chir-3_R1.fastq.gz
│   ├── WT-unstim-1_R1.fastq.gz
│   ├── WT-unstim-2_R1.fastq.gz
│   └── WT-unstim-3_R1.fastq.gz
├── metadata.txt
└── reference
    ├── Homo_sapiens.GRCh38.95.gtf
    ├── Homo_sapiens.GRCh38.cdna.all.fa.gz
    └── Homo_sapiens.GRCh38.dna.primary_assembly.fa
```

Figure 3 The suggested structure for the set of files that need to be organized to run ARMOR on a new dataset. The structure can deviate from this somewhat, since the location of the files can be specified in the corresponding config.yaml file.

including relevant software version information. A detailed summary of the output files generated by the workflow, including the shell command that was used to generate each of them, the time of creation, and information about whether the associated inputs, code or parameters have since been updated, can be obtained at any time by invoking Snakemake with the flag `-D` (or `--detailed-summary`). Using the benchmark directive of Snakemake, ARMOR also automatically generates additional text files summarizing the run time and peak memory usage of each executed rule.

The results from the statistical analyses are combined with the transcript- and gene-level quantifications and saved as SingleCellExperiment objects (Lun and Risso 2019), ensuring easy integration with a large number of Bioconductor packages for downstream analysis and visualization. For example, the results can be interactively explored using the iSEE package (Rue-Albrecht *et al.* 2018) and a template is provided for this.

Multiple project management

When managing multiple projects, the user might run ARMOR in multiple physical locations (*i.e.*, clone the repository in separate places). `snakemake --use-conda` will create a separate conda environment in each subdirectory, which means that the installed software may be duplicated. If disk space is a concern, building and activating a single conda environment (using the `conda env create` command as shown in the **Software management** section), and activating this before invoking each workflow may be beneficial. It is also possible to explicitly specify the path to the desired config.yaml configuration file when snakemake is called:

```
snakemake --configfile config.yaml
```

Thus, the same ARMOR installation can be used for multiple projects, by invoking it with a separate config.yaml file for each project.

By taking advantage of the Snakemake framework, ARMOR makes file and software organization relatively autonomous. Although we recommend using a file structure similar to the one used

```
E-MTAB-7029/output/
├── benchmarks
│   ├── bigwig_b-cat-KO-Chir-1_R1.txt
│   ├── ...
├── FastQC
│   ├── b-cat-KO-Chir-1_R1_fastqc.html
│   ├── b-cat-KO-Chir-1_R1_fastqc.zip
│   ├── b-cat-KO-Chir-1_R1_trimmed_fastqc.html
│   ├── b-cat-KO-Chir-1_R1_trimmed_fastqc.zip
│   └── ...
├── FASTQtrimmed
│   ├── b-cat-KO-Chir-1_R1.fastq.gz_trimming_report.txt
│   ├── b-cat-KO-Chir-1_R1_trimmed.fq.gz
│   ├── ...
├── logs
│   ├── bigwig_b-cat-KO-Chir-1_R1.log
│   ├── ...
├── MultiQC
│   ├── multiqc_data/
│   └── multiqc_report.html
├── outputR
│   ├── camera_dge_results_conditiond4Tcf__chir-conditiond4Tcf__unstim.txt
│   ├── ...
│   ├── camera_gsa.rds
│   ├── DRIMSeq_dtu_files
│   │   └── figure-html/
│   ├── DRIMSeq_dtu.html
│   ├── DRIMSeq_dtu.md
│   ├── DRIMSeq_dtu.rds
│   ├── DRIMSeq_dtu_results_conditiond4Tcf__chir-conditiond4Tcf__unstim.txt
│   ├── ...
│   ├── edgeR_dge_files
│   │   └── figure-html/
│   ├── edgeR_dge.html
│   ├── edgeR_dge.md
│   ├── edgeR_dge.rds
│   ├── edgeR_dge_results_conditiond4Tcf__chir-conditiond4Tcf__unstim.txt
│   ├── ...
│   ├── prepare_shiny.html
│   ├── prepare_shiny.md
│   ├── shiny_sce.rds
│   └── tximeta_se.rds
├── Rout
│   ├── generate_linkedtxome.Rout
│   ├── install_pkgs.Rout
│   ├── pkginstall_state.txt
│   ├── prepare_shiny.Rout
│   ├── run_dge_edgeR.Rout
│   ├── run_dtu_drimseq.Rout
│   └── tximeta_se.Rout
├── salmon
│   ├── b-cat-KO-Chir-1_R1
│   │   ├── aux_info/
│   │   ├── cmd_info.json
│   │   ├── lib_format_counts.json
│   │   ├── libParams/
│   │   ├── logs/
│   │   └── quant.sf
│   ├── ...
├── STAR
│   ├── b-cat-KO-Chir-1_R1
│   │   ├── b-cat-KO-Chir-1_R1_Aligned.sortedByCoord.out.bam
│   │   ├── b-cat-KO-Chir-1_R1_Aligned.sortedByCoord.out.bam.bai
│   │   ├── b-cat-KO-Chir-1_R1_Log.final.out
│   │   ├── b-cat-KO-Chir-1_R1_Log.out
│   │   ├── b-cat-KO-Chir-1_R1_Log.progress.out
│   │   └── b-cat-KO-Chir-1_R1_SJ.out.tab
│   ├── ...
└── STARbigwig
    ├── b-cat-KO-Chir-1_R1_Aligned.sortedByCoord.out.bw
    ├── ...
```

Figure 4 The set of output files from the workflow. This includes log files for every step and all the standard outputs of all the tools, such as R objects and scripts, BAM files, bigWig files and quantification tables. Note that the outputs for only one RNA-seq sample are shown; ... represents the set of output files for the remaining samples or contrasts. Directories ending in / contain extraneous files and are collapsed here.

for the example data provided in the repository (Figure 2), and managing all the software for a project in a conda environment, the user is free to use the same environment for different datasets, even if the files are located in several folders. This alternative is more of a "software-based" structure than the "project-based" structure we present with the pipeline. Either structure has its advantages and depending on the use case and level of expertise, both can be easily implemented using ARMOR.

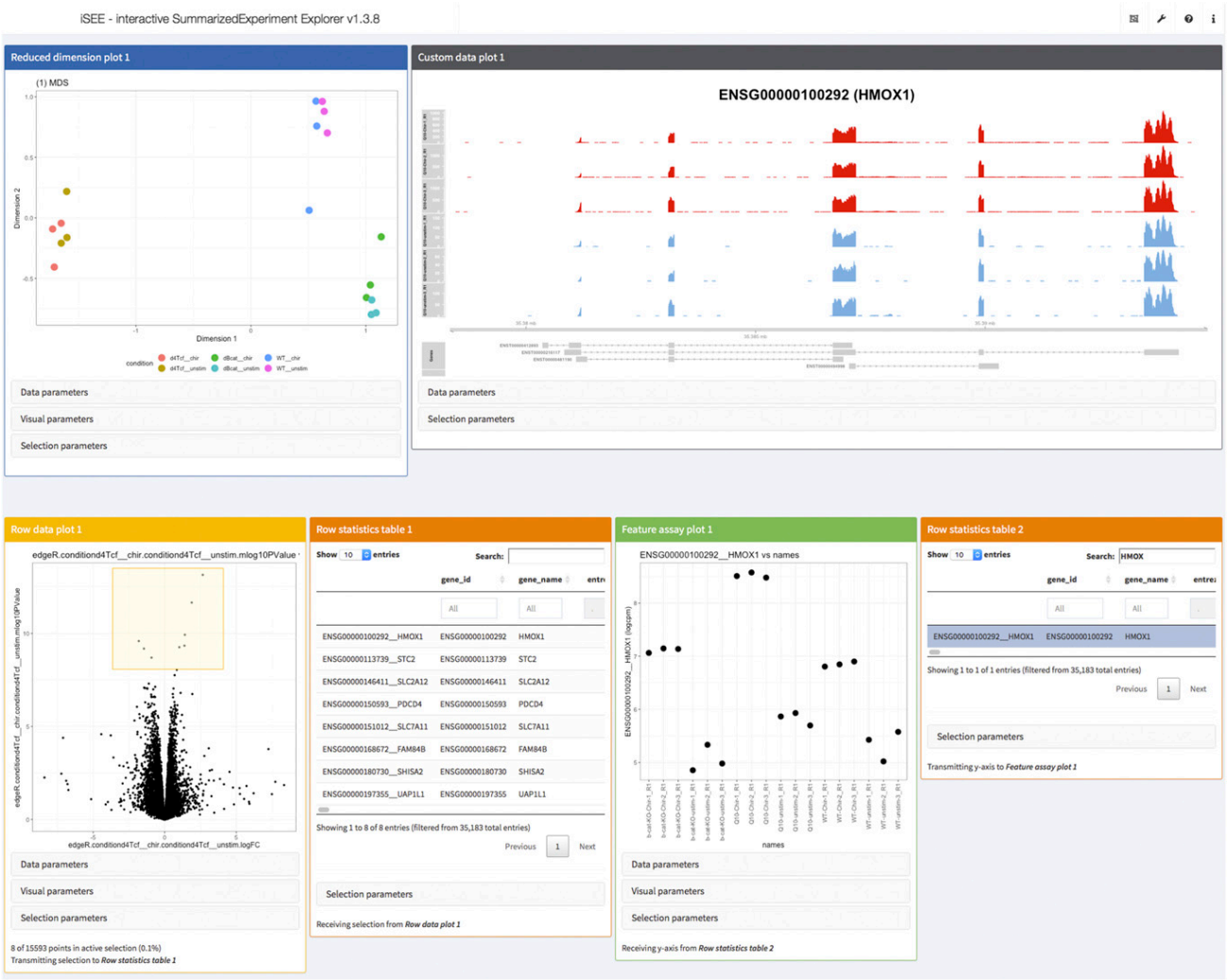


Figure 5 Screenshot of visualization of data and results from the real data walk-through using the iSEE R/Bioconductor package. The interactive application was configured to display an MDS plot colored by the sample condition (top left), a custom panel showing the observed read coverage of a selected gene (top right), a volcano plot for a specified contrast (bottom left, the selected genes are shown in the adjacent table) and an overview of the log-CPM expression values for each sample, for a gene selected in a second table (bottom right).

Code availability

ARMOR is available (under MIT license) from https://github.com/csoneson/ARMOR, together with a detailed walk-through of an example analysis. The repository also contains a wiki (https://github.com/csoneson/ARMOR/wiki), which is the main source of documentation for ARMOR and contains extensive information about the usage of the workflow.

Data Availability

Supplemental file DataS1.html contains the MultiQC report for the data used in the **Real data walk-through** section (ArrayExpress accession number E-MTAB-7029). Supplemental material available at FigShare: https://doi.org/10.25387/g3.8053280.

RESULTS AND DISCUSSION

The ARMOR skeleton

Figure 2 shows the set of files contained within the ARMOR workflow, and what is downloaded to the user's computer when the repository is cloned.

The example_data directory represents a (runnable) template of a very small dataset, which is useful for testing the software setup and the system as well as for having a structure to copy for a real project. The provided config.yaml file is pre-configured for this example dataset. We recommend that users prepare their own config.yaml and a similar directory structure to example_data, with the raw FASTQ files and reference sequence and annotation information in subfolders, perhaps using symbolic links if such files are already available in another location. We present an independent example below in the **Real data walk-through** section.

Once everything is set up, running `snakemake`, which operates on the rules in the Snakefile, will construct the hierarchy of instructions to execute, given the specifications in the config.yaml file. Snakemake automatically determines the dependencies between the rules and will invoke the instructions in a logical order. The scripts and envs directories, and the Snakefile itself, should not need to be modified, unless the user wants to customize certain aspects of the pipeline.

Real data walk-Through

Here, we illustrate the practical usage of ARMOR on a bulk RNA-seq dataset from a study on Wnt signaling (Doumpas *et al.* 2019). For each

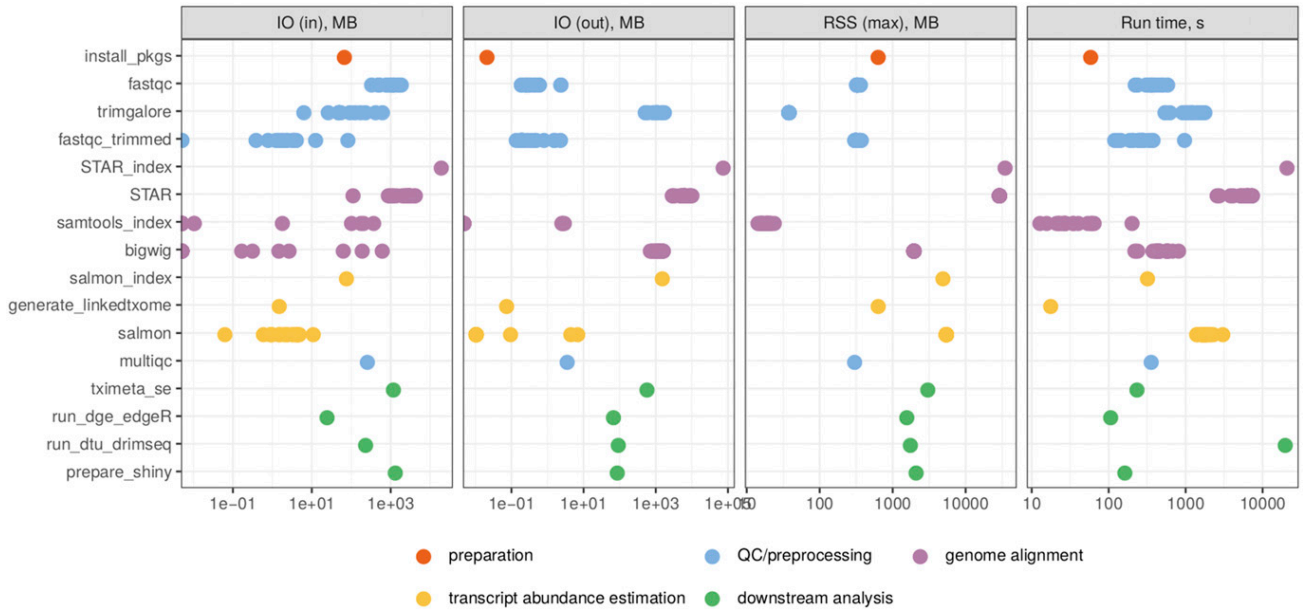


Figure 6 The required resources for the generation of each output file (grouped by the Snakemake rule) in the real data walk-through, as reported by the benchmarking directive of Snakemake. The four panels show the read and written bytes (in MB), the memory usage (in MB) and the run time (in seconds) of each rule. RSS = Resident Set Size. IO = Input/Output.

of three genetic backgrounds (HEK 293T, dBcat and d4TCF) and two experimental conditions (untreated and stimulated using the GSK3 inhibitor CHIRON99021), three biological replicates were measured (18 samples in total). The number of sequenced reads for each individual sample ranges from 12.5 to 41 million. A more detailed overview of the dataset is provided in the MultiQC report generated by the ARMOR run (Supplemental File DataS1.html). An R script (download_files.R, which can be found at https://github.com/csoneson/ARMOR/blob/chiron_realdataworkflow/E-MTAB-7029/download_files.R) was written to download the FASTQ files with raw reads from ArrayExpress (https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-7029/), and create a metadata table detailing the type of library and experimental condition for each sample (Table 1). This table was saved as a tab-delimited text file named metadata.txt.

The raw data and reference files were organized into a directory, E-MTAB-7029, with the structure according to Figure 3. The default config.yaml downloaded with the workflow was copied into a new file called config_E-MTAB-7029.yaml and edited to reflect the location of these files. In addition, the read length was set and the experimental design was specified as "`~ 0 + condition`", where the condition information will be taken from metadata.txt. Then, a set of contrasts of interest (*e.g.*, `conditiond4Tcf__chir-conditiond4Tcf__unstim`) were specified, as well as the set of genesets to use. The final configuration file can be viewed at https://github.com/csoneson/ARMOR/blob/chiron_realdataworkflow/config_E-MTAB-7029.yaml.

The set of files (not including the large data and reference files, which would be downloaded using the download_files.R) used in this setup can be found on the chiron_realdataworkflow branch of the ARMOR repository: https://github.com/csoneson/ARMOR/tree/chiron_realdataworkflow.

After downloading the data, generating the metadata.txt file and editing the config.yaml file, the full workflow was run with the command:

```
snakemake --use-conda--cores 20 \
          --configfile config_E-MTAB-7029.yaml
```

and upon completion of the workflow run, the specified output directory was populated as shown in Figure 4. The MultiQC directory contains a summary report of the quality assessment and alignment steps. In the outputR directory, reports of the statistical analyses (DRIMSeq_dtu.html and edgeR_dge.html), as well as a list of SingleCellExperiment objects (in shiny_sce.rds) are saved. The latter can be imported into R and used for further downstream analysis. Using the template run_iSEE.R (available from https://github.com/csoneson/ARMOR/blob/chiron_realdataworkflow/E-MTAB-7029/run_iSEE.R) and shiny_sce.rds (available from https://doi.org/10.6084/m9.figshare.8040239.v1), an R/shiny web application can be initiated, with various panels to allow the user to interactively explore the data and results (Figure 5).

Figure 6 shows the run time and maximal memory usage for generating each output file. Note that the `ncores` parameter in the configuration file was kept at 1, and thus each rule was run using a single thread. The most memory-intensive parts of the workflow, due to the large size of the reference genome, were the generation of the STAR index and the alignment of reads to the genome. The most time consuming parts were the generation of the STAR index and the DTU analysis with DRIMSeq. However, both of these can be executed using multiple cores, by increasing the value of the `ncores` parameter.

ACKNOWLEDGMENTS

M.D.R. acknowledges support from the University Research Priority Program Evolution in Action at the University of Zurich, the Swiss National Science Foundation (grants 310030_175841, CRSII5_177208) and the Chan Zuckerberg Initiative (grant 2018-182828).

LITERATURE CITED

Afgan, E., D. Baker, B. Batut, M. van den Beek, D. Bouvier *et al.*, 2018 The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res. 46: W537–W544. https://doi.org/10.1093/nar/gky379

Allaire, J., Y. Xie, J. McPherson, J. Luraschi, K. Ushey, *et al.*, 2018 rmarkdown: Dynamic Documents for *R*. R package version 1.11.

Amezquita, R. A., V. J. Carey, L. N. Carpp, L. Geistlinger, A. T. L. Lun *et al.*, 2019 Orchestrating Single-Cell analysis with Bioconductor. bioRxiv. https://doi.org/10.1101/590562

Conesa, A., P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera *et al.*, 2016 A survey of best practices for RNA-seq data analysis. Genome Biol. 17: 13. Erratum: 181. https://doi.org/10.1186/s13059-016-0881-8

Cornwell, M., M. Vangala, L. Taing, Z. Herbert, J. Köster *et al.*, 2018 VIPER: Visualization pipeline for RNA-seq, a snakemake workflow for efficient and complete RNA-seq analysis. BMC Bioinformatics 19: 135. https://doi.org/10.1186/s12859-018-2139-9

Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski *et al.*, 2013 STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29: 15–21. https://doi.org/10.1093/bioinformatics/bts635

Doumpas, N., F. Lampart, M. D. Robinson, A. Lentini, C. E. Nestor *et al.*, 2019 TCF/LEF dependent and independent transcriptional regulation of Wnt/β-catenin target genes. EMBO J. 38. Erratum: e98873. https://doi.org/10.15252/embj.201798873

Ewels, P., R. Hammarén, A. Peltzer, D. Moreno, rfenouil, *et al.*, 2018 nf-core/rnaseq: nf-core/rnaseq version 1.2.

Ewels, P., M. Magnusson, S. Lundin, and M. Käller, 2016 MultiQC: Summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32: 3047–3048. https://doi.org/10.1093/bioinformatics/btw354

Frankish, A., M. Diekhans, A.-M. Ferreira, R. Johnson, I. Jungreis *et al.*, 2019 GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. 47: D766–D773. https://doi.org/10.1093/nar/gky955

Girke, T., 2018 systemPipeRdata: NGS workflow templates and sample data. R package version 1.10.0.

He, W., S. Zhao, C. Zhang, M. S. Vincent, and B. Zhang, 2018 QuickRNASeq: guide for pipeline implementation and for interactive results visualization, pp. 57–70 in *Transcriptome Data Analysis: Methods and Protocols*, Springer New York, New York.

Huber, W., V. J. Carey, R. Gentleman, S. Anders, M. Carlson *et al.*, 2015 Orchestrating high-throughput genomic analysis with Bioconductor. Nat. Methods 12: 115–121. https://doi.org/10.1038/nmeth.3252

Hung, L.-H., J. Hu, T. Meiss, A. Ingersoll, W. Lloyd *et al.*, 2018 Building containerized workflows using the BioDepot-workflow-builder (bwb). bioRxiv. https://doi.org/10.1101/099010

Köster, J., and S. Rahmann, 2012 Snakemake–a scalable bioinformatics workflow engine. Bioinformatics 28: 2520–2522. https://doi.org/10.1093/bioinformatics/bts480

Love, M., R. Patro, P. Hickey, and C. Soneson, 2019 tximeta: Transcript Quantification Import with Automatic Metadata. R package version 1.1.16.

Lun, A. and D. Risso, 2019 SingleCellExperiment: S4 Classes for Single Cell Data. R package version 1.4.1.

Lun, A. T. L., Y. Chen, and G. K. Smyth, 2016 It's DE-licious: A recipe for differential expression analyses of RNA-seq experiments using Quasi-Likelihood methods in edgeR, pp. 391–416, *Statistical Genomics*, edited by E. Mathé and S. Davis, Methods in Molecular Biology, Springer New York, New York. https://doi.org/10.1007/978-1-4939-3578-9_19

Marini, F., 2018 ideal: Interactive Differential Expression AnaLysis. R package version 1.8.0.

Monier, B., A. McDermaid, C. Wang, J. Zhao, A. Miller *et al.*, 2019 IRIS-EDA: An integrated RNA-seq interpretation system for gene expression data analysis. PLoS Comput. Biol. 15: e1006792. https://doi.org/10.1371/journal.pcbi.1006792

Nowicka, M., and M. D. Robinson, 2016 DRIMSeq: a Dirichlet-multinomial framework for multivariate count outcomes in genomics. F1000 Res. 5: 1356. https://doi.org/10.12688/f1000research.8900.2

Patro, R., G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford, 2017 Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods 14: 417–419. https://doi.org/10.1038/nmeth.4197

Powell, D. R., 2018 https://drpowell.github.io/degust/.

Robinson, M. D., D. J. McCarthy, and G. K. Smyth, 2010 edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26: 139–140. https://doi.org/10.1093/bioinformatics/btp616

Rue-Albrecht, K., F. Marini, C. Soneson, and A. T. L. Lun, 2018 iSEE: Interactive SummarizedExperiment explorer. [version 1; referees: 3 approved] F1000 Res. 7: 741. https://doi.org/10.12688/f1000research.14966.1

Soneson, C., M. I. Love, and M. D. Robinson, 2015 Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000 Res. 4: 1521. https://doi.org/10.12688/f1000research.7563.1

Steinbaugh, M. J., L. Pantano, R. D. Kirchner, V. Barrera, B. A. Chapman *et al.*, 2018 bcbioRNASeq: R package for bcbio RNA-seq analysis. F1000 Res. 6: 1976.

Torre, D., A. Lachmann, and A. Ma'ayan, 2018 BioJupies: Automated generation of interactive notebooks for RNA-Seq data analysis in the cloud. Cell Syst. 7: 556–561.e3. https://doi.org/10.1016/j.cels.2018.10.007

Tsyganov, K., A. James Perry, S. Kenneth Archer, and D. Powell, 2018 RNAsik: A pipeline for complete and reproducible RNA-seq analysis that runs anywhere with speed and ease. JOSS 3: 583. https://doi.org/10.21105/joss.00583

Van Den Berge, K., K. Hembach, C. Soneson, S. Tiberi, L. Clement *et al.*, 2018 RNA sequencing data: hitchhiker's guide to expression analysis, PeerJ Preprints 6: e27283v2. https://doi.org/10.7287/peerj.preprints.27283v2

Wang, D., 2018 hppRNA-a snakemake-based handy parameter-free pipeline for RNA-Seq analysis of numerous samples. Brief. Bioinform. 19: 622–626.

Wu, D., and G. K. Smyth, 2012 Camera: a competitive gene set test accounting for inter-gene correlation. Nucleic Acids Res. 40: e133. https://doi.org/10.1093/nar/gks461

Xie, Y., J. Allaire, and G. Grolemund, 2018 *R Markdown: The Definitive Guide*, Chapman and Hall/CRC, Boca Raton, Florida. https://doi.org/10.1201/9781138359444

Zerbino, D. R., P. Achuthan, W. Akanni, M. R. Amode, D. Barrell *et al.*, 2018 Ensembl 2018. Nucleic Acids Res. 46: D754–D761. https://doi.org/10.1093/nar/gkx1098

Communicating editor: A. Whitehead

4 Mutant FUS triggers age-dependent synaptic impairment in presymptomatic ALS-FUS mice[1]

In this paper we studied the precise synaptic location of FUS with microscopy and the synaptic RNA targets by CLIP-seq. We used RNA-seq to analysed gene expression changes in an ALS mouse model with a mutation in the NLS of FUS at different time points. The project was a collaborative effort and I preprocessed and analysed the CLIP-seq and RNA-seq data. Sonu Sahadevan, Magdalini Polymenidou, Mark D. Robinson and I developed the strategy to filter synapse specific CLIP-seq peaks. Sonu Sahadevan, Elena Tantardini, Pierre de Rossi, Magdalini Polymenidou and I wrote the manuscript.

[1]Manuscript in preparation: Sahadevan, S., Hembach, K. M., Tantardini, E., Hruska-Plochan, M., Pérez-Berlanga, M., Weber, J., Schwarz, P., Dupuis, L., Robinson, M. D., De Rossi, P., & Polymenidou, M. "Mutant FUS triggers age-dependent synaptic impairment in presymptomatic ALS-FUS mice".

Synaptic accumulation of FUS triggers age-dependent misregulation of inhibitory synapses in ALS-FUS mice

Sonu Sahadevan[1*], Katharina M. Hembach[1,2*], Elena Tantardini[1], Manuela Pérez-Berlanga[1], Marian Hruska-Plochan[1], Julien Weber[1], Petra Schwarz[3], Luc Dupuis[4], Mark D. Robinson[2], Pierre De Rossi[1], Magdalini Polymenidou[1,#]

[1]Department of Quantitative Biomedicine, University of Zurich
[2]Department of Molecular Life Sciences and SIB Swiss Institute of Bioinformatics, University of Zurich
[3]Institute of Neuropathology, University Hospital Zurich
[4]Inserm, University of Strasbourg

*These authors contributed equally to this work

[#]Author for correspondence: magdalini.polymenidou@uzh.ch

Abstract

FUS is a primarily nuclear RNA-binding protein with important roles in RNA processing and transport. FUS mutations disrupting its nuclear localization characterize a subset of amyotrophic lateral sclerosis (ALS-FUS) patients, through an unidentified pathological mechanism. FUS regulates nuclear RNAs, but its role at the synapse is poorly understood. Here, we used super-resolution imaging to determine the physiological localization of extranuclear, neuronal FUS and found it predominantly near the vesicle reserve pool of presynaptic sites. Using CLIP-seq on synaptoneurosome preparations, we identified synaptic RNA targets of FUS that are associated with synapse organization and plasticity. Synaptic FUS was significantly increased in a knock-in mouse model of ALS-FUS, at presymptomatic stages, accompanied by alterations in density and size of GABAergic synapses. RNA-seq of synaptoneurosomes highlighted age-dependent dysregulation of glutamatergic and GABAergic synapses. Our study indicates that FUS accumulation at the synapse in early stages of ALS-FUS results in synaptic impairment, potentially representing an initial trigger of neurodegeneration.

Keywords: FUS, ALS-FUS, neurodegeneration, RNA-binding proteins, synaptic function, RNA transport, local translation, CLIP-seq, synaptoneurosomes, super-resolution microscopy

Introduction

FUS (Fused in sarcoma) is a nucleic acid binding protein involved in several processes of RNA metabolism[1]. Physiologically, FUS is predominantly localized to the nucleus[2] via active transport by transportin (TNPO)[3] and it can shuttle to the cytoplasm by passive diffusion[4,5]. In amyotrophic lateral sclerosis (ALS) and frontotemporal dementia (FTD), FUS mislocalizes to the cytoplasm where it forms insoluble aggregates[6–8]. In ALS, cytoplasmic mislocalization of FUS is associated with mutations that are mainly clustered in the proline-tyrosine nuclear localization signal (PY-NLS) at the C-terminal site of the protein[9] and lead to mislocalization of the protein to the cytosol. However, in FTD, FUS mislocalization occurs in the absence of mutations[10]. FUS is incorporated in cytoplasmic stress granules[5,11] and undergoes concentration-dependent, liquid-liquid phase separation[12,13], which is modulated by binding of TNPO and arginine methylation of FUS[14–17]. This likely contributes to the role of FUS in forming specific identities of ribonucleoprotein (RNP) granules[18,19] and in transporting RNA cargos[20], which is essential for local translation in neurons[21].

Despite the central role of FUS in neurodegenerative diseases, little is known about its function in specialized neuronal compartments, such as synapses. FUS was shown to mediate RNA transport[20] and is involved in stabilization of RNAs that encode proteins with important synaptic functions[22], such as *GluA1* and *SynGAP1*[23,24]. While the presence of FUS protein in synaptic compartments has been confirmed, its exact subsynaptic localization is debated. Diverging results described the presence of FUS at the pre-synapses in close proximity to synaptic vesicles[25–27], but also in dendritic spines[20] and in association with the postsynaptic density[28]. Confirming a functional role of FUS at the synaptic sites, behavioral and synaptic morphological changes have been observed upon depletion of FUS in mouse models[23,29,30]. Notably, mouse models associated with mislocalization of FUS exhibited reduced axonal translation contributing to synaptic impairments[31]. Synaptic dysfunction has been suggested to be the early event of several neurodegenerative disorders including ALS and FTD[32–36]. The disruption of RNA-binding proteins (RBPs) and RNA regulation could be a central cause of synaptic defects in these disorders.

Previous studies identified nuclear RNA targets of FUS with different cross-linking immunoprecipitation and high-throughput sequencing (CLIP-seq) approaches[22,37–41]. Collectively, these findings showed that FUS binds mainly introns, without a strong sequence specificity, but a preference for either GU-rich regions[22,38,40,41], which is mediated via its zinc finger (ZnF) domain, or a stem-loop RNA[37] via its RNA recognition motif[42]. FUS often binds close to alternatively spliced exons, highlighting its role in splicing regulation[22,38,39]. CLIP-seq studies also identified RNAs bound by FUS at their 3'

untranslated regions (3'UTRs) and exons[22,39,41], suggesting a direct role of FUS in RNA transport and regulating synaptic mRNA stability[23,24] and polyadenylation[40]. However, a precise list of synaptic RNAs directly regulated by FUS is yet to be identified.

In this study, we focused on understanding the role of synaptic FUS in RNA homeostasis and the consequences of ALS-causing mutations in FUS on synaptic maintenance. Using super-resolution imaging, we confirmed the presence of FUS at the synapse. FUS was found at both excitatory and inhibitory synapses, was enriched at the presynapse and rarely associated with postsynaptic structures. Synaptoneurosome preparations from adult mouse cortex, coupled with CLIP-seq uncovered specific synaptic RNA targets of FUS. Computational analyses revealed that most of these targets were associated with both glutamatergic and GABAergic networks. In a heterozygous knock-in FUS mouse model, which harbors a deletion in the NLS of FUS allele, thereby mimicking the majority of ALS-causing mutations[43], we found significant increase of synaptic FUS localization. To test the effect of this elevation in synaptic FUS, we investigated the synaptic organization of the hippocampus, which is enriched in glutamatergic and GABAergic synapses, and found mild and transient changes. However, RNA-seq analysis revealed age-dependent alterations of synaptic RNA composition including glutamatergic and GABAergic synapses. Our data indicate that early synaptic alterations in the GABAergic network precede motor impairments in these ALS-FUS mice[43], and may trigger early behavioral dysfunctions, such as hyperactivity and social disinhibition that these mice develop (Scekic-Zahirovic, Sanjuan-Ruiz et al., co-submitted manuscript).

Altogether, our results demonstrate a critical role for FUS in synaptic RNA homeostasis via direct association with specific synaptic RNAs, such as *Gabra1*, *Grin1* and others. Our study indicates that enhanced synaptic localization of FUS in early stages of ALS-FUS results in synaptic impairment, potentially representing the initial trigger of neurodegeneration. Importantly, we show that increased localization of FUS at the synapses, in the absence of aggregation, suffices to cause synaptic impairment.

Results

FUS is enriched at the presynaptic compartment of mature cortical and hippocampal neurons

While FUS has been shown at synaptic sites, its exact subsynaptic localization is debated. Some studies described a presynaptic enrichment of FUS in cortical neurons and motoneurons[25,27], whereas others have shown an association of FUS with postsynaptic density (PSD) sites[20,28]. To clarify the precise localization of FUS at the synapses, we first performed confocal analysis in mouse cortex (**Fig. 1a-b**) and hippocampus (**Supplementary**

Fig. 1a-b), which confirmed the presence of extranuclear FUS clusters along dendrites and axons (identified with MAP2 and PNF, respectively) and associated with synaptic markers (Synapsin1 and PSD95). To determine the precise subsynaptic localization of FUS, we used super-resolution microscopy (SRM) imaging of mouse hippocampal and cortical synapses. We first explored the distribution of FUS between excitatory and inhibitory synapses of cortical and hippocampal neurons (**Fig. 1c**). STED (Stimulated emission depletion) microscopy was used to precisely determine the localization of FUS clusters compared to synaptic markers: VGAT was used as a marker for inhibitory synapses and PSD95 for excitatory synapses. Image analysis was used to calculate the distance of the closest neighbor (**Supplementary Fig. 1c**). Only FUS clusters within 200 nm from a synaptic marker were considered for this analysis. Our results showed that extranuclear FUS preferentially associates with excitatory synapses, with 46% of the detected ones containing FUS, while only 20% of analyzed inhibitory synapses showing FUS positivity (t-test, p=0.0016) (**Fig. 1d**).

To better define the precise localization of FUS within the synapse, cortical and hippocampal primary cultures were immunolabeled for FUS along with pre- and postsynaptic markers (**Fig. 1e** and **Supplementary Fig. 1d-e**) and their relative distance was analyzed. At the presynapse, Synapsin 1 was used to label the vesicle reserve pool[44], and Bassoon to label the presynaptic active zone[45]. At the postsynaptic site, GluN2B, subunit of NMDA receptors, and GluA1, subunit of AMPA receptors, were used to label glutamatergic synapses. PSD95 was used to label the postsynaptic density zone[46]. Distribution of FUS at the synapse showed a closer association with Synapsin 1 compared to Bassoon, GluA1, BiP (ER marker) and GluN2B (**Supplementary Fig. 1f-g**). FUS also appeared to be closer to Bassoon compared to PSD95 (**Supplementary Fig. 1f-g**). A subset of FUS was also localized at the spine (**Fig. 1e**). To strengthen our analyses and to refine the precise localization of FUS, the relative proportion of FUS within 100 nm was compared for each marker. Our results showed a preferential FUS localization at the presynaptic site (**Fig. 1f**) (t-test, p=0.0006), in accordance with previously reported data[25,27]. Within the presynaptic site (**Fig. 1g**), FUS was significantly enriched in the Synapsin-positive area (One-way ANOVA, p<0.0001, posthoc Tukey, Syn1 vs. PSD95, p<0.0001; Syn1 vs. GluN2B, p=0.0157; Syn1 vs. GluA1, p=0.454; Syn1 vs. Bassoon, p=0.0005). However, no significant difference was found with the ER marker, suggesting that FUS could be localized between Synapsin 1 and ER at the presynapse (**Fig. 1h**). These results are in line with the previously published localization of FUS within 150 nm from the active presynaptic zone[27], but highlight the presence of FUS also at the postsynaptic site, potentially explaining the apparently contradictory results of previous studies[20,28].

Identification of synaptic RNA targets of FUS

The role of FUS in the nucleus has been well studied and previously published CLIP-seq data identified FUS binding preferentially on pre-mRNA, suggesting that these binding events occur in the nucleus[22,47–50]. Given the confirmed synaptic localization of FUS (**Fig. 1**), we wondered if a specific subset of synaptic RNAs are directly bound and regulated by FUS in these compartments. Since synapses contain few copies of different RNAs and only a small fraction of the total cellular FUS is synaptically localized, RNAs specifically bound by FUS at the synapses are likely missed in CLIP-seq datasets from total brain. Therefore, we biochemically isolated synaptoneurosomes that are enriched synaptic fractions from mouse cortex to identify synapse-specific RNA targets of FUS. Electron microscopy analysis confirmed the morphological integrity of our synaptoneurosome preparations, which contained intact pre- and postsynaptic structures (**Fig. 2a**). Immunoblot showed an enrichment of synaptic markers (PSD-95, p-CAMKII, GluN2B, GluA1, SNAP25, NXRN1), absence of nuclear proteins (Lamin B1, Histone H3) and presence of FUS in the synaptoneurosomes (**Fig. 2b** and **Supplementary 2a**). In addition, quantitative reverse transcription polymerase chain reaction (qRT-PCR) analysis showed enrichment of selected synaptic mRNAs (**Fig. 2c**).

Following a previously published method[22,51], we used ultraviolet (UV) crosslinking on isolated synaptoneurosomes and total cortex from 1-month-old wild type mice to stabilize FUS-RNA interactions and to allow stringent immunoprecipitation of the complexes (**Supplementary Fig. 2b**). As FUS is enriched in the nucleus and only a small fraction of the protein is localized at the synapses, we prepared synaptoneurosomes from cortices of 200 mice to achieve sufficient RNA levels for CLIP-seq library preparation. The autoradiograph showed an RNA smear at the expected molecular weight of a single FUS molecule (70 kDa) and lower mobility complexes (above 115 kDa) that may correspond to RNAs bound by more than one FUS molecule or a heterogeneous protein complex (**Fig. 2d**). No complexes were immunoprecipitated in the absence of UV cross-linking or when using nonspecific IgG-coated beads. The efficiency of immunoprecipitation was confirmed by depletion of FUS in post-IP samples (**Supplementary Fig. 2c**). Finally, RNAs purified from the FUS-RNA complexes of cortical synaptoneurosomes and total cortex were sequenced and analyzed. We obtained 29,057,026 and 27,734,233 reads for the total cortex and cortical synaptoneurosome samples, respectively. 91% of the total cortex and 66% of the synaptoneurosome reads could be mapped to a unique location in the mouse reference genome (GRCm38) (**Supplementary Fig. 2d**). After removing PCR duplicates, we identified peaks using a previously published tool called CLIPper[52], resulting in 619,728 total cortex and 408,918 synaptoneurosome peaks.

Before comparing the peaks in the two samples, we normalized the data to correct for

different sequencing depths and signal-to-noise ratios[53] (see Methods). This is especially important in our case, because the synaptoneurosome sample should contain only a subset of the FUS targets from total cortex. We wanted to filter the predicted peaks of the synaptoneurosome sample to identify genomic regions with high log2 fold-change between the synaptoneurosome and total cortex samples. Peaks with low number of reads (or no reads) in the total cortex, but high read coverage in the synaptoneurosomes correspond to regions that are putatively bound by FUS in the synapse. However, the observable number of reads per RNA in each sample strongly depends on gene expression and the number of localized RNA copies. Therefore, we did not want to use a simple read count threshold to filter and identify synapse specific peaks. Instead, we fit a count model and computed peak-specific p-values to test for differences between the synaptoneurosome and total cortex CLIP-seq enrichment (**Fig. 2e**). The normalization highlights the expected association between p-values (yellow) and log2 CPM (**Fig. 2e**).

We ranked the peaks by p-values and used a stringent cutoff of 1e-5 (**Fig. 2e**) to ensure enrichment of synaptic FUS targets. Indeed, the resulting peaks were largely devoid of intronic regions, but were enriched in exons and 3'UTRs, as was expected for synaptic FUS targets, which are mature and fully processed RNAs (**Fig. 2e** and **Supplementary Fig. 2g**). The same normalization and filtering of CLIPper peaks identified in the total cortex highlighted RNAs primarily bound by FUS in the nucleus, where the vast majority of FUS protein resides (**Supplementary Fig. 2e**). After selecting an equal number of top peaks as obtained for the synaptoneurosome sample (1560 peaks in 517 genes), corresponding to a p-value cutoff of 0.0029 (**Supplementary Fig. 2f**), we confirmed the previously reported[22] preferential binding of FUS within intronic regions of pre-mRNAs (**Fig. 2g** and **Supplementary Fig. 2h**).

The final list of synapse-specific FUS binding sites consists of 1560 peaks in 307 RNAs (**Supplementary Table 1**), primarily localized to exons and 3'UTRs of RNAs specific to the synapses. Among those, FUS peaks on the exon of *Grin1* (Glutamate ionotropic NMDA type subunit 1) and 3'UTR of a long isoform of *Gabra1* (Gamma aminobutyric acid receptor subunit alpha-1) were exclusively detected in synaptoneurosomes, but not in total cortex (**Fig. 2h-i**). Direct binding of FUS to 3'UTR and exonic regions of its targets suggests a potential role in regulating RNA transport, local translation and/or stabilization.

Synaptic FUS RNA targets encode essential protein components of synapse

We then wondered if the 307 synaptic FUS target RNAs were collectively highlighting any known cellular localization and function. Most RNAs are localized to either the pre- or postsynapse or they are known astrocytic markers (**Fig. 2j**). Among those are RNAs encoding essential protein members of glutamatergic (*Grin1*, *Gria2*, *Gria3*) and GABAergic

synapses (*Gabra1*, *Gabrb3*, *Gabbr1*, *Gabbr2*), transporters, as well as components of the calcium signaling pathway, which are important for plasticity of glutamatergic synapses. An overrepresentation analysis (ORA) comparing the synaptic FUS targets to all synaptic RNAs detected in cortical mouse synaptoneurosomes by RNA-seq (logCPM >1, 1-month-old mice), revealed that FUS targets were enriched for synaptic - both pre- and postsynaptic - localization. Synaptic FUS target RNAs were enriched for gene ontology categories, such as transport, localization and trans-synaptic signaling, as well as signaling receptor binding and transmembrane transporter activity (**Supplementary Fig. 2i**).

Here we identified for the first time specific synaptic RNA targets directly bound by FUS, including those associated with glutamatergic and GABAergic networks. Our data suggests that FUS plays a critical role in maintaining synaptic integrity and organization.

FUS binds GU-rich sequences at the synapse

While FUS has been shown to be a relatively promiscuous RNA-binding protein, preference towards GU-rich motifs has been reported in previous CLIP-seq studies[22,38,40,41], a binding mediated via its ZnF domain[42]. To understand if FUS binding to synaptic RNA targets follows the same modalities as its nuclear targets, we explored the sequence specificity of FUS in the synapse and predicted motifs with HOMER[54], comparing the FUS peak sequences of cortical synaptoneurosomes and total cortex samples. In accordance with previous studies, we found a degenerate GU-rich motif for intronic FUS binding sites in the total cortex (**Table 1**). The sequences of the synaptic FUS peaks in exons and 5' UTRs revealed a "AGGUAAGU" motif which was only found in 11% and 6% of the peaks, respectively. We conclude that FUS does not have a stronger sequence preference in the synapse than in the nucleus.

Increased synaptic localization of mutant FUS protein in *Fus*$^{\Delta NLS/+}$ mice

In order to explore synaptic impairments associated with FUS mislocalization, we used the *Fus*$^{\Delta NLS/+}$ mouse model[55]. This mouse model shows partial cytoplasmic mislocalization of FUS due to a lack of the nuclear localization (NLS) in one copy of the FUS allele, closely mimicking ALS-causing mutations reported in patients. Taking advantage of two antibodies that recognize either total FUS (both full length and mutant) or only the full length protein (**Fig. 3a**), we assessed FUS protein levels in synaptoneurosomes isolated from *Fus*$^{\Delta NLS/+}$ mice and wild type (*Fus*$^{+/+}$) of 1 and 6 months of age. We detected higher levels of total FUS in synaptoneurosomes from *Fus*$^{\Delta NLS/+}$ at both ages compared to *Fus*$^{+/+}$ (**Fig. 3b-c, Supplementary Fig. 3a-b**). However, full length FUS levels were decreased in synaptoneurosomes of *Fus*$^{\Delta NLS/+}$ compared to *Fus*$^{+/+}$ indicating that the truncated FUS protein is misaccumulated at the synaptic sites of *Fus*$^{\Delta NLS/+}$ mice.

Confirming our biochemical evidence, immunofluorescence analyses of *Fus*$^{\Delta NLS/+}$ mice showed higher levels of FUS in dendritic compartments of CA1 pyramidal cells. *Fus*$^{+/+}$ mice at both 1 month (**Supplementary Fig. 3c-d**) and 6 months of age (**Fig. 3d-e**) showed prominent expression of FUS in the nucleus. High magnification images highlighted the presence of FUS at the synapses, identified by co-labeling with Synapsin1. *Fus*$^{\Delta NLS/+}$ mice at 1 (**Supplementary Fig. 3c-d**) and 6 months of age (**Fig. 3d-e**) showed higher levels of FUS within the dendritic tree (identified with MAP2) and at the synapse compared to *Fus*$^{+/+}$ mice, confirming our previous quantifications by immunoblot.

Dysregulation of inhibitory synapses in *Fus*$^{\Delta NLS/+}$ mouse model

To explore a possible synaptic disorganization associated with mislocalization of FUS, we performed synaptic density and size analyses. Based on evidence that the hippocampal/prefrontal cortex connectome participates in memory encoding and recalling[56] and that CA1 hippocampal excitatory and inhibitory synapses are highly similar to the cortical synapses[57–60], we explored the possible synaptic changes triggered by FUS mislocalization in the CA1 hippocampal region. We analyzed both *Fus*$^{+/+}$ and *Fus*$^{\Delta NLS/+}$ mice, using presynaptic and postsynaptic markers. Density and area analyses were performed as shown in **Supplementary Fig. 3e**. At the presynapse, we quantified the density of the SNARE associated protein SNAP25[61] (synaptic RNA target of FUS) and the presynaptic active zone marker Bassoon[45]. The density of inhibitory synapses was assessed using VGAT[62] (presynaptic). At the postsynapse, we quantified the density of postsynaptic glutamatergic receptor GluN1[63] (synaptic RNA target of FUS and obligatory subunit of all NMDAR) and GluA1[64] (obligatory subunit of AMPAR), as well as postsynaptic GABAergic receptors containing α1 subunit ($GABA_A$α1; synaptic RNA target of FUS) and α3 ($GABA_A$α3)[65]. We also assessed the number of active excitatory synapses using phospho-CaMKII (pCaMKII) as well as functional inhibitory synapses using Gephyrin[66].

At 1 month of age in *Fus*$^{\Delta NLS/+}$ mice, we did not observe significant changes at the presynaptic site, suggesting a normal axonal and axon terminal development and functions. However, at the postsynaptic sites, we observed a significant increase of NMDAR (p=0.0219) and a significant decrease of $GABA_A$α3 receptors (p=0.0156) (**Fig. 3f-g, Supplementary Fig. 3f** and **Table 2**). Moreover at 1 month of age, *Fus*$^{\Delta NLS/+}$ mice showed significantly more NMDAR located at the extrasynaptic site (p=0.0433) (**Fig. 3h**). Interestingly, the size of the $GABA_A$α3 clusters was significantly decreased in *Fus*$^{\Delta NLS/+}$ mice (p=0.0053) at 1 month of age (**Fig. 3f, i, Supplementary Fig. 3h** and **Table 3**). We did not record changes in the number of Synapsin1, Bassoon, SNAP25, VGAT, GluA1, $GABA_A$α1, Gephyrin or pCaMKII, suggesting either an increase of silent synapses, immature synapses

or an increase of the number of NMDAR in the dendritic shaft together with a decrease of $GABA_A\alpha3$ synaptic clustering. These results suggested a hyperexcitability profile during developmental stages.

At 6 months of age, we did not observe significant changes in the density of pre or postsynaptic markers (**Fig. 3f-g** a**nd Supplementary Fig. 3g**), suggesting a normal maturation of the synaptic network despite developmental synaptic dysregulation described above. However, SNAP25 (p=0.085) and VGAT (p=0.0792) trended towards an increased density, suggesting a potential alteration at inhibitory presynaptic sites (**Supplementary Fig. 3g and Table 2**). This interpretation was confirmed by an increase of the area of the presynaptic marker VGAT (p=0.0028) and of the size of $GABA_A\alpha3$ clusters at the postsynaptic site (p=0.0166) (**Fig. 3i, Supplementary Fig. 3i** and **Table 3**), while GluN1 clusters appeared unaffected. Increase in VGAT suggested an elevated number of presynaptic GABAergic vesicles, which was confirmed by EM analyses in older mice (Scekic-Zahirovic, Sanjuan-Ruiz et al., co-submitted manuscript). Correlatively, increase of $GABA_A\alpha3$ cluster size suggested an increase in the trafficking of $GABA_AR$ at the postsynaptic site. This occurred, however, without an increase of the anchoring protein Gephyrin, suggesting instable structure of the inhibitory postsynaptic sites. Altogether, our results show alterations of both glutamatergic and GABAergic synapses during developmental synaptogenesis (1 month of age), while only GABAergic synapses appeared affected at a later time point (6 months of age). This suggests a potential role for FUS in synaptogenesis and network wiring and synaptic maintenance, with a selective exacerbation of inhibitory synaptic defects with age.

Fus$^{\Delta NLS/+}$ mice show age-dependent synaptic RNA alterations

FUS plays an essential role in RNA stabilization[23,24] and transport[20]. Therefore, we used RNA-seq to investigate the consequences of increased synaptic levels of mutated FUS in *Fus*$^{\Delta NLS/+}$ mice (**Fig. 4a**). We isolated RNA from six biological replicates of synaptoneurosomes and paired total cortex samples from *Fus*$^{+/+}$ and *Fus*$^{\Delta NLS/+}$ mice at 1 and 6 months of age and prepared poly-A-selected libraries for high-throughput sequencing. As a control, we also sequenced the nuclear fraction from 4 biological replicates of *Fus*$^{+/+}$ mice at 1 month of age. For quality control, we computed principal components of all samples and all expressed genes (see methods for details) and found a clustering by sample condition and age (**Supplementary Fig. 4b-c).**

We compared the expressed genes in our synaptoneurosomes (15087 genes) with the forebrain synaptic transcriptome[67] (14073 genes) and the vast majority of detected RNAs (13475) were identical between the two studies (**Supplementary Fig. 4a**). The small

differences in the two transcriptomes can be explained by differences in the used synaptoneurosome protocols and the brain region (frontal cortex versus forebrain).

We conducted four differential gene expression analyses, comparing *Fus*$^{\Delta NLS/+}$ to *Fus*$^{+/+}$ replicates separately for the total cortex and synaptoneurosomes at both time points (for full lists see **Supplementary Tables 2-5**). A false discovery rate (FDR) cutoff of 0.05 was used to define significant differential expression. Only three and five RNAs were differentially expressed (DE) in the *Fus*$^{\Delta NLS/+}$ samples of the total cortex at 1 and 6 months of age, respectively (**Supplementary Fig. 4f and Supplementary Tables 2-3**). However, in the synaptoneurosomes, we identified 11 and 594 RNAs differentially abundant at 1 and 6 months, respectively (**Supplementary Tables 4-5**). 136 RNAs were decreased and 485 RNAs were increased in the *Fus*$^{\Delta NLS/+}$ mice at 6 months of age compared to synaptoneurosomes from *Fus*$^{+/+}$ mice (**Fig. 4b**). The significantly increased RNAs in *Fus*$^{\Delta NLS/+}$ mice at 6 months were enriched in gene ontology (GO) categories such as synaptic signaling, intrinsic component of membrane and transporter activity (**Supplementary Fig. 4d**), while those that were decreased in abundance were associated with cytoskeletal organization and RNA metabolism (**Supplementary Fig. 4e**).

At 6 months of age, the log2 fold changes of the altered RNAs are consistently negative or positive in all *Fus*$^{\Delta NLS/+}$ synaptoneurosome replicates (**Fig. 4c**). At 1 month of age, the log2 fold changes of the *Fus*$^{\Delta NLS/+}$ synaptoneurosome replicates are mostly neutral (white color on the heatmap) indicating that alterations in RNA abundance are age-dependent and not detectable as early as 1 month of age. In the total cortical samples at 6 months of age, some of the replicates show a similar trend as the synaptoneurosome samples, but it seems that the effects cannot be detected because synaptic RNAs are too diluted (**Supplementary Fig. 4g**). Overall, we found synapse-specific differential RNA abundance at 6 months in the *Fus*$^{\Delta NLS/+}$ mice, but not in the total cortex.

While most of the 594 differentially abundant RNAs (**Supplementary Table 5**) were not direct FUS targets, 33 altered RNAs are synaptic targets of FUS. The altered synaptic transcriptome, along with the impaired expression of a subset of FUS RNA targets in *Fus*$^{\Delta NLS/+}$ mice, suggests direct and indirect effects of mutant FUS at the synapses (**Fig. 4d**). FUS targets with known synaptic functions that are altered in *Fus*$^{\Delta NLS/+}$ are represented in **Fig. 4e**. Most of those RNAs show exonic FUS binding on our CLIP-seq analysis (**Supplementary Fig. 5-6, Supplementary Table 1**), with the exception of *Gria 3*, *Spock1, Spock2* (**Supplementary Fig. 6b, f-g**) and *Gabra1* (**Supplementary Fig. 7**), which are bound by FUS at their 3'UTR. Altered FUS targets include RNAs encoding presynaptic vesicle associated proteins, transsynaptic proteins, membrane proteins, receptors associated with glutamatergic and GABAergic pathways. Our results suggest that mislocalization of FUS leads to mild alterations in the synaptic RNA profile that may affect

synaptic signaling and plasticity. Our data indicate that synaptic RNA alterations may occur at an asymptomatic age and represent one of the early events in disease pathogenesis.

Discussion

In this study, we identified for the first-time synaptic RNA targets of FUS combining cortical synaptoneurosome preparations with CLIP-seq. Additionally, synaptic RNA levels were found to be altered in a *Fus*$^{\Delta NLS/+}$ mouse model at 6 months of age. Along with these results, we assessed FUS localization at the synaptic site using a combination of super-resolution microscopy approaches. Altogether, our results point to a critical role for FUS at the synapse and indicate that increased synaptic FUS localization at presymptomatic stages of ALS-FUS mice triggers early alterations of synaptic RNA content and misregulation of the GABAergic network. These early synaptic changes mechanistically explain the behavioral dysfunctions that these mice develop (Scekic-Zahirovic, Sanjuan-Ruiz et al., co-submitted manuscript).

RNA transport and local translation ensure fast responses with locally synthesized proteins essential for plasticity[21,68,69]. CLIP-seq using synaptoneurosome preparations from mouse cortex demonstrated that FUS not only binds nuclear RNAs, but also those that are localized at the synapses. Both pre- and postsynaptic localization of the identified targets correlated with the subcellular localization of FUS in both synaptic compartments. Moreover, by CLIP-seq on synaptoneurosomes, we identified that FUS binds RNAs encoding GABA receptor subunits (*Gabra1, Gabrb3, Gabbr1, Gabbr2*) and glutamatergic receptors (*Gria2, Gria3, Grin1*) previously known to be localized at dendritic neuropils[70]. FUS binding on synaptic RNAs is enriched on 3'UTRs and/or exonic regions, as revealed by our synaptoneurosome CLIP-seq dataset, suggesting that FUS might play a role in regulating local translation or transport of these targets.

Synaptic analyses at presymptomatic ages of *Fus*$^{\Delta NLS/+}$ mice revealed interesting changes. Our results showed a major effect on inhibitory synapses at 1 and 6 months of age. We explored $GABA_AR$ density and found changes in α3-containing $GABA_AR$. $GABA_A\alpha3$ is expressed at the postsynaptic site of monoaminergic synapses[71], and have been shown to be involved in fear and anxiety behavior, and mutations in the *Gabra3* subunit resulted in an absence of inhibition behavior[72–74]. Changes in $GABA_A\alpha3$ and not $GABA_A\alpha1$-containing receptor suggested that only monoaminergic neurons were affected in the *Fus*$^{\Delta NLS/+}$ mouse model. These results are well aligned with a contemporaneous study (Scekic-Zahirovic, Sanjuan-Ruiz et al., co-submitted manuscript), which showed specific behavioral changes that can be linked to monoaminergic networks. Interestingly at 1 month of age, *Fus*$^{\Delta NLS/+}$ mice showed an increase of NMDAR associated with a decrease in $GABA_A\alpha3$. These results suggested a role for FUS during synaptogenesis in regulating postsynaptic receptor

composition as previously suggested[23,28,75]. In 1-month-old *Fus*$^{\Delta NLS/+}$ mice, NMDARs were enriched at the extrasynaptic sites, which, together with the decrease in $GABA_A\alpha3$, suggested an hyperexcitability profile during development. We hypothesize that abnormal activity during developmental stages could result in abnormal network connection. *Fus*$^{\Delta NLS/+}$ mice at 6 months of age showed higher density of presynaptic inhibitory boutons, pointing toward a compensatory mechanism at the GABAergic synapses to overcome the hyperexcitability profile observed during development. Moreover at 6 months of age, *Fus*$^{\Delta NLS/+}$ mice also displayed higher density of SNAP25, present at both inhibitory and excitatory synapses[61,76], but we did not explore if this increase was specific for the GABAergic network.

Interestingly, the cluster size of VGAT, which is involved in the transport of GABA in the presynaptic vesicles[77], was increased in *Fus*$^{\Delta NLS/+}$ mice at 6 months of age. Increase of the cluster size would suggest that either more vesicles were present at the presynapse, or an increase of VGAT protein per vesicle. We also observed an increase in $GABA_A\alpha3$ cluster size and their density in 6-month-old *Fus*$^{\Delta NLS/+}$ mice. Surprisingly, we did not observe an increase in Gephyrin, a postsynaptic protein responsible for anchoring GABAR at the postsynaptic site[78,79]. Gephyrin interacts at the postsynaptic site with GABAR at a ratio 1:1[80], suggesting that inhibitory synapses in the *Fus*$^{\Delta NLS/+}$ model were unstable at 6 months of age with an excess of GABAR poorly anchored at the postsynaptic site, which could lead to malfunction of the inhibitory network. In correlation, *Fus*$^{\Delta NLS/+}$ mice showed behavioral changes overtime with disinhibition and hyperactivity behaviors as early as 4 months of age, associated with a decrease in the number of inhibitory neurons at 22-month-old (Scekic-Zahirovic, Sanjuan-Ruiz et al., co-submitted manuscript). Altogether, these results suggest that increased level of extranuclear FUS during development led to abnormal synaptogenesis affecting the GABAergic system over time.

Using the *Fus*$^{\Delta NLS/+}$ mouse model, we found that accumulation of mislocalized mutant FUS at the synapses altered the synaptic RNA content as early at 6 months of age. These alterations include FUS target RNAs that are associated with glutamatergic (*Grin1, Gria2, Gria3*) and GABAergic (*Gabra1*) synapses. These targets were found with increased synaptic localization in *Fus*$^{\Delta NLS/+}$. An impairment of genes associated with the GABAergic network in the frontal cortex of both young (5-month-old) and old (22-month-old) *Fus*$^{\Delta NLS/+}$ mice has been shown by an independent study (Scekic-Zahirovic, Sanjuan-Ruiz et al., co-submitted manuscript). Importantly, this ALS-FUS mouse model developed behavioral deficits, including hyperactivity and social disinhibition, suggesting defects in cortical inhibition. Our data supports that phenotypic manifestations in *Fus*$^{\Delta NLS/+}$ mice could be due to synaptic RNA alterations caused by mutant FUS at synapses. Moreover, mutant FUS-associated synaptic RNA alterations precede in ALS-FUS mice as suggested in our data.

However, the precise mechanism of how FUS regulates these targets is yet to be determined.

CLIP-seq from synaptoneurosomes showed that FUS binds selectively to specific GABA receptor subunits encoding mRNAs: *Gabra1, Gabrb3, Gabbr1, Gabbr2*. Other RNA-binding proteins, such as fragile X mental retardation protein (FMRP), Pumilio 1, 2 and cytoplasmic polyadenylation binding element binding protein (CPEB) have also been shown to bind GABAR subunit mRNAs by CLIP-seq[81]. Whether all these proteins act in concert to locally regulate the expression of GABAR subunits at synapses needs to be investigated. Interestingly, FUS interacts with FMRP, a well-studied protein known to regulate local translation[82]. Long 3' UTRs have been suggested to promote increased binding of RBPs and miRNAs which control the translation of these mRNAs[83]. Our CLIP-seq from synaptoneurosomes showed that FUS binds to the long 3' UTR containing isoform of *Gabra1* (**Supplementary Fig. 7**) indicating that FUS may be directly involved in regulating the protein expression of *Gabra1* at the synapses. Furthermore, we found increased levels of *Gabra1* mRNA in synaptoneurosome preparations from $Fus^{\Delta NLS/+}$ mice. It is important to study whether elevated levels of FUS at the synapse may directly impact *Gabra1* levels via mRNA stabilization or local translation leading to altered regulation of inhibitory network. Overall, our findings highlight the role of FUS in synaptic RNA homeostasis possibly through regulating RNA transport, RNA stabilization and local translation.

Materials and Methods

Experimental models

Mice housing and breeding were in accordance with the Swiss Animal Welfare Law and in compliance with the regulations of the Cantonal Veterinary Office, Zurich. We used 1- to 6-month-old C57/Bl6 mice or $Fus^{+/+}$/$Fus^{\Delta NLS/+}$ mice with genetic background (C57/Bl6). Wild type and heterozygous $Fus^{\Delta NLS/+}$ mice with genetic background (C57/Bl6)[55] were bred and housed in the animal facility of the University of Zurich.

Immunofluorescence staining for brain sections

Mice were anesthetized by CO_2 inhalation before perfusion with PBS containing 4% paraformaldehyde and 4% sucrose. Brains were harvested and post-fixed overnight in the same fixative and then stored at 4°C in PBS containing 30% sucrose. Sixty m-thick coronal sections were cut on a cryostat and processed for free-floating immunofluorescence staining. Brain sections were incubated with the indicated primary antibodies for 48 h at 4°C followed by secondary antibodies for 24h at 4°C. The antibodies were diluted in 1X Tris Buffer Saline solution containing 10% donkey serum, 3% BSA, and 0.25% Triton-X100. Sections were then mounted on slides with Prolong Diamond (Life Technologies) before confocal microscopy.

STED super-resolution imaging and analysis

Super-resolution STED (Stimulated emission depletion microscopy) images of FUS and synaptic markers were acquired on a Leica SP8 3D, 3-color gated STED laser scanning confocal microscope. Images were acquired in the retrospenial cortical area in the layer 5 and in the molecular layer of the hippocampal CA1 area. A 775 nm depletion laser was used to deplete both 647 and 594 dyes. The powers used for depletion lasers, the excitation laser parameters, and the gating parameters necessary to obtain STED resolution were assessed for each marker. 1 m-thick Z-stacks of 1024 X 1024-pixel images at 40 nm step size were acquired at 1800 kHz bidirectional scan rate with a line averaging of 32 and 3 frame accumulation, using a 100X (1.45) objective with a digital zoom factor of 7.5, yielding 15.15 nm pixels resolution.

STED microscopy data were quantified from at least 2 image stacks acquired from 2 *Fus*$^{+/+}$ adult mice. The STED images were deconvolved using Huygens Professional software (Scientific Volume Imaging). Images were subsequently analyzed using Imaris software. Volumes for each marker were generated using smooth surfaces with details set up at 0.01 m. The diameter of the largest sphere was set up at 1 m. Threshold background subtraction methods were used to create the surface, and the threshold was calculated for each marker and kept constant. Surfaces were then filtered by setting up the number of voxels >10 and <2000 pixels. Closest neighbor distance was calculated using integrated distance transformation tool in Imaris. Distances were then organized and statistically analyzed using mean comparison and t-test comparison. Distances greater than 200 nm were removed from the analysis, and average distance were analyzed.

Neuronal primary cultures

Primary neuronal cell cultures were prepared from postnatal (P0) pups. Briefly, hippocampus and cortex were isolated. Hippocampi were treated with trypsin (0.5% w/v) in HBSS-Glucose (D-Glucose, 0.65 mg/ml) and triturated with glass pipettes to dissociate tissue in Neurobasal medium (NB) supplemented with glutamine (2 mM), 2% B27, 2.5% Horse Serum, 100U penicillin-streptomycin and D-Glucose (0.65 mg/ml). Hippocampal cells were then plated onto poly-D-lysine coated 18x18 mm coverslips (REF) at 6 x 10^4 cells/cm² for imaging, and for biochemistry at high density (8 x 10^4 cells/cm^2). Cells were subsequently cultured in supplemented Neurobasal (NB) medium at 37°C under 5% CO_2, one-half of the medium changed every 5 days, and used after 15 days in vitro (DIV). Cortex were dissociated and plated similarly to hippocampal cells in NB supplemented with 2% B27, 5% horse serum, 1% N2, 1% glutamax, 100U penicillin-streptomycin and D-Glucose (0.65 mg/ml).

Direct Stochastic Optical Reconstruction Microscopy (dSTORM)

Super-resolution images were acquired on a Leica SR Ground State Depletion 3D / 3 color TIRFM microscope with an Andor iXon Ultra 897 EMCCD camera (Andor Technology PLC). DIV15-18 mouse primary neurons were fixed for 20 min in 4% PFA - 4% sucrose in PBS. Primary antibodies were incubated overnight at 4% in PBS containing 10% donkey serum, 3% BSA, and 0.25% Triton X-100. Secondary antibodies were incubated at RT for 3 h in the same buffer. After 3 washes in PBS, the cells were re-fixed with 4%PFA for 5 min. The coverslips were then washed over a period of 2 days at 4°C in PBS to remove non-specific binding of the secondary antibodies. Coverslips were mounted temporarily in an oxygen scavenger buffer (200mM phosphate buffer, 40% glucose, 1M cysteamine hydrochloride (M6500 Sigma), 0.5mg/mL Glucose-oxydase, 40ug/mL Catalase) to limit oxidation of the fluorophores during image acquisition. The areas of capture were blindly selected by direct observation in DIC. Images were acquired using a 160X (NA 1.43) objective in the TIRF mode North direction with a penetration of 200 nm. Far red channels (Alexa 647 or 660) were acquired using a 642 nm laser. Red channels (Alexa 568 or 555) were acquired using a 532 nm laser. Green channel (Alexa 488) was acquired using 488 nm laser. Images were acquired in 2D. The irradiation intensity was adjusted until the single molecule detection reached a frame correlation <0.25. Detection particle threshold was defined between 20-60 depending on the marker and adjusted to obtain a number of events per frame between 0 and 25. The exposure was maintained at 7.07 ms and the EM gain was set at 300. The power of depletion and acquisition was defined for each marker and kept constant during acquisition. The number of particles collected were maintained constant per markers and between experiments. At least 3 independent cultures or coverslips were imaged per marker.

Super-resolution image processing and analysis

Raw GSD images were processed using a custom-made macro in Fiji to remove background by subtraction of a running median of frames (300 renewed every 300 frames) and subtracting the previously processed image once background was removed[84]. A blur (0.7-pixel radius) per slice prior to median subtraction was applied to reduce the noise further. These images were then processed using Thunderstorm plugin in Imagej. Image filtering was performed using Wavelet filter (B-spline, order 3/scale2.0). The molecules were localized using centroid of connected components, and the peak intensity threshold was determined per marker/dye to maintain an XY uncertainty <50. Sub-pixel localization of molecules was performed using PSF elliptical gaussian and least squared fitting methods with a fitting radius of 5 pixels and initial sigma of 1.6 pixels. Images were analyzed using Bitplane Imaris software v.9.3.0 (Andor Technology PLC). Volumes for each marker were generated using smooth surfaces with details set up at 0.005. The diameter of the largest sphere was set up at 1 m. A threshold background subtraction method was used to create the surface and threshold was calculated and applied to all the images of the same experiment. Surfaces were then filtered by setting up the area between 0.01-1 m^2. The closest neighbor distance was processed using the integrated distance transformation tool in Imaris. Distances were then organized and statistically analyzed using median comparison and ANOVA and Fisher's Least Significant Difference (LSD) test. Distances greater than 100 nm were removed from the analysis, and average distance were analyzed.

Preparation of synaptoneurosomes from mouse brain tissues

Synaptoneurosomes were prepared based on previously published protocols[85,86] with slight modifications. The freshly harvested cortex tissue homogenized using dounce homogenizer for 12 strokes at 4°C in buffer (10%w/v) containing pH 7.4, 10 mM 4-(2 hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES; Biosolve 08042359), 0.35 M Sucrose, 1 mM ethylenediaminetetraacetic acid (EDTA; VWR 0105), 0.25 mM dithiothreitol (Thermo Fisher Scientific R0861), 30 U/ml RNAse inhibitor (Life Technologies N8080119) and complete-EDTA free protease inhibitor cocktail (Roche 11836170001, PhosSTOP (Roche 04906845001). 200ul of the total homogenate were saved for RNA extraction or western blot analysis. The remaining homogenate was spun at 1000g, 15 min at 4°C to remove the nuclear and cell debris. The supernatant was sequentially passed through three 100 µm nylon net filters (Millipore NY1H02500), followed by one 5 m filter (Millipore SMWP013000). The filtrate was resuspended in 3 volumes of SNS buffer without sucrose and spun at 2000g, 15 min at 4°C to collect the pellet containing synaptoneurosomes. The pellets were resuspended in RIPA buffer for western blot or in qiazol reagent for RNA extraction.

Cross-Linking Immunoprecipitation and high-throughput sequencing (CLIP-seq)

Total lysate and synaptoneurosomes isolated from cortex tissue of 1-month-old C57Bl/6 mice were UV crosslinked (100 mJ/cm^2 for 2 cycles) using UV Stratalinker 2400 (Stratagene) and stored at -80°C until use. For the total sample, cortex tissue was dissociated using a cell strainer of pore size 100 m before crosslinking. We used cortex from 200 mice to prepare SNS and two mice for the total cortex sample. We used a mouse monoclonal antibody specific for the C-terminus of FUS (Santa Cruz) to pull down FUS associated RNAs using magnetic beads. After immunoprecipitation, FUS-RNA complexes were treated with MNAse in mild conditions and the 5' end of RNAs were radiolabeled with P^{32}-gamma ATP. Samples run on SDS-gel (10% Bis Tris) were transferred to nitrocellulose membrane and visualized using FLA phosphorimager. RNAs corresponding to FUS-RNA complexes were purified from the nitrocellulose membrane and strand-specific paired-end CLIP libraries were sequenced on HiSeq 2500 for 15 cycles.

Bioinformatic analysis of CLIP-seq data and identification of FUS targets

Low quality reads were filtered and adapter sequences were removed with Trim Galore! (Krueger, F., TrimGalore. Retrieved February 24, 2010, from https://github.com/FelixKrueger/TrimGalore). Reads were aligned to the mouse reference genome (build GRCm38) using STAR version 2.4.2a[87] and Ensembl gene annotations (version 90). We allowed a maximum of two mismatches per read (--outFilterMismatchNmax 2) and removed all multimapping reads (--outFilterMultimapNmax 1). PCR duplicates were removed with Picard tools version 2.18.4 ("Picard Toolkit." 2019. Broad Institute, GitHub Repository. http://broadinstitute.github.io/picard/; Broad Institute). Peaks were called separately on each sample with CLIPper[52] using default parameters.

To identify regions that are specifically bound by FUS in the SNS sample but not the total cortex sample, we filtered the peaks based on an MA plot. For each peak, we counted the number of overlapping reads in the SNS (x) and total cortex samples (y). M (log2 fold change) and A (average log2 counts) were calculated as follows:

$$M = \log2[(x + o)/(lib.size\_x + o)] - \log2[(y + o)/(lib.size\_y + o)]$$

$$A = [\log2(x + o) + \log2(y + o)] / 2$$

where o = 1 is an offset to prevent a division by 0 and lib.size_x and lib.size_y is the effective library size of the two samples: the library size (number of reads mapping to the peaks) multiplied by the normalization factor obtained from "calcNormFactors" using the trimmed mean of M-values[88] method. The M and A values of all CLIPper peaks identified in

665 the SNS sample were plotted against each other (x-axis A, y-axis M). The plot was not
666 centered at a log2FC of 0. Therefore, we fitted a LOESS (locally estimated scatterplot
667 smoothing) curve for normalization (loess (formula=M~A, span=1/4, family="symmetric",
668 degree=1, iterations=4)). We computed the predicted M values (fitted) for each A value and
669 adjusted the M values by the fit (adjusted M = M - fitted M). After adjustment, the fitted
670 LOESS line crosses the y-axis at 0 with slope = 0 in the adjusted MA-plot.
671 For ranking purposes, we computed p-values for each peak with the Bioconductor edgeR
672 package[88]. We computed the common dispersion of the peaks at the center of the main
673 point cloud (-3 < y < 1 in raw MA-plot) and not the tagwise dispersion because we are
674 lacking replicate information. Peak specific offsets were computed as log
675 (lib.size*norm.factors) where norm.factors are the normalization factors. The fitted M-values
676 were subtracted from the peak specific offsets to use the adjustments from the LOESS fit for
677 the statistical inference. We fit a negative binomial generalized linear model to the peak
678 specific read counts using the adjusted offsets. We want to test for differential read counts
679 between the synaptoneurosome and total cortex sample (~group). A likelihood ratio test[89]
680 was run on each peak to test for synaptoneurosome versus total cortex differences.
681 We compared the sets of peaks obtained from different p-value cutoffs (Supplementary **Fig.**
682 **2g**) and choose the most stringed cutoff of 1e-5 because it showed the strongest depletion
683 of intronic peaks and strongest enrichment of exonic and 3'UTR peaks. CLIPper annotated
684 each peak to a gene and we manually inspected the assigned genes and removed wrong
685 assignments caused by overlapping gene annotations.
686 Total cortex-specific peaks (regions that are exclusively bound in the total cortex sample but
687 not the SNS sample) were computed with the same approach: the M values were computed
688 as
689 M = log2((y + o)/(lib.size_y + o)) - log2((x + o)/(lib.size_x + o))
690 and we used a p-value cutoff of 0.0029825 because that resulted in an identical number of
691 SNS-specific peaks.
692 For the over representation analysis (ORA) we applied the "goana" function from the limma
693 R package using the gene length as covariate[90]. As background set, we used all genes with
694 a cpm of at least 1 in all RNA-seq samples of synaptoneurosomes from 1-month-old mice.
695 RNA motifs of length 2-8 were predicted with HOMER[54]. To help with the motif finding, we
696 decided to use input sequences of equal length because the lengths of the predicted peaks
697 varied a lot. We define the peak center as the median position with maximum read
698 coverage. Then, we centered a window of size 41 on the peak center of each selected peak
699 and extracted the genomic sequence. We generated background sequences for each set of
700 target sequences. A background set consists of 200,000 sequences of length 41 from
701 random locations with the same annotation as the corresponding target set (intron, exon, 3'

UTR or 5' UTR). All background sequences are from regions without any read coverage in the corresponding CLIP-seq sample to ensure that the background sequences are not bound by FUS.

RNA extraction and high-throughput sequencing (RNA-seq)

Cortex tissue was isolated from 1 and/or 6-month-old *Fus*$^{\Delta NLS/+}$ and *Fus*$^{+/+}$ mice. Paired total cortex (200 l) and SNS sample was obtained from a single mouse per condition using filtration protocol as previously described. Briefly, frozen total and SNS samples were mixed with Qiazol reagent following the manufacturer's recommendations and incubated at RT for 5 min. Two hundred microliters of chloroform were added to the samples and mixed for 15s and then centrifuged for 15 min (12,000g, 4°C). To the upper aqueous phase collected, five hundred microliters of isopropanol and 0.8 l of glycogen was added and incubated at RT for 15 minutes. The samples were centrifuged at 10,000 rpm for 10 min. After centrifugation at 12,000g for 15 min, the isopropanol was removed and the pellet was washed with 1 ml of 70% ethanol and samples were centrifuged for 5 min at 7500g. Ethanol was discarded and the RNA pellet was air-dried and dissolved in nuclease free water and further purified using the RNeasy Mini Kit including the DNAse I digestion step. The concentration and the RIN values were determined by Bioanalyzer. 150 ng of total RNA were used for Poly A library preparation. Strand specific cDNA libraries were prepared and sequenced on Illumina NovaSeq6000 platform (2x150bp, paired end) from Eurofins Genomics, Konstanz, Germany.

Bioinformatic analysis of RNA-seq data

The preprocessing, gene quantification and differential gene expression analysis was performed with the ARMOR workflow[91]. In brief, reads were quality filtered and adapters were removed with Trim Galore! (Krueger, F., TrimGalore. Retrieved February 24, 2010, from https://github.com/FelixKrueger/TrimGalore). For visualization purposes, reads were mapped to the mouse reference genome GRCm38 with STAR version 2.4.2a[87] and default parameters using Ensembl gene annotations (version 90). BAM files were converted to BigWig files with bedtools[92]. Transcript abundance estimates were computed with Salmon version 0.10.2[93] and summarized to gene level with the tximeta R package[94]. All downstream analyses were performed in R and the edgeR package[88] was used for differential gene expression analysis. We filtered the lowly expressed genes and kept all genes with a CPM of at least 10/median_library_size*1e6 in 4 replicates (the size of the smallest group, here the nuclear samples). Additionally, each kept gene is required to have at least 15 counts across all samples. The filtered set of genes was used for the PCA plot and differential gene expression analysis.

cDNA synthesis and Quantitative Real-Time PCR

Total RNA was reverse transcribed using Superscript III kit (Invitrogen). For qRT-PCR, 2x SYBR master mix (Thermoscientific) were used and the reaction was run in Thermocycler (Applied Biosystems ViiA 7) following the manufacturer's instructions.

Primer list

Gene	Forward primer sequence	Reverse primer sequence
Actin B	GGTGGGTATGGGTCAGAAGGAC	GGCTGGGGTGTTGAAGGTCTC
CamkIIα	AATGGCAGATCGTCCACTTC	ATGAGAGGTGCCCTCAACAC
Psd-95	GTGGGCGGCGAGGATGGTGAA	CCGCCGTTTGCTGGGAATGAA

SDS-PAGE and Western blotting

Protein concentrations were determined using the Pierce BCA Protein Assay (Thermo Fisher Scientific) prior to SDS-PAGE. 20 g for total protein were used for western blots. The samples were resuspended in 1X SDS loading buffer with 1X final sample reducing reagent and boiled at 95°C, 10 mins. Samples were separated by Bolt 4-12% Bis-Tris pre-cast gels and transferred onto nitrocellulose membranes using iBlot® transfer NC stacks with iBlot Dry Blotting system (Invitrogen). Membranes were blocked with buffer containing 0.05% v/v Tween-20 (Sigma P1379) prepared in PBS (PBST) with 5% w/v non-fat skimmed powdered milk and probed with primary antibodies (list attached) overnight at 4°C in PBST with 1% w/v milk. Following three washes with PBST, membranes were incubated with secondary HRP-conjugated goat anti mouse or rabbit AffiniPure IgG antibodies (1:5000, 1:10000, respectively) (Jackson ImmunoResearch 115-035-146 and 111-035-144, respectively) in PBST with 1% w/v milk, for 1.5 hours at RT. Membranes were washed with PBST, and the bands were visualized using Amersham Imager 600RGB (GE Healthcare Life Sciences 29083467).

Transmission Electron Microscopy

SNS pellets were prepared from cortical tissue of 1-month-old C57/Bl6 mice as previously mentioned before and submitted to imaging facility at ZMB UZH. Briefly, SNS pellet prepared were re-suspended in 2X fixative (5% Glutaraldehyde in 0.2 M Cacodylate buffer) and fixed at RT for 30 mins. Sample was then washed twice with 0.1 M Cacodylate buffer before embedding into 2% Agar Nobile. Post-fixation was performed with 1% Osmium 1 hour on ice, washed three times with ddH_2O, dehydrated with 70% ethanol for 20 mins, followed by 80% ethanol for 20 mins, 100% for 30 mins and finally Propylene for 30 mins.

Propylene: Epon Araldite at 1:1 were added overnight followed by addition of Epon Araldite for 1 hour at RT. Sample was then embedded via 28 hours incubation at 60°C. The resulting block was then cut into 60 nm ultrathin sections using ultramicrotome. Ribbons of sections were then put onto TEM grid and imaged on TEM - FEI CM100 electron microscope (modify).

Confocal image acquisition and analysis

Confocal images were acquired on a Leica SP8 Falcon microscope using 63X (NA 1.4) with a zoom power of 3. Images were acquired at a 2048x2048 pixel size, yielding to a 30.05 nm/pixel resolution. To quantify the density of synaptic markers, images were acquired in CA1 region in the apical dendrite area, ~50 m from the soma, at the bifurcation of the apical dendrite of pyramidal cells, using the same parameters for both genotypes. Images were acquired from top to bottom with a Z step size of 500 nm. Images were deconvoluted using Huygens Professional software (Scientific Volume Imaging). Images were then analyzed as described previously[84]. Briefly, stacks were analyzed using the built-in particle analysis function in Fiji[95]. The size of the particles was defined according to previously published studies[80,96,97]. To assess the number of clusters, images were thresholded (same threshold per marker and experiment), and a binary mask was generated. A low size threshold of 0.01 m diameter and high pass threshold of 1 m diameter was applied. Top and bottom stacks were removed from the analysis to only keep the 40 middle stacks. For the analysis, the number of clusters per 40z stacks was summed and normalized by the volume imaged (75153.8 m3). The density was normalized by the control group. The densities were compared by t test for 1- and 6-month-old mice. GluN1 synaptic localization was analyzed by counting the number of colocalized GluN1 clusters with Synapsin 1. Colocalization clusters were generated using ImageJ plugin colocalization highlighter. The default parameters were applied to quantify the colocalization. The number of colocalized clusters were quantified using the built-in particle analysis function in Fiji[95].

Synaptic density and composition imaging and analysis of primary neuronal culture

Imaging and quantification were performed as previously reported[98]. Briefly, synaptic density and synapse composition was assayed in 22 DIV neuronal cell cultures. Cultures were fixed in cold 4% PFA with 4% sucrose for 20 minutes at RT. Primary antibodies were incubated overnight at 4°C. secondary antibodies were incubated for 3h at RT. Hippocampal primary culture: pyramidal cells were selected based on their morphology and confocal images were acquired on a Leica SP8 Falcon microscope using 63X (NA 1.4) with a zoom power of 3 and analyzed with Fiji software. After deconvolution (huygens professional), images were

subsequently thresholded, and subsequent analyses were performed by an investigator blind to cell culture treatment.

Antibody list

Antibody	Species, Source	STORM dilution	Confocal dilution	Western blot dilution
FUS	Rb, A300-293A, Bethyl		1:500	1:1000
FUS	Rb, A300-294A, Bethyl			1:1000
FUS	Ms, 4H11, Santa Cruz	1:200		
PSD-95	Ms, Invitrogen	1:200	1:1000	1:1000
P-CAMKIIa	Ms, D21E4, Cell signaling	1:500	1:500	1:1000
PNF	Ms, SMI31, Covance		1:1000	
Spinophilin	Rb, Synaptic Systems	1:500		
Synapsin 1	Ms, Synaptic Systems	1:200	1:500	
GluA1	Rb, Sigma Aldrich	1:200	1:200	1:1000
GluN1	Ms, Covance		1:500	
GluN2B	Rb, Sigma Aldrich	1:500		1:2000
Bassoon	Gp, Synaptic Systems	1:500	1:500	
GRP78 BiP (ER)	Rb, Abcam	1:200		

MAP2	Ms, Sigma Aldrich		1:1000	
SYP	Ms, Santa cruz			1:200
GABAA/alpha1	Gp, Synaptic Systems		1:500	
GABAA/alpha3	Rb, Synaptic Systems		1:500	
Gephyrin	Ms, Synaptic Systems		1:500	
Vgat	Gp, Synaptic Systems		1:500	
β-Actin	Ms, Sigma			1:5000
SNAP25	Gp, Synaptic Systems		1:500	1:1000

Author Contribution

Conceptualization of the study was carried by S.S., K.M.H., and M.P.. S.S. performed synaptosome isolation, CLIP-seq sample preparation and RNA-seq sample preparation. K.M.H. analyzed the data from CLIP-seq and RNA-seq. S.S., K.M.H., M.D.R. and M.P. developed the strategy to analyze the sequencing data. E.T., M.H.P., M.P.B., J.W., and P.S. provided experimental support for the experiments. L.D. provided the mouse model and input on the study. P.D.R. performed immunostaining and image analyses including confocal, STED and dSTORM. S.S., K.M.H, E.T, P.D.R and M.P wrote and edited the manuscript. M.D.R, P.D.R and M.P. provided supervision. M.P directed the entire study. All authors read, edited, and approved the final manuscript.

Acknowledgments

We gratefully acknowledge the support of the National Centre for Competence in Research (NCCR) RNA & Disease funded by the Swiss National Science Foundation. SSMK was supported by Swiss Government Excellence Scholarships for Foreign Scholars. The authors would like to thank Prof. Adriano Aguzzi and Dr. Claudia Scheckel for helpful discussions and Dr. Dorothee Dormann for critical comments on the manuscript. We thank Gery Barmettler and Dr. José María Mateos from ZMB UZH for technical help with TEM. We also thank Catharina Aquino and Lucy Poveda from FGCZ for discussions and technical help on CLIP library preparation and sequencing.

850

851

852

853

854

855

856

References

1. Lagier-Tourenne, C., Polymenidou, M. & Cleveland, D. W. TDP-43 and FUS/TLS: Emerging roles in RNA processing and neurodegeneration. *Hum. Mol. Genet.* **19**, 46–64 (2010).
2. Andersson, M. K. *et al.* The multifunctional FUS, EWS and TAF15 proto-oncoproteins show cell type-specific expression patterns and involvement in cell spreading and stress response. *BMC Cell Biol.* (2008). doi:10.1186/1471-2121-9-37
3. Dormann, D. *et al.* ALS-associated fused in sarcoma (FUS) mutations disrupt transportin-mediated nuclear import. *EMBO J.* (2010). doi:10.1038/emboj.2010.143
4. Ederle, H. *et al.* Nuclear egress of TDP-43 and FUS occurs independently of Exportin-1/CRM1. *Sci. Rep.* (2018). doi:10.1038/s41598-018-25007-5
5. Hock, E. M. *et al.* Hypertonic Stress Causes Cytoplasmic Translocation of Neuronal, but Not Astrocytic, FUS due to Impaired Transportin Function. *Cell Rep.* (2018). doi:10.1016/j.celrep.2018.06.094
6. Kwiatkowski, T. J. *et al.* Mutations in the FUS/TLS gene on chromosome 16 cause familial amyotrophic lateral sclerosis. *Science (80-.).* **323**, 1205–1208 (2009).
7. Vance, C. *et al.* Mutations in FUS, an RNA processing protein, cause familial amyotrophic lateral sclerosis type 6. *Science (80-.).* (2009). doi:10.1126/science.1165942
8. Neumann, M. *et al.* A new subtype of frontotemporal lobar degeneration with FUS pathology. *Brain* **132**, 2922–2931 (2009).
9. Lee, B. J. *et al.* Rules for Nuclear Localization Sequence Recognition by Karyopherinβ2. *Cell* (2006). doi:10.1016/j.cell.2006.05.049
10. Mackenzie, I. R. A., Rademakers, R. & Neumann, M. TDP-43 and FUS in amyotrophic lateral sclerosis and frontotemporal dementia. *Lancet Neurol.* **9**, 995–1007 (2010).
11. Sama, R. R. K. *et al.* FUS/TLS assembles into stress granules and is a prosurvival factor during hyperosmolar stress. *J. Cell. Physiol.* (2013). doi:10.1002/jcp.24395
12. Murakami, T. *et al.* ALS/FTD Mutation-Induced Phase Transition of FUS Liquid Droplets and Reversible Hydrogels into Irreversible Hydrogels Impairs RNP Granule Function. *Neuron* (2015). doi:10.1016/j.neuron.2015.10.030
13. Patel, A. *et al.* A Liquid-to-Solid Phase Transition of the ALS Protein FUS Accelerated by Disease Mutation. *Cell* (2015). doi:10.1016/j.cell.2015.07.047
14. Guo, L. *et al.* Nuclear-Import Receptors Reverse Aberrant Phase Transitions of RNA-Binding Proteins with Prion-like Domains. *Cell* (2018). doi:10.1016/j.cell.2018.03.002
15. Yoshizawa, T. *et al.* Nuclear Import Receptor Inhibits Phase Separation of FUS through Binding to Multiple Sites. *Cell* (2018). doi:10.1016/j.cell.2018.03.003
16. Hofweber, M. *et al.* Phase Separation of FUS Is Suppressed by Its Nuclear Import Receptor and Arginine Methylation. *Cell* (2018). doi:10.1016/j.cell.2018.03.004
17. Qamar, S. *et al.* FUS Phase Separation Is Modulated by a Molecular Chaperone and Methylation of Arginine Cation-π Interactions. *Cell* (2018). doi:10.1016/j.cell.2018.03.056
18. Maharana, S. *et al.* RNA buffers the phase separation behavior of prion-like RNA binding

proteins. *Science (80-.).* (2018). doi:10.1126/science.aar7366
19. Langdon, E. M. *et al.* mRNA structure determines specificity of a polyQ-driven phase separation. *Science (80-.).* (2018). doi:10.1126/science.aar7432
20. Fujii, R. *et al.* The RNA binding protein TLS is translocated to dendritic spines by mGluR5 activation and regulates spine morphology. *Curr. Biol.* (2005). doi:10.1016/j.cub.2005.01.058
21. Biever, A., Donlin-Asp, P. G. & Schuman, E. M. Local translation in neuronal processes. *Current Opinion in Neurobiology* (2019). doi:10.1016/j.conb.2019.02.008
22. Lagier-Tourenne, C. *et al.* Divergent roles of ALS-linked proteins FUS/TLS and TDP-43 intersect in processing long pre-mRNAs. *Nat. Neurosci.* **15**, 1488–1497 (2012).
23. Udagawa, T. *et al.* FUS regulates AMPA receptor function and FTLD/ALS-associated behaviour via GluA1 mRNA stabilization. *Nat. Commun.* **6**, (2015).
24. Yokoi, S. *et al.* 3′UTR Length-Dependent Control of SynGAP Isoform α2 mRNA by FUS and ELAV-like Proteins Promotes Dendritic Spine Maturation and Cognitive Function. *Cell Rep.* **20**, 3071–3084 (2017).
25. Schoen, M. *et al.* Super-resolution microscopy reveals presynaptic localization of the ALS/FTD related protein FUS in hippocampal neurons. *Front. Cell. Neurosci.* **9**, 1–16 (2016).
26. So, E. *et al.* Mitochondrial abnormalities and disruption of the neuromuscular junction precede the clinical phenotype and motor neuron loss in hFUSWT transgenic mice. *Hum. Mol. Genet.* **27**, 463–474 (2018).
27. Deshpande, D. *et al.* Synaptic FUS localization during motoneuron development and its accumulation in human ALS synapses. *Front. Cell. Neurosci.* **13**, 1–17 (2019).
28. Aoki, N. *et al.* Localization of fused in sarcoma (FUS) protein to the post-synaptic density in the brain. *Acta Neuropathol.* **124**, 383–394 (2012).
29. Hicks, G. G. *et al.* Fus deficiency in mice results in defective B-lymphocyte development and activation, high levels of chromosomal instability and perinatal death. *Nat. Genet.* (2000). doi:10.1038/72842
30. Kino, Y. *et al.* FUS/TLS deficiency causes behavioral and pathological abnormalities distinct from amyotrophic lateral sclerosis. *Acta Neuropathol. Commun.* **3**, 24 (2015).
31. López-Erauskin, J. *et al.* ALS/FTD-Linked Mutation in FUS Suppresses Intra-axonal Protein Synthesis and Drives Disease Without Nuclear Loss-of-Function of FUS. *Neuron* **100**, 816-830.e7 (2018).
32. Fogarty, M. J. Driven to decay: Excitability and synaptic abnormalities in amyotrophic lateral sclerosis. *Brain Research Bulletin* (2018). doi:10.1016/j.brainresbull.2018.05.023
33. Starr, A. & Sattler, R. Synaptic dysfunction and altered excitability in C9ORF72 ALS/FTD. *Brain Research* (2018). doi:10.1016/j.brainres.2018.02.011
34. Henstridge, C. M. *et al.* Synapse loss in the prefrontal cortex is associated with cognitive decline in amyotrophic lateral sclerosis. *Acta Neuropathol.* (2018). doi:10.1007/s00401-017-1797-4
35. Sephton, C. F. & Yu, G. The function of RNA-binding proteins at the synapse: Implications for neurodegeneration. *Cellular and Molecular Life Sciences* (2015). doi:10.1007/s00018-015-

1943-x
36. Selkoe, D. J. Alzheimer's disease is a synaptic failure. *Science* (2002). doi:10.1126/science.1074069
37. Hoell, J. I. *et al.* RNA targets of wild-type and mutant FET family proteins. *Nat. Struct. Mol. Biol.* (2011). doi:10.1038/nsmb.2163
38. Rogelj, B. *et al.* Widespread binding of FUS along nascent RNA regulates alternative splicing in the brain. *Sci. Rep.* **2**, 1–10 (2012).
39. Ishigaki, S. *et al.* Position-dependent FUS-RNA interactions regulate alternative splicing events and transcriptions. *Sci. Rep.* **2**, 1–8 (2012).
40. Masuda, A. *et al.* Position-specific binding of FUS to nascent RNA regulates mRNA length. *Genes Dev.* **29**, 1045–1057 (2015).
41. Nakaya, T., Alexiou, P., Maragkakis, M., Chang, A. & Mourelatos, Z. FUS regulates genes coding for RNA-binding proteins in neurons by binding to their highly conserved introns. *RNA* (2013). doi:10.1261/rna.037804.112
42. Loughlin, F. E. *et al.* The Solution Structure of FUS Bound to RNA Reveals a Bipartite Mode of RNA Recognition with Both Sequence and Shape Specificity. *Mol. Cell* (2019). doi:10.1016/j.molcel.2018.11.012
43. Scekic-Zahirovic, J. *et al.* Toxic gain of function from mutant FUS protein is crucial to trigger cell autonomous motor neuron loss. *EMBO J.* **35**, 1077–97 (2016).
44. Gerth, F. *et al.* Intersectin associates with synapsin and regulates its nanoscale localization and function. *Proc. Natl. Acad. Sci. U. S. A.* (2017). doi:10.1073/pnas.1715341114
45. Nishimune, H., Badawi, Y., Mori, S. & Shigemoto, K. Dual-color STED microscopy reveals a sandwich structure of Bassoon and Piccolo in active zones of adult and aged mice. *Sci. Rep.* (2016). doi:10.1038/srep27935
46. Fukata, Y. *et al.* Local palmitoylation cycles define activity-regulated postsynaptic subdomains. *J. Cell Biol.* (2013). doi:10.1083/jcb.201302071
47. Wang, W. Y. *et al.* Interaction of FUS and HDAC1 regulates DNA damage response and repair in neurons. *Nat. Neurosci.* (2013). doi:10.1038/nn.3514
48. Rulten, S. L. *et al.* PARP-1 dependent recruitment of the amyotrophic lateral sclerosis-associated protein FUS/TLS to sites of oxidative DNA damage. *Nucleic Acids Res.* (2014). doi:10.1093/nar/gkt835
49. Tan, A. Y. & Manley, J. L. TLS Inhibits RNA Polymerase III Transcription. *Mol. Cell. Biol.* (2010). doi:10.1128/mcb.00884-09
50. Schwartz, J. C. *et al.* FUS binds the CTD of RNA polymerase II and regulates its phosphorylation at Ser2. *Genes Dev.* (2012). doi:10.1101/gad.204602.112
51. Polymenidou, M. *et al.* Long pre-mRNA depletion and RNA missplicing contribute to neuronal vulnerability from loss of TDP-43. *Nat. Neurosci.* (2011). doi:10.1038/nn.2779
52. Lovci, M. T. *et al.* Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. *Nat. Struct. Mol. Biol.* **20**, 1434–1442 (2013).
53. Wang, T., Xie, Y. & Xiao, G. dCLIP: A computational approach for comparative CLIP-seq

analyses. *Genome Biol.* (2014). doi:10.1186/gb-2014-15-1-r11

54. Heinz, S. *et al.* Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol. Cell* (2010). doi:10.1016/j.molcel.2010.05.004

55. Scekic□Zahirovic, J. *et al.* Toxic gain of function from mutant FUS protein is crucial to trigger cell autonomous motor neuron loss . *EMBO J.* (2016). doi:10.15252/embj.201592559

56. Preston, A. R. & Eichenbaum, H. Interplay of hippocampus and prefrontal cortex in memory. *Current Biology* (2013). doi:10.1016/j.cub.2013.05.041

57. Somogyi, P., Tamás, G., Lujan, R. & Buhl, E. H. Salient features of synaptic organisation in the cerebral cortex. in *Brain Research Reviews* (1998). doi:10.1016/S0165-0173(97)00061-1

58. Zhu, F. *et al.* Architecture of the Mouse Brain Synaptome. *Neuron* (2018). doi:10.1016/j.neuron.2018.07.007

59. Contreras, A., Hines, D. J. & Hines, R. M. Molecular specialization of GABAergic synapses on the soma and axon in cortical and hippocampal circuit function and dysfunction. *Frontiers in Molecular Neuroscience* (2019). doi:10.3389/fnmol.2019.00154

60. Distler *et al.* Proteomic Analysis of Brain Region and Sex-Specific Synaptic Protein Expression in the Adult Mouse Brain. *Cells* **9**, 313 (2020).

61. Irfan, M. *et al.* SNAP-25 isoforms differentially regulate synaptic transmission and long-term synaptic plasticity at central synapses. *Sci. Rep.* (2019). doi:10.1038/s41598-019-42833-3

62. Chaudhry, F. A. *et al.* The vesicular GABA transporter, VGAT, localizes to synaptic vesicles in sets of glycinergic as well as GABAergic neurons. *J. Neurosci.* (1998). doi:10.1523/jneurosci.18-23-09733.1998

63. McIlhinney, R. A. J. *et al.* Assembly of N-methyl-D-aspartate (NMDA) receptors. in *Biochemical Society Transactions* (2003). doi:10.1042/BST0310865

64. Diering, G. H. & Huganir, R. L. The AMPA Receptor Code of Synaptic Plasticity. *Neuron* (2018). doi:10.1016/j.neuron.2018.10.018

65. Sigel, E. & Steinmann, M. E. Structure, function, and modulation of GABAA receptors. *Journal of Biological Chemistry* (2012). doi:10.1074/jbc.R112.386664

66. Choii, G. & Ko, J. Gephyrin: a central GABAergic synapse organizer. *Experimental & molecular medicine* (2015). doi:10.1038/emm.2015.5

67. Brüning, F. *et al.* Sleep-wake cycles drive daily dynamics of synaptic phosphorylation. *Science (80-.).* **366**, (2019).

68. Holt, C. E., Martin, K. C. & Schuman, E. M. Local translation in neurons: visualization and function. *Nat. Struct. Mol. Biol.* **26**, 557–566 (2019).

69. Kosik, K. S. Life at Low Copy Number: How Dendrites Manage with So Few mRNAs. *Neuron* (2016). doi:10.1016/j.neuron.2016.11.002

70. Cajigas, I. J. *et al.* The Local Transcriptome in the Synaptic Neuropil Revealed by Deep Sequencing and High-Resolution Imaging. *Neuron* **74**, 453–466 (2012).

71. Fritschy, J. □M & Mohler, H. GABAA□receptor heterogeneity in the adult rat brain: Differential regional and cellular distribution of seven major subunits. *J. Comp. Neurol.* (1995).

doi:10.1002/cne.903590111
72. Dias, R. *et al.* Evidence for a significant role of α3-containing GABAA receptors in mediating the anxiolytic effects of benzodiazepines. *J. Neurosci.* (2005). doi:10.1523/JNEUROSCI.1166-05.2005
73. Smith, K. S., Engin, E., Meloni, E. G. & Rudolph, U. Benzodiazepine-induced anxiolysis and reduction of conditioned fear are mediated by distinct GABAA receptor subtypes in mice. *Neuropharmacology* (2012). doi:10.1016/j.neuropharm.2012.03.001
74. Fischer, B. D. *et al.* Contribution of GABAA receptors containing α3 subunits to the therapeutic-related and side effects of benzodiazepine-type drugs in monkeys. *Psychopharmacology (Berl).* (2011). doi:10.1007/s00213-010-2142-y
75. Husi, H., Ward, M. A., Choudhary, J. S., Blackstock, W. P. & Grant, S. G. N. Proteomic analysis of NMDA receptor-adhesion protein signaling complexes. *Nat. Neurosci.* **3**, 661–669 (2000).
76. Tafoya, L. C. R. *et al.* Expression and function of SNAP-25 as a universal SNARE component in GABAergic neurons. *J. Neurosci.* (2006). doi:10.1523/JNEUROSCI.1866-06.2006
77. Wojcik, S. M. *et al.* A Shared Vesicular Carrier Allows Synaptic Corelease of GABA and Glycine. *Neuron* (2006). doi:10.1016/j.neuron.2006.04.016
78. Mukherjee, J. *et al.* The residence time of GABA ARs at inhibitory synapses is determined by direct binding of the receptor α1 subunit to gephyrin. *J. Neurosci.* (2011). doi:10.1523/JNEUROSCI.2001-11.2011
79. Tretter, V. *et al.* Molecular basis of the γ-aminobutyric acid a receptor α3 subunit interaction with the clustering protein gephyrin. *J. Biol. Chem.* (2011). doi:10.1074/jbc.M111.291336
80. Specht, C. G. *et al.* Quantitative nanoscopy of inhibitory synapses: Counting gephyrin molecules and receptor binding sites. *Neuron* (2013). doi:10.1016/j.neuron.2013.05.013
81. Schieweck, R. & Kiebler, M. A. Posttranscriptional gene regulation of the GABA receptor to control neuronal inhibition. *Front. Mol. Neurosci.* (2019). doi:10.3389/fnmol.2019.00152
82. He, Q. & Ge, W. The tandem Agenet domain of fragile X mental retardation protein interacts with FUS. *Sci. Rep.* (2017). doi:10.1038/s41598-017-01175-8
83. Heraud-Farlow, J. E. & Kiebler, M. A. The multifunctional Staufen proteins: Conserved roles from neurogenesis to synaptic plasticity. *Trends in Neurosciences* (2014). doi:10.1016/j.tins.2014.05.009
84. Rossi, P. De, Nomura, T., Andrew, R. J. & Nicholson, D. A. Neuronal BIN1 Regulates Presynaptic Neurotransmitter Release and Memory Article Neuronal BIN1 Regulates Presynaptic Neurotransmitter Release and Memory Consolidation. *CellReports* **30**, 3520-3535.e7 (2020).
85. Williams, C. *et al.* Transcriptome analysis of synaptoneurosomes identifies neuroplasticity genes overexpressed in incipient Alzheimer's disease. *PLoS One* (2009). doi:10.1371/journal.pone.0004936
86. Muddashetty, R. S., Kelić, S., Gross, C., Xu, M. & Bassell, G. J. Dysregulated metabotropic glutamate receptor-dependent translation of AMPA receptor and postsynaptic density-95

mRNAs at synapses in a mouse model of fragile X syndrome. *J. Neurosci.* (2007). doi:10.1523/JNEUROSCI.0937-07.2007

87. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* (2013). doi:10.1093/bioinformatics/bts635

88. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* (2010). doi:10.1186/gb-2010-11-3-r25

89. McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* (2012). doi:10.1093/nar/gks042

90. Young, M. D., Wakefield, M. J., Smyth, G. K. & Oshlack, A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol.* (2010). doi:10.1186/gb-2010-11-2-r14

91. Orjuela, S., Huang, R., Hembach, K. M., Robinson, M. D. & Soneson, C. ARMOR: An automated reproducible modular workflow for preprocessing and differential analysis of RNA-seq data. *G3 Genes, Genomes, Genet.* (2019). doi:10.1534/g3.119.400185

92. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* (2010). doi:10.1093/bioinformatics/btq033

93. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* (2017). doi:10.1038/nmeth.4197

94. Love, M. I. *et al.* Tximeta: reference sequence checksums for provenance identification in RNA-seq. *bioRxiv* (2019). doi:10.1101/777888

95. Rueden, C. T. *et al.* ImageJ2: ImageJ for the next generation of scientific image data. *BMC Bioinformatics* (2017). doi:10.1186/s12859-017-1934-z

96. MacGillavry, H. D., Song, Y., Raghavachari, S. & Blanpied, T. A. Nanoscale scaffolding domains within the postsynaptic density concentrate synaptic ampa receptors. *Neuron* (2013). doi:10.1016/j.neuron.2013.03.009

97. Nair, D. *et al.* Super-resolution imaging reveals that AMPA receptors inside synapses are dynamically organized in nanodomains regulated by PSD95. *J. Neurosci.* (2013). doi:10.1523/JNEUROSCI.2381-12.2013

98. De Rossi, P. *et al.* Predominant expression of Alzheimer's disease-associated BIN1 in mature oligodendrocytes and localization to white matter tracts. *Mol. Neurodegener.* (2016). doi:10.1186/s13024-016-0124-1

99. Noya, S. B. *et al.* The forebrain synaptic transcriptome is organized by clocks but its proteome is driven by sleep. *Science (80-.).* **366**, (2019).

Figure legends

Fig. 1 FUS is enriched at the presynaptic compartment

(**a**) Confocal images showing the distribution of FUS (green) in the pyramidal layer of the retrosplenial cortical area along with MAP2 (blue) and PNF (magenta). Left panel shows the overview and the right panel the zoomed in area labelled with the red box on the left panel. (**b**) Similar confocal images showing FUS (green) along with PSD95 (orange) and Synapsin 1 (Syn, blue. (**c**) Synaptic localization of FUS was assessed by STED microscopy using excitatory (PSD95) and inhibitory (VGAT) markers for synapses. 60 m brain sections were analyzed and distance between FUS and the synaptic markers was analyzed using Imaris. (**d**) Bar graph representing the percentage of synapses within 200 nm of FUS clusters and showing an enrichment of FUS at the excitatory synapses. (**e**) dSTORM was used to explore more precisely the FUS localization within the synapse, using primary culture. Bassoon and Synapsin 1 (Syn) were used to label the presynaptic compartment and GluN1, GluA1 and PSD95 were used to label the postsynapse. Spinophilin (Spino) was used to label the spines. (**f**) Bar graph representing the percentage of FUS localized within 100nm from presynaptic or postsynaptic markers. (**g**) Bar graph representing the distribution of FUS in the synapse. (**h**) Schematic summarizing the FUS localization within the synapse. Graph bar showing mean + SD. *p>0.05, **p>0.01, ***p>0.001, ****p>0.000.

Fig. 2 CLIP-seq on cortical synaptoneurosomes identified FUS-associated pre- and postsynaptic RNAs

(a) Electron microscopic images of synaptoneurosomes (SNS) from mouse cortex showing intact pre- and postsynaptic compartments. **(b)** Western blot of synaptic proteins (PSD95, p-CamKII), nuclear protein (Lamin B1) and FUS in total and SNS. **(c)** qPCR shows enrichment of PSD95, CamKII mRNAs in SNS. **(d)** Autoradiograph of FUS-RNA complexes immunoprecipitated from total homogenate and SNS and trimmed by different concentrations of micrococcal nuclease (MNase). **(e)** MA-plot of CLIPper peaks predicted in the SNS CLIP-seq sample. logCPM is the average log2CPM of each peak in the total cortex and SNS sample and logFC is the log2 fold-change between the number of reads in the SNS and total cortex sample. **(f)** Same MA-plot as E showing the selected, SNS specific peaks (p-value cutoff of 1e-05) in red. **(g)** Barplot with the percentage of SNS and total cortex specific peaks located in exons, 5'UTRs, 3'UTRs or introns. FUS binding in *Grin1* **(h)**, *Gabra1* **(i)** in total cortex (green) and SNS (blue). **(j)** Schematic with the cellular localization and function of some of the selected FUS targets.

Fig. 3 Increased synaptic FUS localization in *Fus*$^{\Delta NLS/+}$ mice affect GABAergic synapses (**a**) Schematic showing specificity of antibodies used for western blot against protein domains of FUS. (**b**) Western blot of total FUS, full length FUS and actin in

synaptoneurosomes isolated from *Fus*$^{+/+}$ and *Fus*$^{\Delta NLS/+}$ mice at 6 months of age. (**c**) Quantification of total FUS and full length FUS levels in synaptoneurosomes from *Fus*$^{+/+}$ and *Fus*$^{\Delta NLS/+}$ at 6 months of age. (**d**) Confocal images of the hippocampal CA1 area from 6-month-old mice showing higher level of FUS in the dendritic tree and synaptic compartment in *Fus*$^{\Delta NLS/+}$ mouse-model. On the top, low magnification pictures show the dendritic area of pyramidal cells stained with FUS (green), MAP2 (dendritic marker, magenta), Synapsin 1 (Syn, Synaptic marker, Cyan) and DAPI (Blue). Red box indicates the area imaged in the high magnification images below. (**e**) Higher magnification equivalent to the area highlighted in red in (**d**). (**f**) Representative images of staining using synaptic markers Synapsin 1, VGAT, $GABA_A\alpha3$ and GluN1 in *Fus*$^{+/+}$ and *Fus*$^{\Delta NLS/+}$ at 1 and 6 months of age. Images were generated with Imaris and display volume view used for quantification with statistically coded surface area. Density and cluster area were analyzed. (**g**) Graph bar representation of the synaptic density of Synapsin 1, VGAT, $GABA_A\alpha3$ and GluN1 from *Fus*$^{+/+}$ and *Fus*$^{\Delta NLS/+}$ at 1 and 6 months of age. Graph bar showing mean + SD. *p<0.05. Graphs are extracted from the same analysis shown in **Supplementary Fig. 3e-f**. The statistical analysis can be found in **Table 2.** (**h**) Colocalization analysis of GluN1 with Synapsin 1 to identify synaptic NMDAR and extrasynaptic NMDAR. Results were normalized by the control of each group. Graph bar showing mean + SD. *p<0.05. (**i**) Box and Whiskers representation of the average cluster area for each marker (Synapsin1, VGAT, $GABA_A\alpha3$ and GluN1) from 1-month and 6-month-old *Fus*$^{+/+}$ and *Fus*$^{\Delta NLS/+}$ mice. Box showing Min to Max, *p<0.05 **p<0.01. Graphs are extracted from the same analysis shown in **Supplementary Fig. 3f-i.** The statistical analysis can be found in **Table 3.**

Fig. 4 Age-dependent alterations in the synaptic RNA profile of *Fus*$^{\Delta NLS/+}$ mouse cortex (**a**) Outline of the RNA-seq experiment. (**b**) Heatmap from the set of up- and downregulated genes in SNS of *Fus*$^{\Delta NLS/+}$ at 6-months compared to *Fus*$^{+/+}$. Genes are on the rows and the different samples on the columns. The color scale indicates the log2FC between the CPM of each sample and mean CPM of the corresponding *Fus*$^{+/+}$ samples at each time point [sample logCPM – mean (logCPM of *Fus*$^{+/+}$ samples)]. (**c**) Volcano plots showing the log2 fold change of each gene and the corresponding minus log10 (FDR) of the differential gene expression analysis comparing *Fus*$^{\Delta NLS/+}$ SNS to *Fus*$^{+/+}$ SNS at 1 month (left panel) and 6 months of age (right panel). The horizontal line marks the significance threshold of 0.05. Significantly downregulated genes are highlighted in green, upregulated genes in purple and all FUS targets identified in the CLIP-seq data in blue. (**d**) Venn diagram of the sets of significantly up- and downregulated genes (SNS of *Fus*$^{\Delta NLS/+}$ vs. *Fus*$^{+/+}$ at 6 months of age) and the SNS FUS target genes identified by our FUS CLIP-seq. (**e**) Schematic of the cellular localization of the differentially expressed FUS targets in SNS of *Fus*$^{\Delta NLS/+}$ mice at 6 months

of age.

Table 1: FUS binds GU-rich sequences at the synapse

Predicted sequence motifs (HOMER) in windows of size 41 centered on the position with maximum coverage in each peak. Each set of target sequences has a corresponding background set with 200,000 sequences without any CLIP-seq read coverage (they are not bound by FUS). Note: These are all motifs that were not marked as possible false positives by HOMER and that occur in more than 1% of the target sequences.

Table 2. Statistical analysis of synaptic density

The table reports statistical analysis of density of the synaptic markers analyzed from a minimum of 2 images from at least 4 animals per genotype ($Fus^{+/+}$ and $Fus^{\Delta NLS/+}$) at 1 and 6 months of age. Unpaired t-test statistics, p-values, specific t-distribution (t), degrees of freedom (DF) and sample size are listed.

Table 3. Statistical analysis of synaptic cluster area

The table reports statistical analysis of area of the synaptic markers analyzed from a minimum of 2 images from at least 4 animals per genotype ($Fus^{+/+}$ and $Fus^{\Delta NLS/+}$) at 1 and 6 months of age. Unpaired t-test statistics, p-values, specific t-distribution (t), degrees of freedom (DF) and sample size are listed.

Supplemental Figures titles and legends

Supplementary Fig. 1 FUS is enriched at the presynaptic compartment

(**a**) Confocal images showing the distribution of FUS (green) in the molecular layer of the CA1 hippocampal area along with MAP2 (blue) and PNF (magenta). Left panel shows the overview and the right panel, the zoomed in area labelled with the red box on the left panel. (**b**) Similar confocal images showing FUS (green) along with PSD95 (orange) and Synapsin 1 (Syn, blue). (**c**) Schematic of the workflow for distance calculation after STED imaging. (**d**) Schematic of the workflow for distance calculation after STORM imaging. (**e**) Representative images of STORM imaging for FUS-GluN2B-Synapsin1 and FUS-PSD95-Bassoon. (**f**) Violin graph representing the distance distribution between FUS and synaptic markers. (**g**) Binning distribution showing the distance between FUS and the markers (in relative frequency) for PSD95, GluN2b, GluA1, Bassoon, Synapsin and BiP.

Supplementary Fig. 2 CLIP-seq on cortical synaptoneurosomes identified FUS-associated pre- and postsynaptic RNAs

(**a**) Western blot of synaptic proteins (GluN2b, SNAP25, GluA1, NRXN1), nuclear protein (Histone H3) in total cortex and synaptoneurosomes (SNS). (**b**) Schematic of CLIP-seq workflow from total homogenate and SNS from mouse cortex. (**c**) Immunoblot showing efficient immunoprecipitation of FUS from total cortex and SNS. (**d**) Flow chart illustrating the reads analyzed to define FUS peaks in total and SNS. (**e**) MA-plot of CLIPper peaks predicted in the total cortex CLIP-seq sample. logCPM is the average log2CPM of each peak in the total cortex and SNS sample and logFC is the log2 fold-change between the number of reads in the total cortex and SNS sample. (**f**) Same MA-plot as (**e**) showing the selected, total cortex specific peaks (p-value cutoff of 3e-03) in red. (**g**) Bar plot of different sets of SNS peaks and their location in genes. The p-value cutoff of each set is on the x-axis and no cutoff refers to the full list of all predicted SNS CLIPper peaks. The selected cutoff is in bold. (**h**) Bar plot of different sets of total cortex peaks and their location in genes. The p-value cutoff of each set is on the x-axis and no cutoff refers to the full list of all predicted SNS CLIPper peaks. The selected cutoff is in bold. (**i**) GO terms enriched among the synapse specific FUS RNA targets.

Supplementary Fig. 3 Increased synaptic FUS localization in *Fus*$^{\Delta NLS/+}$ mice affect GABAergic synapses

(**a**) Western blot of total FUS, full length FUS and actin in synaptoneurosomes isolated from 1-month-old *Fus*$^{+/+}$ and *Fus*$^{\Delta NLS/+}$ mice. (**b**) Quantification of total FUS and full length FUS levels in synaptoneurosomes from *Fus*$^{+/+}$ and *Fus*$^{\Delta NLS/+}$ at 1 month of age. (**c**) Confocal images of the hippocampal CA1 area from 1-month-old mice showing higher level of FUS in the dendritic tree and synaptic compartment in *Fus*$^{\Delta NLS/+}$ mouse-model. On the top, low magnification pictures show the dendritic area of pyramidal cells stained with FUS (green), MAP2 (dendritic marker, magenta), Synapsin 1 (Syn, Synaptic marker, Cyan) and DAPI (Blue). Red box indicates the area imaged in the high magnification images below. (**d**) Higher magnification equivalent to the area highlighted in red in (**c**). (**e**) Workflow for synaptic marker quantification. Molecular layer of CA1 hippocampal area was imaged by confocal microscopy. Z-stacks were imaged from top (higher Z step with specific signal) to bottom (last step with specific signal) with a Z-step of 0.5 m. The 40 middle steps were used for quantification. Confocal images were then processed with Huygens professional software for deconvolution. Fiji was used for quantification. Images were first thresholded to only select the specific signal. Images were then binarized and quantification of size and density of synaptic markers was performed using the built-in “Analyze particles”, with size exclusion threshold (as described in the Method section). Data were then compiled in open-

office and analyzed using Graphpad Prism software. (**f**) Heatmap summarizing the density of the different synaptic markers quantified in the CA1 hippocampal area from 1-month-old *Fus*$^{\Delta NLS/+}$ mice. Densities were normalized by the respective control. Mean value of each marker is indicated. Shade of color code for mean variation from 0 (white) to 2 (dark blue). *p<0.05. (**g**) Heatmap summarizing the density of the different synaptic markers quantified in the CA1 hippocampal area from 6-month-old *Fus*$^{\Delta NLS/+}$ mice. Densities were normalized by the respective control (*Fus*$^{+/+}$). Mean value of each marker is indicated. Shade of color code for mean variation from 0 (white) to 2 (dark blue). *p<0.05. (**h**) Heatmap summarizing the cluster area of the different synaptic markers quantified in the CA1 hippocampal area from 1-month-old *Fus*$^{+/+}$ and *Fus*$^{\Delta NLS/+}$ mice. Mean value of each marker is indicated. Shade of color code for mean variation from 0.01 (white) to 1 (dark red). *p<0.05. (**i**) Heatmap summarizing the cluster area of the different synaptic markers quantified in the CA1 hippocampal area from 6-month-old *Fus*$^{+/+}$ and *Fus*$^{\Delta NLS/+}$ mice. Mean value of each marker is indicated. Shade of color code for mean variation from 0.01 (white) to 1 (dark red). *p<0.05 **p<0.01.

Supplementary Fig. 4 Age-dependent alterations in the synaptic RNA profile of *Fus*$^{\Delta NLS/+}$ mouse cortex.

(**a**) Overlap between transcripts expressed in SNS RNA-seq and expressed genes in forebrain synaptic transcriptome reported previously[99]. Expressed genes are all genes with > 10 reads in 2/3 of the replicates (as defined previously[99]). (**b**) Plot of the first and second principal component of all RNA-seq samples and all expressed genes. The genotype is indicated by the symbol and the preparation and age by the color: 1-month-old mice in light and 6-month-old mice in dark colors. (**c**) Plot of the first and third principal component of all RNA-seq samples. (**d**) GO terms enriched among the significantly upregulated genes at 6 months of age in synaptoneurosomes of *Fus*$^{\Delta NLS/+}$ compared to *Fus*$^{+/+}$. (**e**) Gene ontology (GO) terms enriched among the significantly increased RNAs at 6 months of age in synaptoneurosomes of *Fus*$^{\Delta NLS/+}$ compared to *Fus*$^{+/+}$ (**f**) Heatmap from the set of up- and downregulated genes between total cortex samples from *Fus*$^{\Delta NLS/+}$ and *Fus*$^{+/+}$ at 6 months of age. Genes are on the rows and the different total cortex samples on the columns. The color scale indicates the log2FC between the CPM of each sample and mean CPM of the corresponding *Fus*$^{+/+}$ samples at each time point [sample logCPM – mean (logCPM of *Fus*$^{+/+}$ samples)]. (**g**) Volcano plots showing the log2 fold change of each gene and the corresponding -log10 (FDR) of the differential gene expression analysis comparing total cortex from *Fus*$^{\Delta NLS/+}$ to *Fus*$^{+/+}$ at 1 month (left panel) and 6 months (right panel) of age. The horizontal line marks the significance threshold of 0.05. Significantly downregulated genes are highlighted in green, upregulated genes in purple.

Supplementary Fig. 5. FUS peak locations on presynaptic and transsynaptic FUS RNA targets altered in *Fus*$^{\Delta NLS/+}$ mice.

CLIP-traces showing FUS binding on (**a**) *Syp* (**b**) *Robo2* (**c**) *Sv2a* (**d**) *Syt1* (**e**) *Chl1* (**f**) *App* (**g**) *Aplp2*

Supplementary Figure 6. FUS peak locations on postsynaptic FUS RNA targets altered in *Fus*$^{\Delta NLS/+}$ mice.

CLIP-traces showing FUS binding on (**a**) *Gria2* (**b**) *Gria3* (**c**) *Atp1a1* (**d**) *Atp1a3* (**e**) *Atp1b1* (**f**) *Spock1* (**g**) *Spock2* (**h**) *Clstn1*

Supplementary Figure 7. FUS binding on *Gabra1* RNA.

CLIP-traces showing FUS binding to the long 3'UTR containing isoform of *Gabra1*

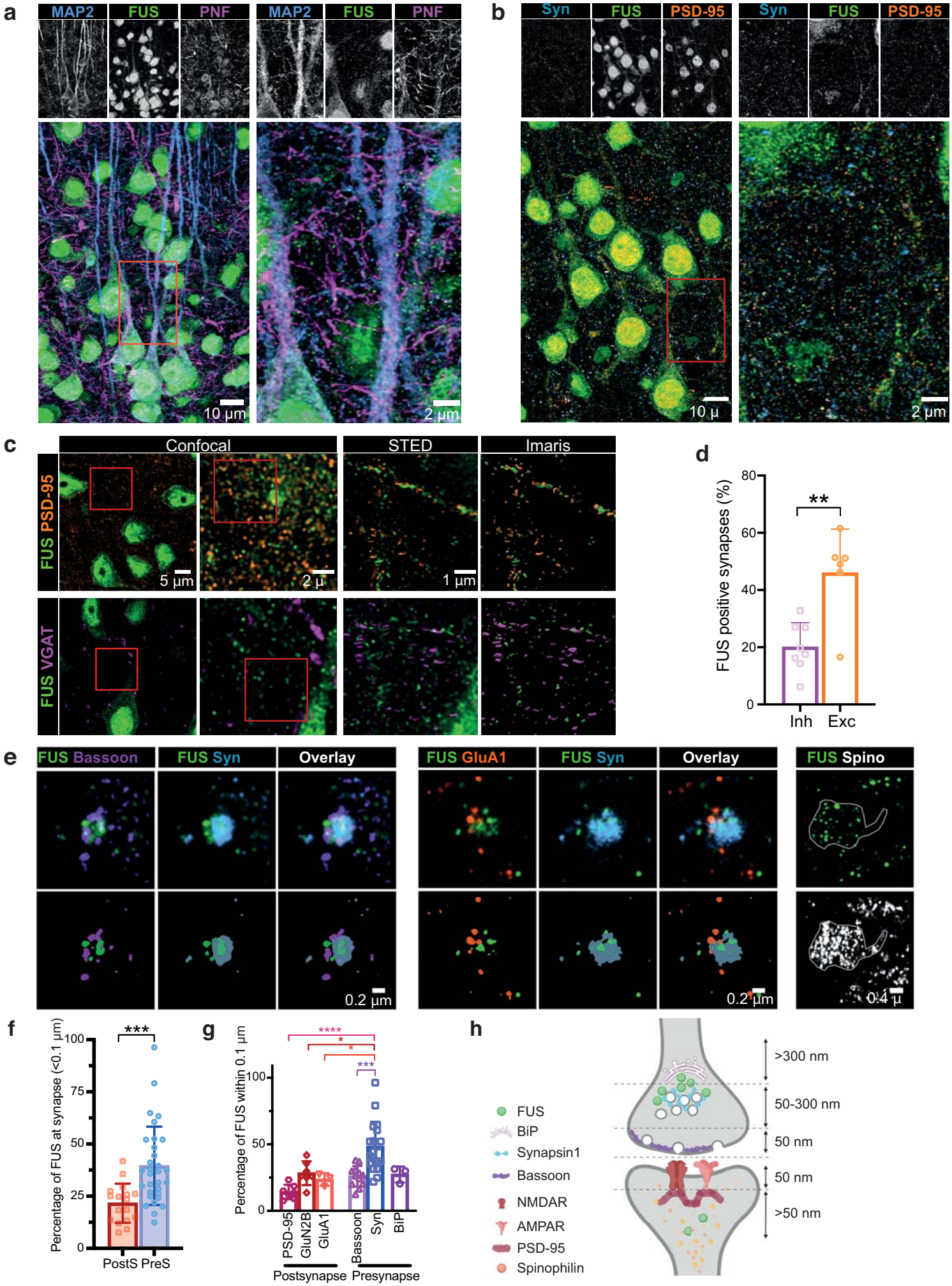


Fig 1. FUS is enriched at the presynaptic compartment

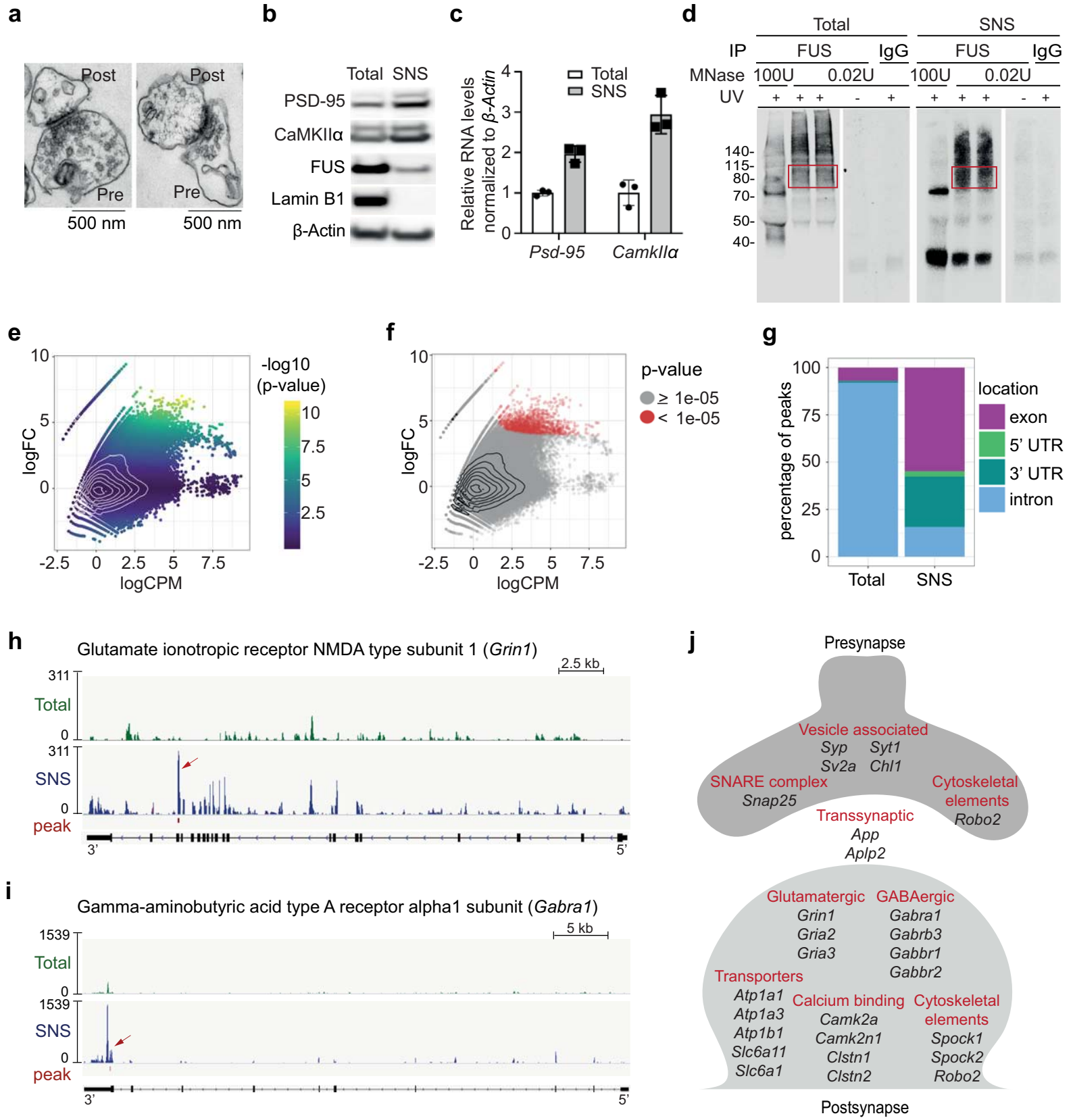


Fig. 2 CLIP-seq on cortical synaptoneurosomes identified FUS-associated pre- and postsynaptic RNAs

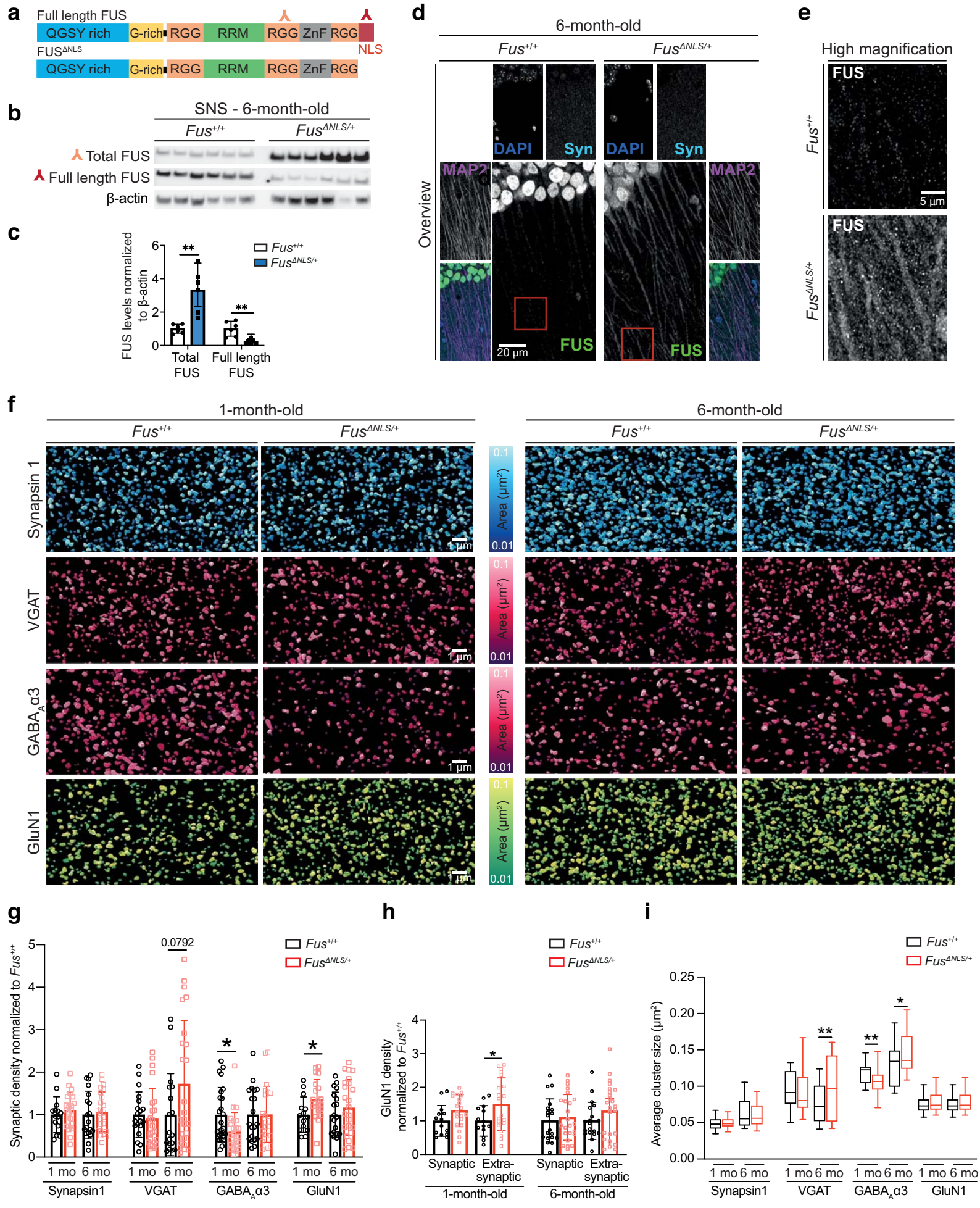


Fig 3. Increased synaptic FUS localization in *Fus*$^{\Delta NLS/+}$ mice affect GABAergic synapses

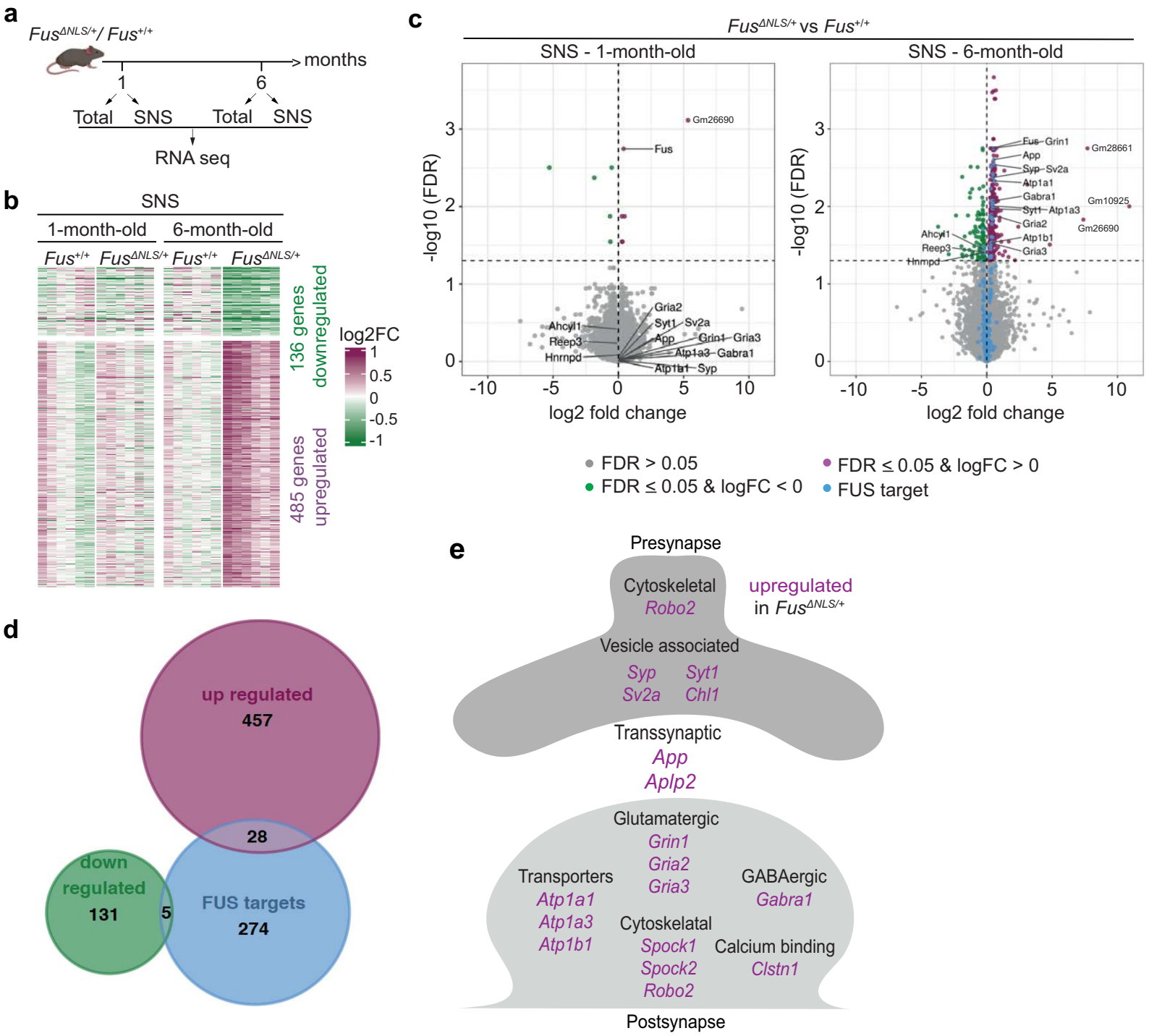


Fig. 4 Age-dependent alterations in the synaptic RNA profile of *Fus*$^{\Delta NLS/+}$ mouse cortex

Target sequences	Motif	p value	% of targets	% of background
Total cortex, intron		1e-17	7.23	1.39
SNS, intron		1e-13	3.67	0.15
SNS, intron		1e-12	13.84	4.24
SNS, 3' UTR		1e-17	7.23	1.39
SNS, exon		1e-40	10.83	1.46
SNS 5' UTR		1e-15	6.21	0.11

Table 1: FUS binds GU-rich sequences at the synapse
Predicted sequence motifs (HOMER) in windows of size 41 centered on the position with maximum coverage in each peak. Each set of target sequences has a corresponding background set with 200,000 sequences without any CLIP-seq read coverage (they are not bound by FUS). Note: These are all motifs that were not marked as possible false positives by HOMER and that occur in more than 1% of the target sequences.

Synaptic density analysis

unpaired t-test	1 month			6 months		
	p value	t, df	sample size	p value	t, df	sample size
Synapsin1	0.4556	t=0.7553, df=32	+/+=14 ΔNLS/+=20	0.6812	t=0.4138, df=41	+/+=19 ΔNLS/+=24
SNAP25	0.5320	t=0.6319, df=32	+/+=14 ΔNLS/+=20	0.085	t=1.765, df=41	+/+=19 ΔNLS/+=24
Bassoon	0.5821	t=0.5567, df=28	+/+=18 ΔNLS/+=12	0.4460	t=0.7708, df=35	+/+=18 ΔNLS/+=19
VGAT	0.6368	t=0.4758, df=40	+/+=19 ΔNLS/+=23	0.0792	t=1.801, df=40	+/+=18 ΔNLS/+=24
GluN1	0.0219	t=2.409, df=32	+/+=14 ΔNLS/+=20	0.3786	t=0.8900, df=41	+/+=19 ΔNLS/+=24
GluA1	0.6009	t=0.5292, df=28	+/+=18 ΔNLS/+=12	0.4885	t=0.7000, df=35	+/+=18 ΔNLS/+=19
pCaMKII	0.9055	t=0.1195, df=40	+/+=19 ΔNLS/+=23	0.2160	t=1.257, df=40	+/+=18 ΔNLS/+=24
Gephyrin	0.9878	t=0.1531, df=88	+/+=43 ΔNLS/+=47	0.5778	t=0.5591, df=74	+/+=34 ΔNLS/+=42
GABAARα1	0.1368	t=1.514, df=46	+/+=24 ΔNLS/+=24	0.9611	t=0.04906, df=44	+/+=20 ΔNLS/+=26
GABAARα3	0.0156	t=2.512, df=46	+/+=24 ΔNLS/+=24	0.9744	t=0.03234, df=40	+/+=20 ΔNLS/+=22

Table 2. Statistical analysis of synaptic density

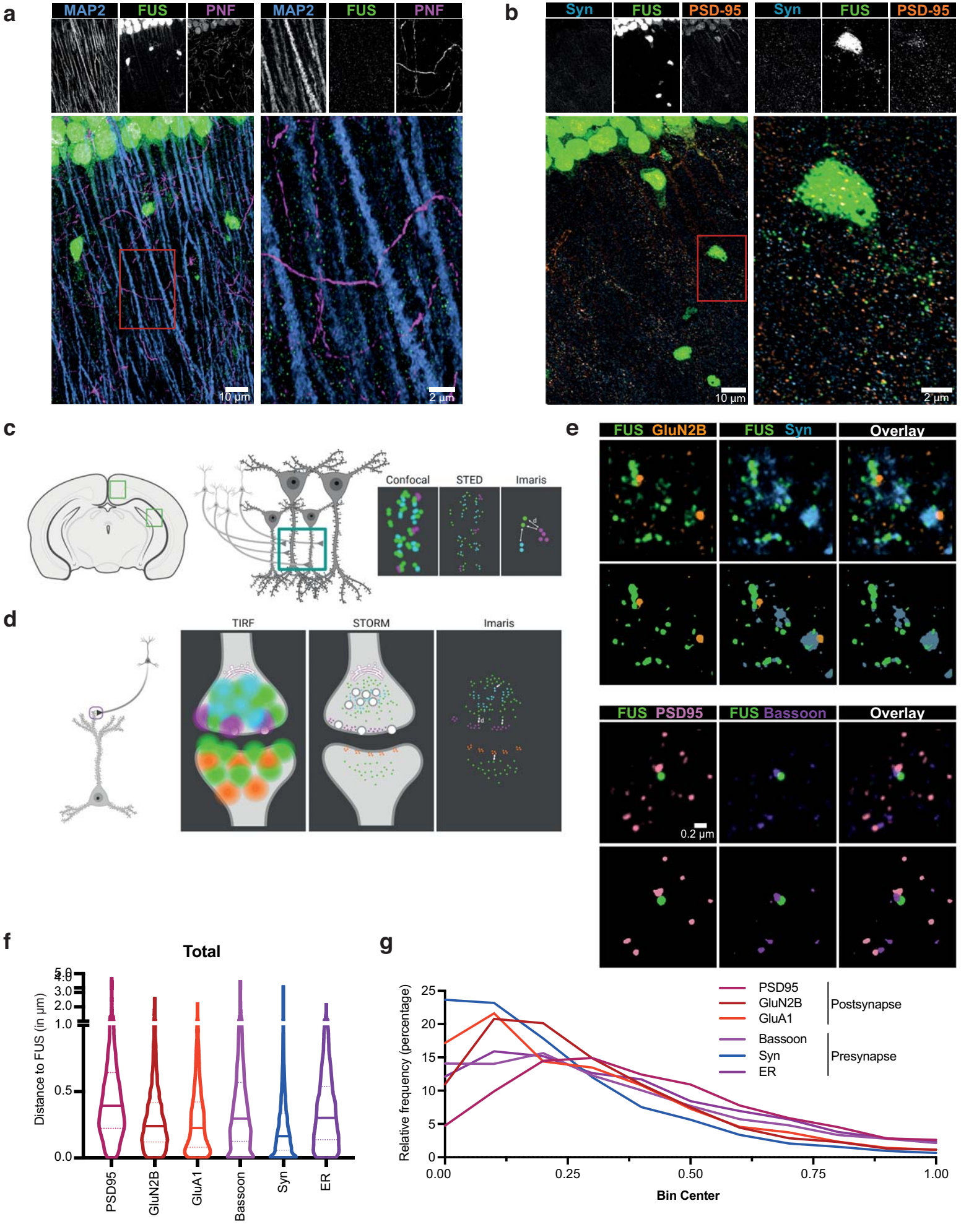
The table reports statistical analysis of density of the synaptic markers analyzed from a minimum of 2 images from at least 4 animals per genotype (*Fus*$^{+/+}$ and *Fus*$^{\Delta NLS/+}$) at 1 and 6 months of age. Unpaired t-test statistics, p-values, specific t-distribution (t), degrees of freedom (DF) and sample size are listed.

Synaptic cluster area analysis

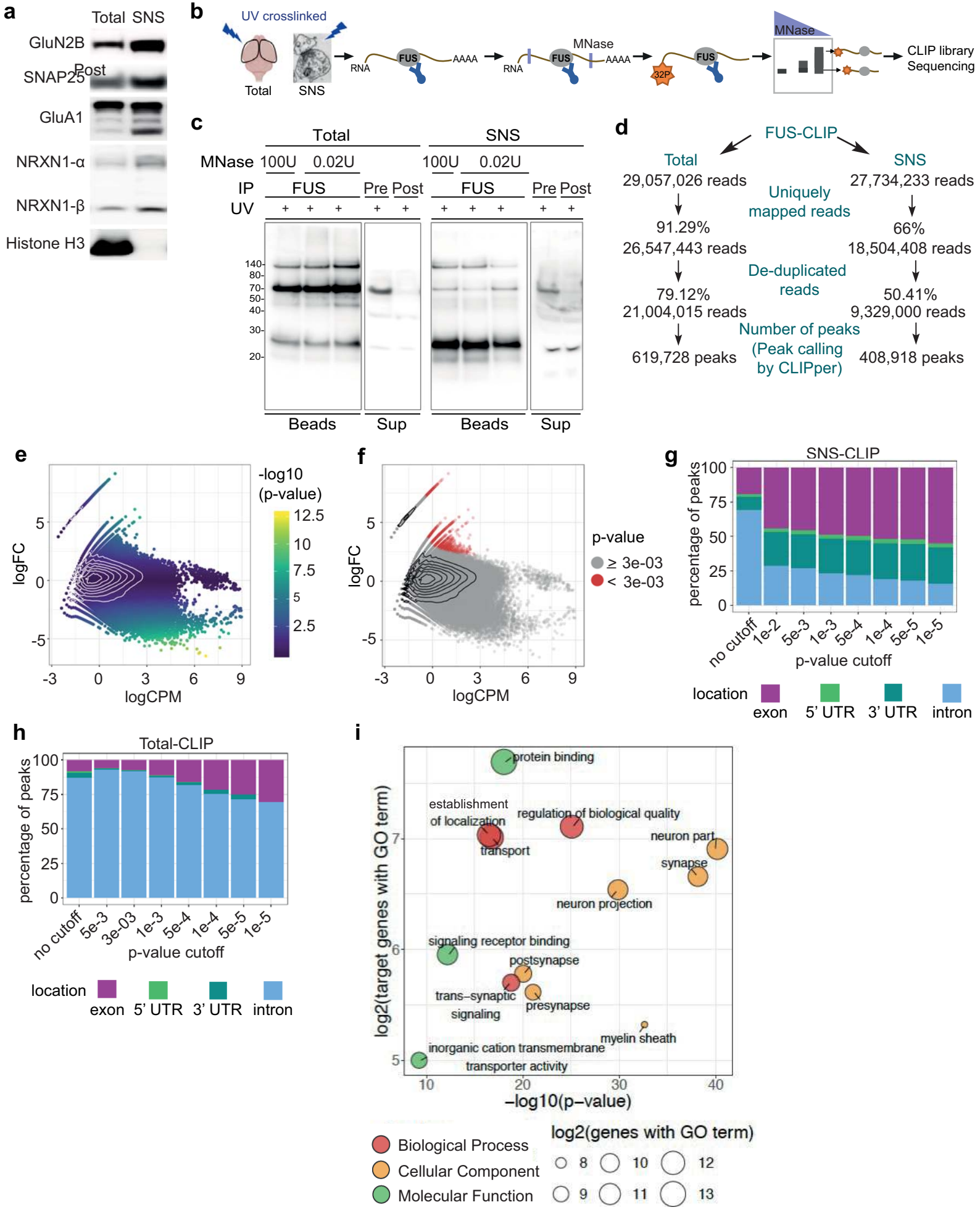
unpaired t-test	1 month			6 months		
	p value	t, df	sample size	p value	t, df	sample size
Synapsin1	0.8249	t=0.2214, df=363	+/+=14 ΔNLS/+=20	0.643	t=0.4639, df=393	+/+=19 ΔNLS/+=24
SNAP25	0.3834	t=0.8727, df=363	+/+=14 ΔNLS/+=20	0.5015	t=0.6727, df=393	+/+=19 ΔNLS/+=24
Bassoon	0.6022	t=0.5217, df=363	+/+=18 ΔNLS/+=12	0.7529	t=0.315 df=393	+/+=18 ΔNLS/+=19
VGAT	0.2819	t=1.078, df=363	+/+=19 ΔNLS/+=23	0.0028	t=3.005, df=393	+/+=18 ΔNLS/+=24
GluN1	0.5437	t=6078, df=363	+/+=14 ΔNLS/+=20	0.5694	t=0.5694, df=393	+/+=19 ΔNLS/+=24
GluA1	0.4303	t=0.7896, df=363	+/+=18 ΔNLS/+=12	0.4517	t=0.7533, df=393	+/+=18 ΔNLS/+=19
pCaMKII	0.242	t=1.172, df=363	+/+=19 ΔNLS/+=23	0.4150	t=0.8159, df=393	+/+=18 ΔNLS/+=24
Gephyrin	0.7467	t=0.3233, df=363	+/+=43 ΔNLS/+=47	0.2614	t=1.125, df=393	+/+=34 ΔNLS/+=42
GABAARα1	0.374	t=0.8902, df=363	+/+=24 ΔNLS/+=24	0.3204	t=0.9950 df=393	+/+=20 ΔNLS/+=26
GABAARα3	0.0053	t=2.807, df=363	+/+=24 ΔNLS/+=24	0.0166	t=2.407, df=393	+/+=20 ΔNLS/+=22

Table 3. Statistical analysis of synaptic cluster area

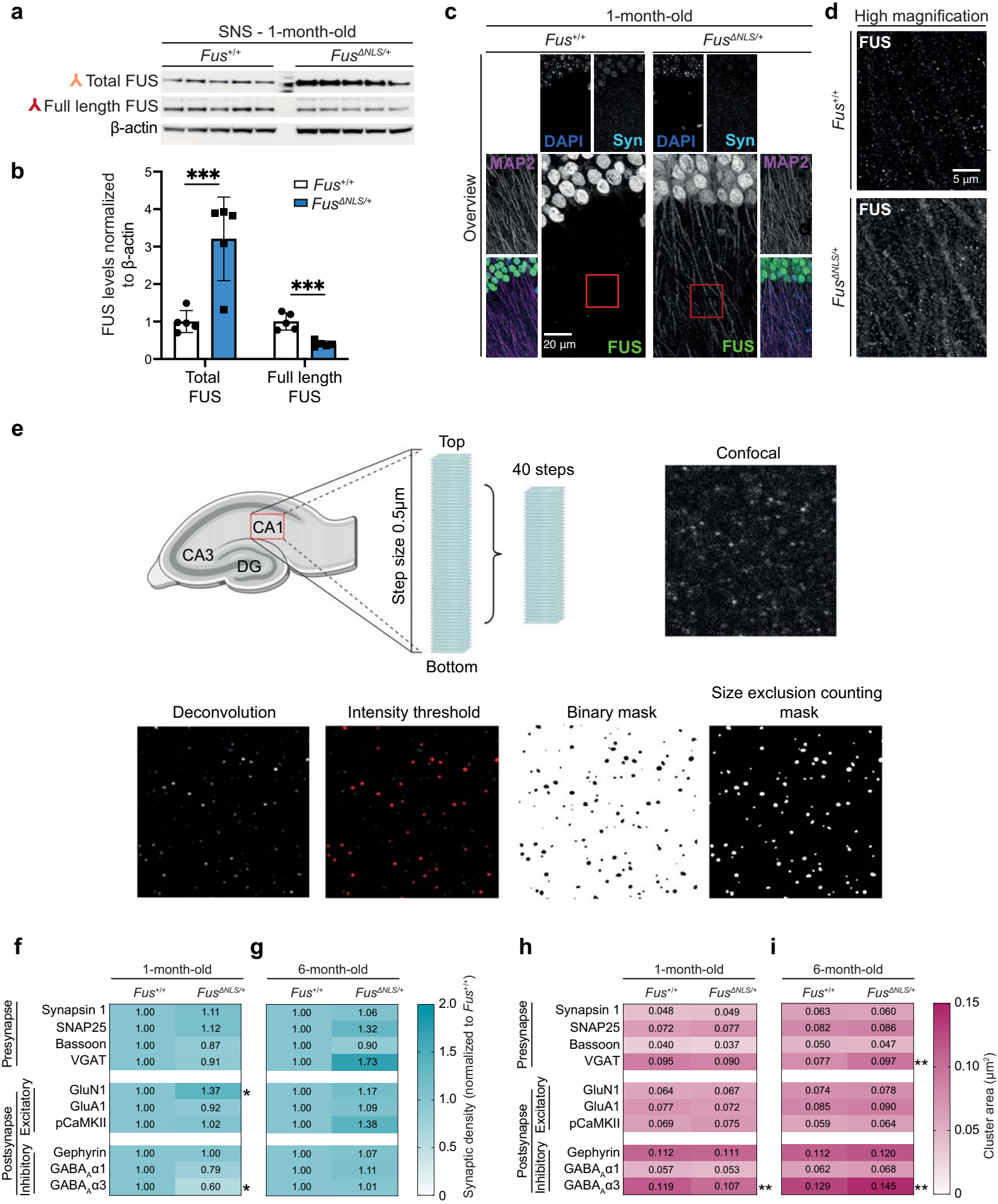
The table reports statistical analysis of area of the synaptic markers analyzed from a minimum of 2 images from at least 4 animals per genotype (*Fus*$^{+/+}$ and *Fus*$^{\Delta NLS/+}$) at 1 and 6 months of age. Unpaired t-test statistics, p-values, specific t-distribution (t), degrees of freedom (DF) and sample size are listed.



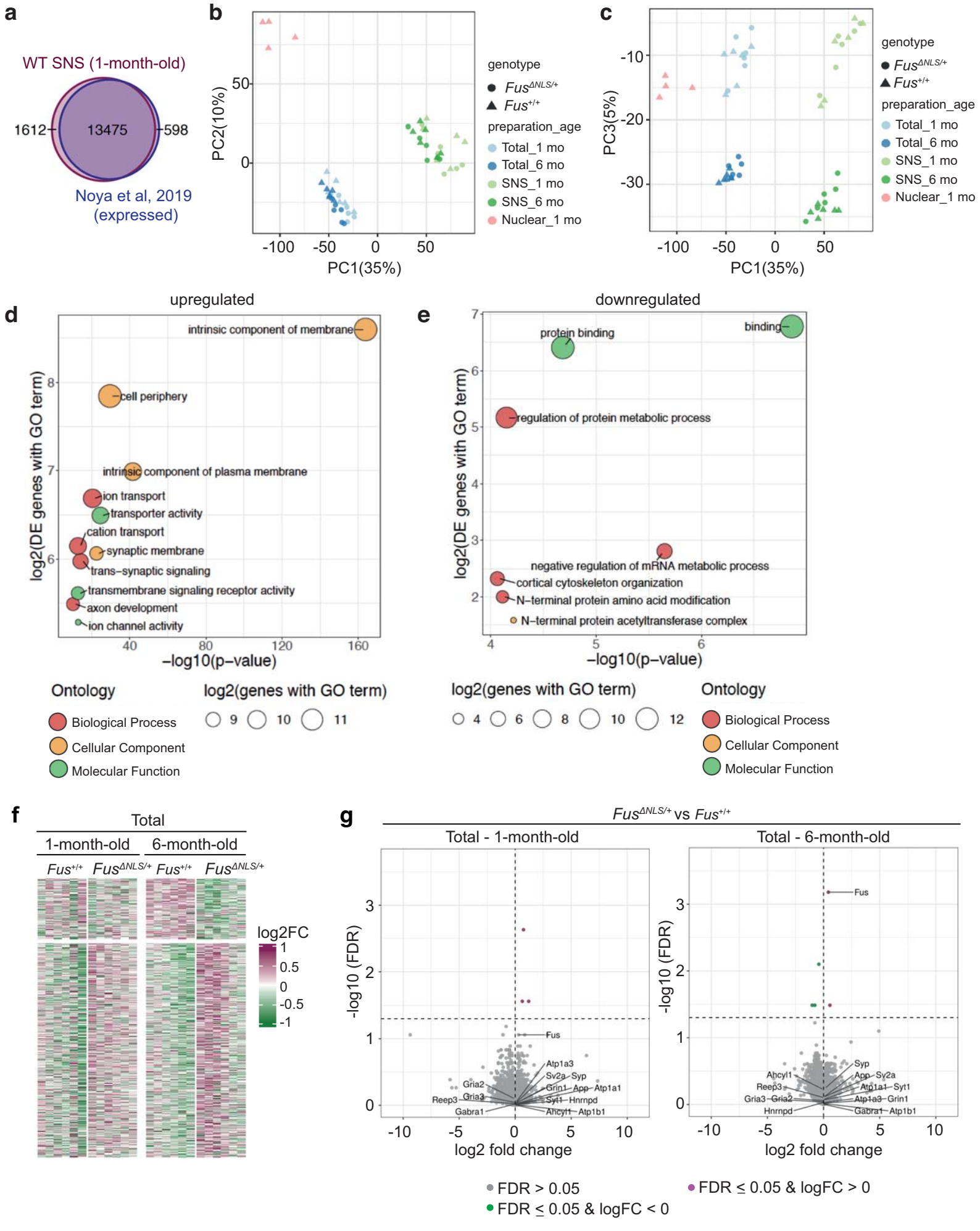
Supplementary Fig. 1 FUS is enriched at the presynaptic compartment



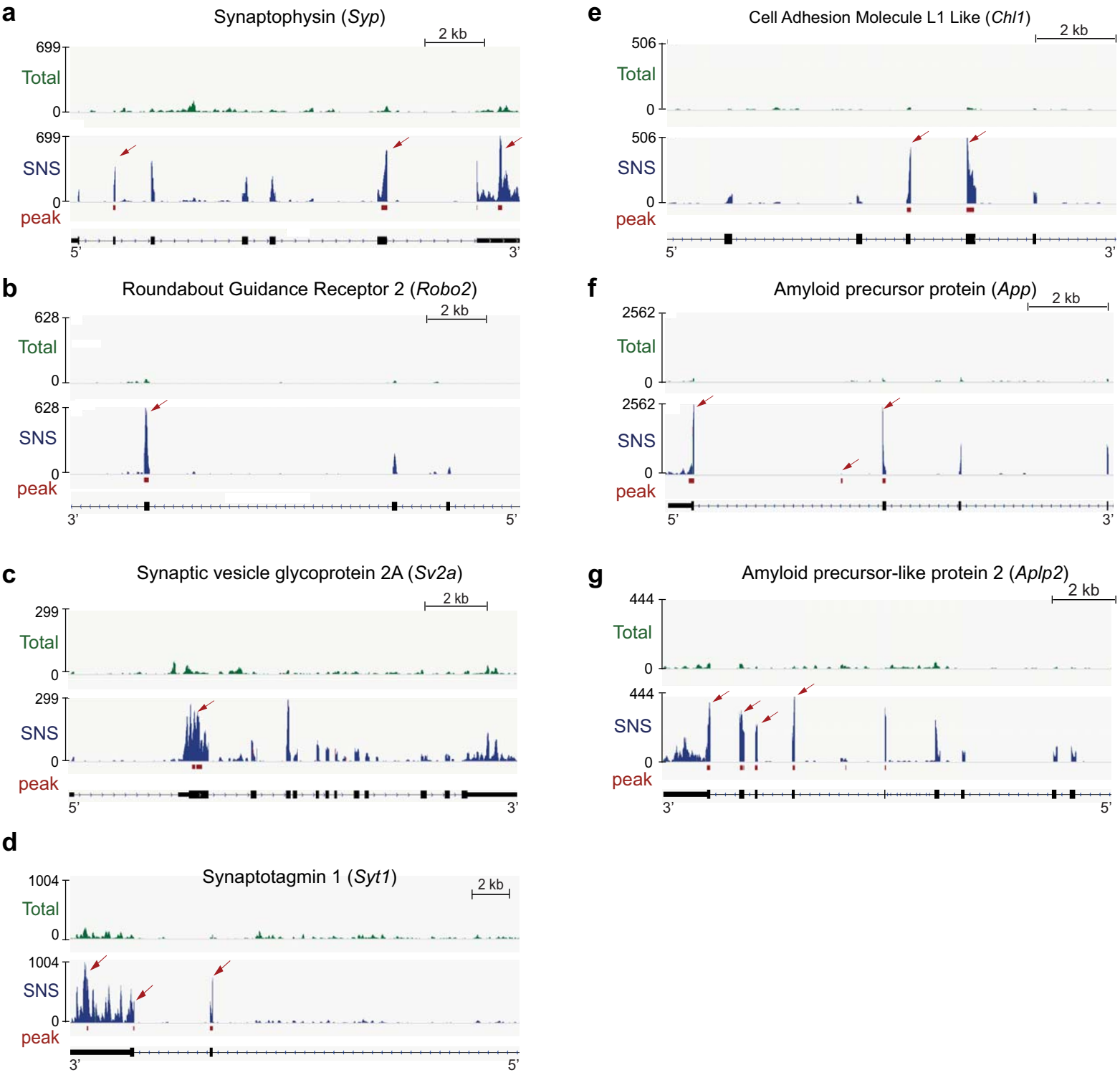
Supplementary Fig. 2 CLIP-seq on cortical synaptoneurosomes identified FUS-associated pre- and postsynaptic RNAs



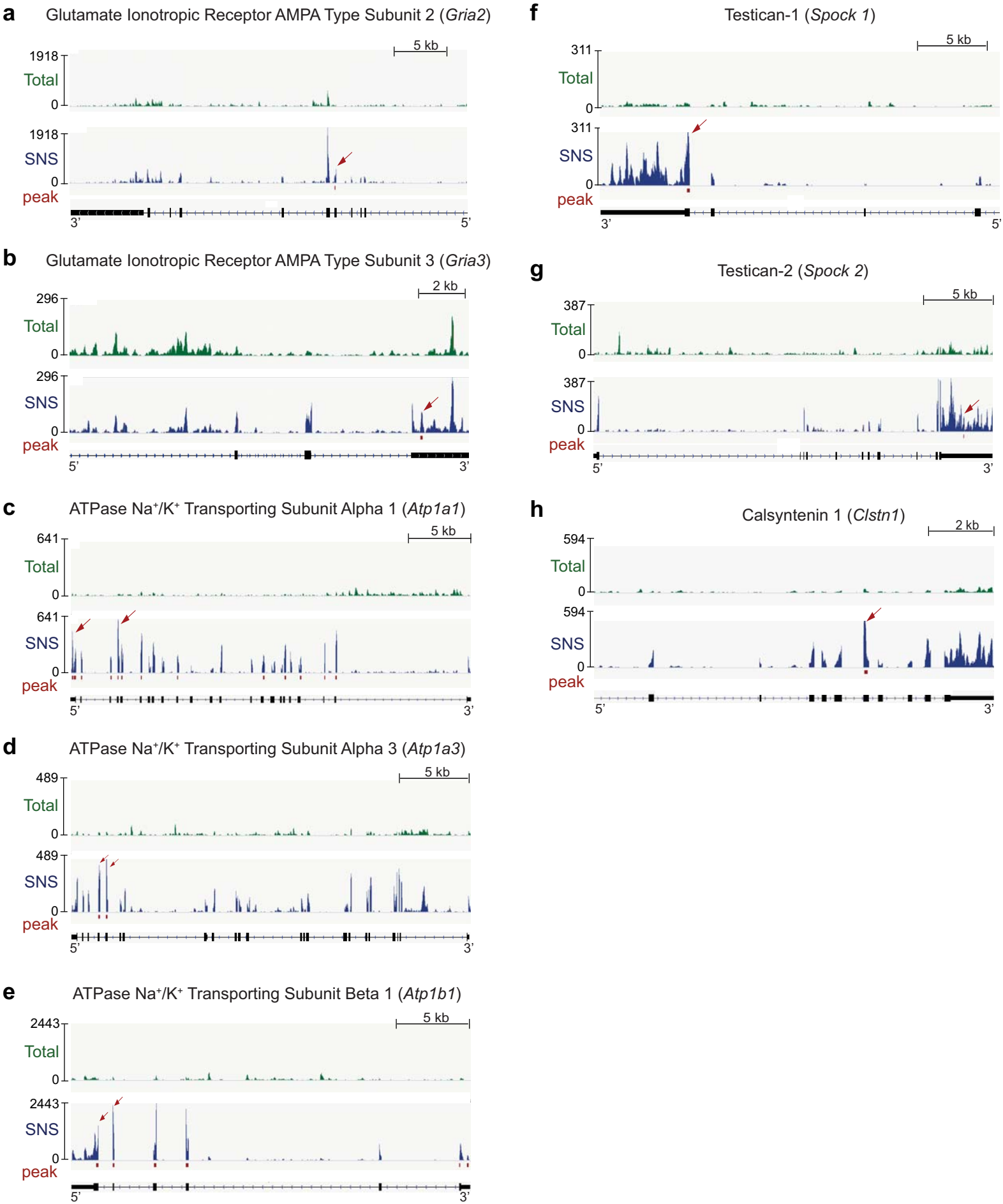
Supplementary Fig. 3 Age-dependent alterations in the synaptic RNA profile of *Fus*$^{\Delta NLS/+}$ mouse cortex



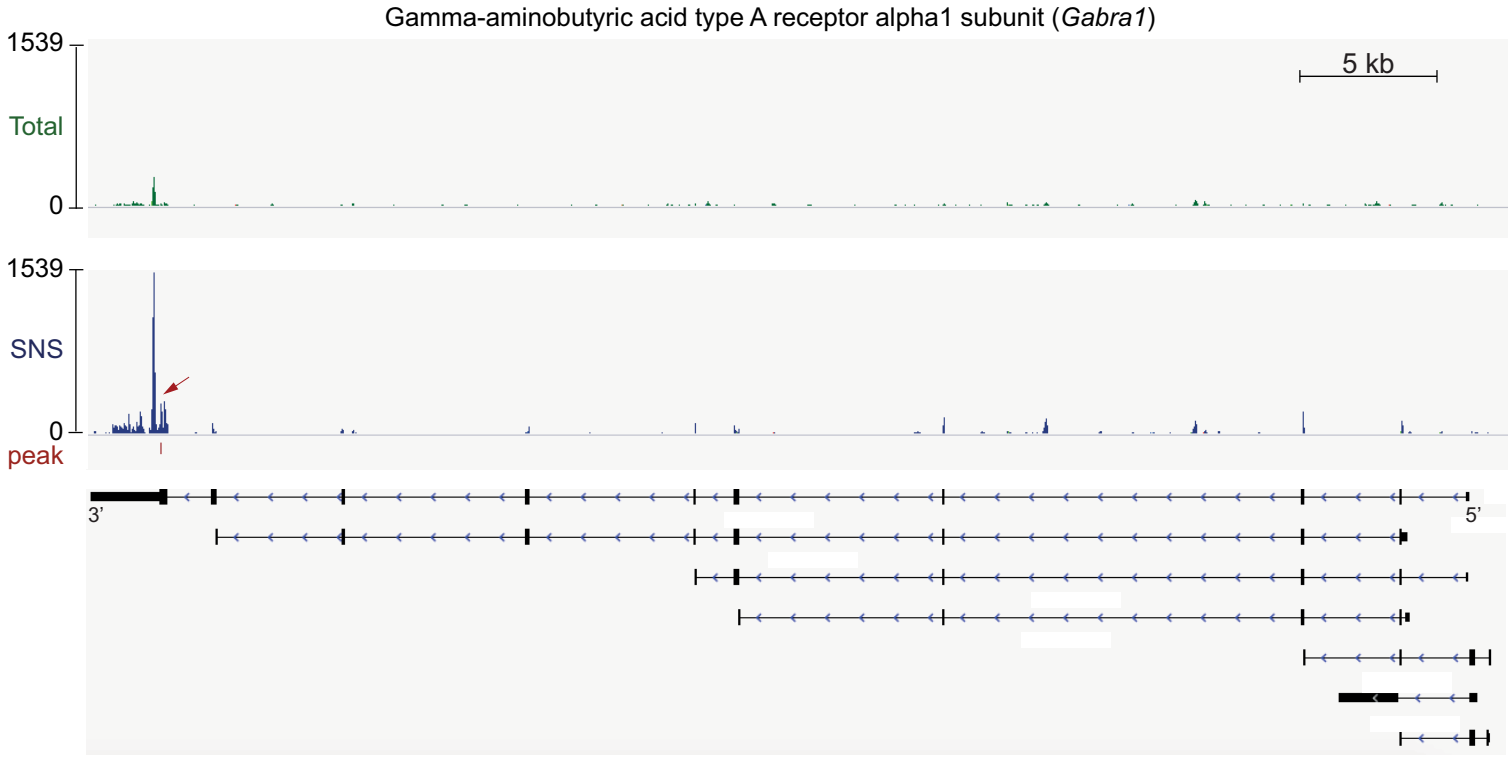
Supplementary Fig. 4 Age-dependent alterations in the synaptic RNA profile of *Fus*$^{\Delta NLS/+}$ mouse cortex



Supplementary Fig. 5 FUS peak locations on postsynaptic FUS RNA targets altered in *Fus*$^{\Delta NLS/+}$ mice



Supplementary Fig. 6 FUS peak locations on postsynaptic FUS RNA targets altered in *Fus*$^{\Delta NLS/+}$ mice



Supplementary Fig. 7 FUS binding on *Gabra1* RNA

5 DISCERNS: a pipeline for the DISCovery of unannotated Exons from RNa-seq Splice junction reads[1]

In this paper, we present DISCERNS an R package for the discovery of unannotated exons in RNA-seq data. We utilize information from splice-junction reads to identify the genomic coordinates of novel exons and their neighbouring exons in the corresponding transcripts. We test the performance of DISCERNS on simulated data and show that we can identify previously published novel splicing events in real RNA-seq datasets. Vladimir Barbosa C. de Souza, Mark Robinson and I developed the simulation. Mark Robinson and I developed the novel exon prediction strategy. I implemented the R package and performed all analyses. I wrote the manuscript.

[1] Manuscript in preparation: Hembach, K. M., Barbosa C. de Souza, V., Polymendiou, M., & Robinson, M. D. "DISCERNS: a pipeline for the DISCovery of unannotated Exons from RNa-seq Splice junction reads."

DISCERNS: a pipeline for the DISCovery of unannotated Exons from RNa-seq Splice junction reads

Katharina M. Hembach[1,2,*], Vladimir Barbosa C. de Souza[1], Magdalini Polymenidou[2] & Mark D. Robinson[1,*]

[1] *Department of Molecular Life Sciences and SIB Swiss Institute of Bioinformatics, University of Zürich, Switzerland*

[2] *Department of Quantitative Biomedicine, University of Zürich, Switzerland*

[*] *Correspondence to Katharina M. Hembach (katharina.hembach@uzh.ch) and Mark D. Robinson (mark.robinson@imls.uzh.ch)*

Abstract

Motivation: Alternative splicing is a tightly regulated process that generates diverse transcriptomes from a limited set of genes. Defects in alternative splicing can lead to aberrant transcripts that are often associated with diseases. Unfortunately, the exact splicing changes are often not well studied or even known. While various tools for the detection and quantification of alternative splicing have been developed, most are not able to predict novel splicing events or provide functionality to extend existing annotation catalogs with predicted exons.

Results: We developed a pipeline for the discovery of novel exons from RNA-seq data. We simulated an RNA-seq dataset with known novel splicing events (unknown to the mapping algorithms) to evaluate whether STAR and hisat2, over a range of parameter combinations, can effectively discover novel splice junctions. We present DISCERNS, an R package for the prediction of unannotated exons using information from reads with novel splice junctions. DISCERNS has higher precision on the simulated data than StringTie. We show that DISCERNS correctly predicts known and published cryptic exons in real RNA-seq data sets.

Availability: The code for the simulation, mapping comparison and exon discovery evaluation can be accessed from `https://github.com/khembach/novel_exon_pipeline`. The DISCERNS R-package is available from `https://github.com/khembach/DISCERNS`.

Contact: katharina.hembach@uzh.ch

1 Introduction

Alternative splicing is an important phenomenon that creates a vast diversity of observed transcripts from a limited pool of genes. Most transcribed pre-mRNAs are processed by the spliceosome, a complex of different proteins and RNAs, that removes intronic sequences and connects the exons to form the final mRNA molecule; thus, a single pre-mRNA can give rise to many different mRNAs through a process termed alternative splicing. Alternative splicing is regulated by splicing factors, RNA-binding proteins (RBPs), that bind to regulatory sequences in the pre-mRNA and thereby determine which splice sites are recognized by the spliceosome. The six most common types of alternative splicing are [1,2]:

1. Cassette exons/exon skipping: an already-defined exon that is spliced out of a transcript
2. Alternative 3' acceptor splice sites: the usage of different 3' acceptors results in alternative 5' starts of the downstream exon
3. Alternative 5' donor splice sites: the usage of a different 5' donor results in a different 3' end of the upstream exon
4. Mutually exclusive exons: two exons that never occur together in an mRNA
5. Intron retention: an intron that is not spliced out but contained in the mature mRNA
6. Alternative 3' (or 5') UTR usage: the usage of different 3' (or 5') UTRs results in different first or last exons in a transcript

A related phenomenon are microexons [3], which are very short exons that are extensively regulated and alternatively spliced, for example, in neurodevelopment [4]. In particular, the inclusion or exclusion of a microexon can modulate the structure of a protein domain or interaction site. In this paper, we adopt the microexon definition from Irimia *et al.*[4] of 3-27 nucleotides.

Alternative splicing directly determines which proteins can be produced in a cell by generating the mRNA templates for translation. Complex regulatory networks of cis elements and trans-acting factors, as well as different steps in the spliceosome assembly [5], control splicing and ensure that a cell can fulfill its function under normal conditions but also that it can react to stress. Consequently, defects in splicing are associated with many different diseases. For example, in amyotrophic lateral sclerosis (ALS), the cytoplasmic aggregation of the RBP TAR DNA-binding protein 43 (TDP-43) results in nuclear loss of function leading to splicing defects and the inclusion of cryptic exons in

mRNAs [6]. In autism spectrum disorder (ASD), reduced SRRM4 levels lead to a downregulation of microexons [4]. In cancer, mutations in RBP binding sites can disrupt the motif and affect RNA expression and splicing of cancer-driver genes [7]. On the protein level, alternative splicing changes can add or remove protein domains and disrupt protein-protein interactions in cancer-related pathways [8].

Fortunately, rapid improvements in the accuracy, length and depth of (predominantly Illumina) cDNA library sequencing (commonly known as RNA-seq) means that alternative splicing events can be widely observed. However, the vast majority of RNA-seq analysis pipelines depend on annotation catalogs, independent of how they quantify gene or exon expression (genome alignment or transcript-level estimation). As shown by Soneson *et al.*[9], even well annotated genomes such as human or mouse contain genes with low agreement between annotated and observed splice junctions (SJ). Consequently, the selected reference annotation can have a strong influence on the quantification and ultimately all downstream analyses such as differential gene expression or transcript usage. All unannotated events will be missed by pipelines that only rely on annotation and that do not predict novel splicing events or assemble the transcriptome. Thus, there is clearly a need to extend current catalogs according to the novel events seen in a data set.

Long-read transcriptomic sequencing (mainly PacBio and Oxford Nanopore) is an emerging technology that, in theory, enables the sequencing of full-length transcripts. However, currently the technologies have much lower sequencing depth, higher error rates and higher cost per base than "short read" RNA-seq [10]. Fragmentation of the cDNA during library preparation and early termination while sequencing lead to biases in transcript detection [11]. Therefore, short read RNA-seq is still the main technology to study alternative splicing.

Currently, there are different tools available that can detect and/or quantify alternative splicing (Table 1). Most tools focus on the detection and quantification of annotated alternative splicing events. Some tools, such as ASGAL, SplAdder or Whippet, can predict (specific types of) alternative splicing events and in theory, they are able to find novel unannotated exons. However, none of tools will output a list with the predicted novel exons. This information has to be parsed from the output files that are not standardized and vary between tools. Transcriptome assembly methods are not specialized to detect alternative splicing or novel exons, but they can detect all splicing events present in the analysed samples. Thus, the final transcripts can in principle include unannotated events. The prediction of microexons poses an additional problem since they are often misaligned or completely missed due to their small size. To our knowledge, none of the

Table 1: Comparison of alternative splicing (AS) analysis methods. AS, alternative splicing; Tr., transcriptome.

Method	AS detection	AS quantification	AS prediction	Tr. assembly	GTF extension	Input	Output
Yanagi [12]	+	+	-	-	-	GTF, genome FASTA, FASTQs	table with PSI counts
MISO [13]	-	+	-	-	-	BAM, GFF with alternative events	table with counts and PSI
rMATS [14]	-	+	-	-	-	FASTQ or BAM, GTF	table with event counts
SUPPA2 [15]	+	+	-	-	-	GTF, transcript expression	event specific GTF
JUM [16]	-	+	+	-	-	BAM, SJ.out.tab (STAR) of replicates	table with PSI per events
VAST-TOOLS [4]	+	-	-	-	-	FASTQ, VAST-DB	tabel with counts and PSI per event
ASGAL [17]	+	-	+	-	-	raw FASTQ (single-end), GTF	table with event counts
SpliceGrapher [18]	-	-	+	-	-	SAM, GTF	GFF with splice graph
SGSeq [19]	+	+	+	-	-	BAM (GFF/TxDb optional)	R objects with event counts and PSI
SplAdder [20]	+	+	+	-	-	BAM, GTF	GFF per event + table with event counts
Whippet [21]	+	+	+	-	-	GTF (BAM optional)	table: PSI per node; exon coordinates (novel are labeled)
StringTie [22]	-	-	-	+	+	SAM (GTF optional)	GTF with assembled transcripts
Scripture [23]	-	-	-	+	-	SAM, genome FASTA	BED with transcripts, transcript graph (.dot)
Cufflinks [24]	-	-	-	+	+ (cuffcompare)	BAM, GTF	GTF with assembled transcripts
DISCERNS	-	-	+	-	+	BAM, SJ.out.tab (STAR)	table with novel exons; extended GTF (optional)

alternative splicing analysis methods outputs genome annotations (GTF or GFF format) that are augmented with the predicted novel events. The only exception are transcriptome assemblers, such as StringTie or cufflinks.

Using simulated data with novel splicing events, we compared the two most common RNA-

seq genome alignment methods hisat2 [25] and STAR [26] in their ability to correctly align splice-junction reads. Furthermore, we developed DISCERNS, an R-package for the discovery of novel splicing events based on splice-junction RNA-seq reads from genome alignments. We show that DISCERNS outperforms StringTie on simulated data in terms of precision. In ALS related RNA-seq data sets, DISCERNS is able to systematically find known cryptic exons.

2 Materials and Methods

2.1 Reference data and simulation

We simulated an RNA-seq data set to evaluate the two most commonly-used read alignment methods STAR and hisat2, in their ability to detect novel splice junctions. Simulated data is essential for such an evaluation, because in real datasets, we do not know the true location of novel splice junctions. As reference, we used the human Ensembl genome sequence and gene annotation GRCh37.85 [27]. We simulated RNA-seq data set using RSEM [28] and quantifications from a single sample, ENCODE [29] biosample ENCBS049RNA (GEO accession GSM2072377), which contains stranded paired-end reads (101 bp) from rRNA depleted total RNA of human fetal (37 weeks) cerebellum tissue. We decided to simulate reads from only two chromosomes (19 and 22) to ensure fast runtimes. We chose chromosome 19 and 22 because they have the highest gene density of all human chromosomes [30]. Gene expression and RNA-seq parameters were estimated with `rsem-calculate-expression`. The resulting model file was modified by removing the probability of generating reads with quality 2 to prevent sampling of reads with overall low quality. Paired-end reads of length 101 from chromosomes 19 and 22 were created with `rsem-simulate-reads`. The fraction of reads that come from background noise, the theta parameter, was set to 0.05 and the number of reads to be simulated was set to 1,118,017 (number of reads mapped to chromosomes 19 and 22 by `rsem-calculate-expression` using STAR for alignment). The overall outline of the simulation is depicted in Figure 1.

2.2 Reduced gene annotation and novel exons

The RNA-seq data set was simulated using the full GTF file from Ensembl. We created three reduced GTF files, after removing specific exons from the annotation to test the performance of different algorithms. This enabled us to evaluate if the different methods were able to recover the removed (unannotated) splice junctions and exons. The different sets of unannotated novel

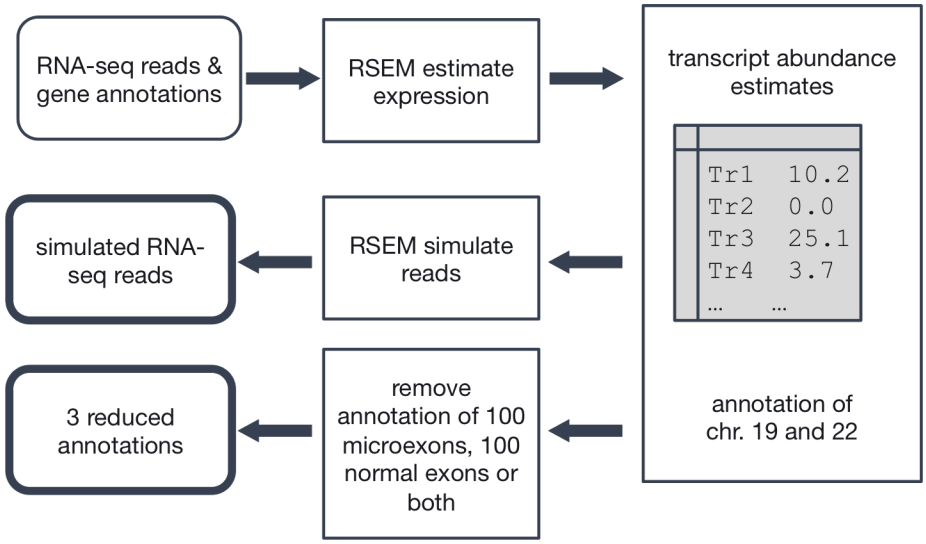


Figure 1: Simulation outline. The abundance of annotated transcripts in an RNA-seq sample is estimated with RSEM. RNA-seq reads for chromosome 19 and 22 are simulated with RSEM using the estimated abundances. Three different sets of exons/microexons are removed from the reference annotation to create reduced sets.

exons were created by removing random exons from the Ensembl GTF file. 100 microexons ($\leq$ 27 nts) and 100 exons ($>$ 27 nts) from chromosomes 19 and 22 were randomly chosen from the set of expressed exons, i.e. exons with at least one read in the simulated RNA-seq data set, using random seed 42. The two sets of 100 microexons and 100 exons were deleted from the reference annotation individually and together by removing the corresponding exons from all transcripts. This resulted in three reduced GTF files with:

1. 100 missing microexons (me)
2. 100 missing exons (exon)
3. 100 missing microexons and 100 missing exons (me_exon).

The only difference between the three reduced GTF files are the sets of missing exons. The three different sets of deleted exons (100 microexons, 100 exons, 100 microexons and 100 exons) define the sets of novel exons that can be identified in the simulated data when using either of the reduced annotations.

2.3 Read alignments

Reads were aligned with STAR version 2.5.3a [26] and hisat2 version 2.1.0 [25]. hisat2 was run with default parameters except for `-k`, which was set to 1 to search for a single primary alignment

Table 2: Additional STAR parameters.

Name	Additional parameter setting	Parameter description
default		no additional parameter
outSJfilterOverhangMin9	`--outSJfilterOverhangMin 30 9 9 9`	default: 30 12 12 12 Minimum overhang length for novel splice junctions on both sides for: (1) non-canonical motifs, (2) GT/AG and CT/AC motif, (3) GC/AG and CT/GC motif, (4) AT/AC and GT/AT motif.
outSJfilterOverhangMin6	`--outSJfilterOverhangMin 30 6 6 6`	see above
outSJfilterCountTotalMin3	`--outSJfilterCountTotalMin 3 3 3 3`	default: 3 1 1 1 Minimum total (multi-mapping+unique) read count per junction for: (1) non-canonical motifs, (2) GT/AG and CT/AC motif, (3) GC/AG and CT/GC motif, (4) AT/AC and GT/AT motif.
scoreGenomicLength	`--scoreGenomicLength`	default -0.25
Log2scale0	`Log2scale 0`	extra alignment score logarithmically scaled with genomic length of the alignment: `scoreGenomicLengthLog2scale *log2(genomicLength)`
alignSJoverhangMin3	`--alignSJoverhangMin 3`	default 5 minimum overhang (i.e. block size) for spliced alignments

per read. STAR was run with with `--outFilterMultimapNmax 1` to not allow read multimappings and `--outSJfilterDistToOtherSJmin 10 0 0 0` to lower the minimum allowed distance between a novel splice junction with canonical motif and other junctions donor/acceptor to 0. All other STAR parameters were left at default values, except for some additional parameters that were varied one at a time (Table 2). All STAR alignments were run in 2-pass mode, which means that after read mapping, the list of novel splice junctions is inserted in the genome index and all reads are mapped a second time against the new genome index. The hisat2 output files in SAM format were converted to BAM files and all BAM files were sorted and indexed with samtools version 1.4.1 [31]. Each read alignment was performed three times using the three different reduced GTF files (see Section 2.2) resulting in three different BAM files per mapper and parameter combination.

2.4 Computation of precision, recall and F1 score for the mapping comparison

The true genomic start and end coordinates of splice junctions were determined for each of the simulated reads. For each of the BAM files, true positive (TP), false positive (FP), true negative (TN) and false negative (FN) splice junctions were computed by comparing the splice junctions in the mapped reads to the true splice junctions (Figure 4A). A TP is an observed splice junction that is also present in the same read in the truth; a FP splice junction is not present in the truth; a TN is a read without a splice junction in both the mapping and the truth; and a FN splice junction is only observed in the truth but not the mapping. Precision was calculated as $TP/(TP+FP)$, recall as $TP/(TP+FN)$, and the F1 score as $2*(precision*recall)/(precision+recall)$. All scores were separately computed for the first and second read in the pair.

2.5 Novel exon classification

All novel exons in the simulation (see Section 2.2) are defined by their genomic start and end position, as well as the end of the upstream exon and the start of the downstream exon in a transcript. We classified the novel exons based on their number of novel splice junctions (see Figure 2A) and thus the identification difficulty:

- novel cassette exons with two novel splice junctions: easy
- alternative 3' or 5' splice site or terminal exons (first or last exon in a transcript) with one novel splice junction: medium
- novel exon without a novel splice junction because the splice junctions are shared with annotated exons: complicated

2.6 Novel exon discovery

We developed DISCERNS, an R package for the discovery of novel exons. It depends on the SJ.out.tab file from the STAR read alignment, the corresponding BAM file (optional) and the GTF annotation file. We developed different strategies to predict novel exons from each of the three classes (see 2.5). In the following, we will explain the three strategies:

- Easy (Figure 2A): Novel cassette exons have two novel splice junctions and we can thus predict them using only the splice junctions. The novel splice junctions from the SJ.out.tab

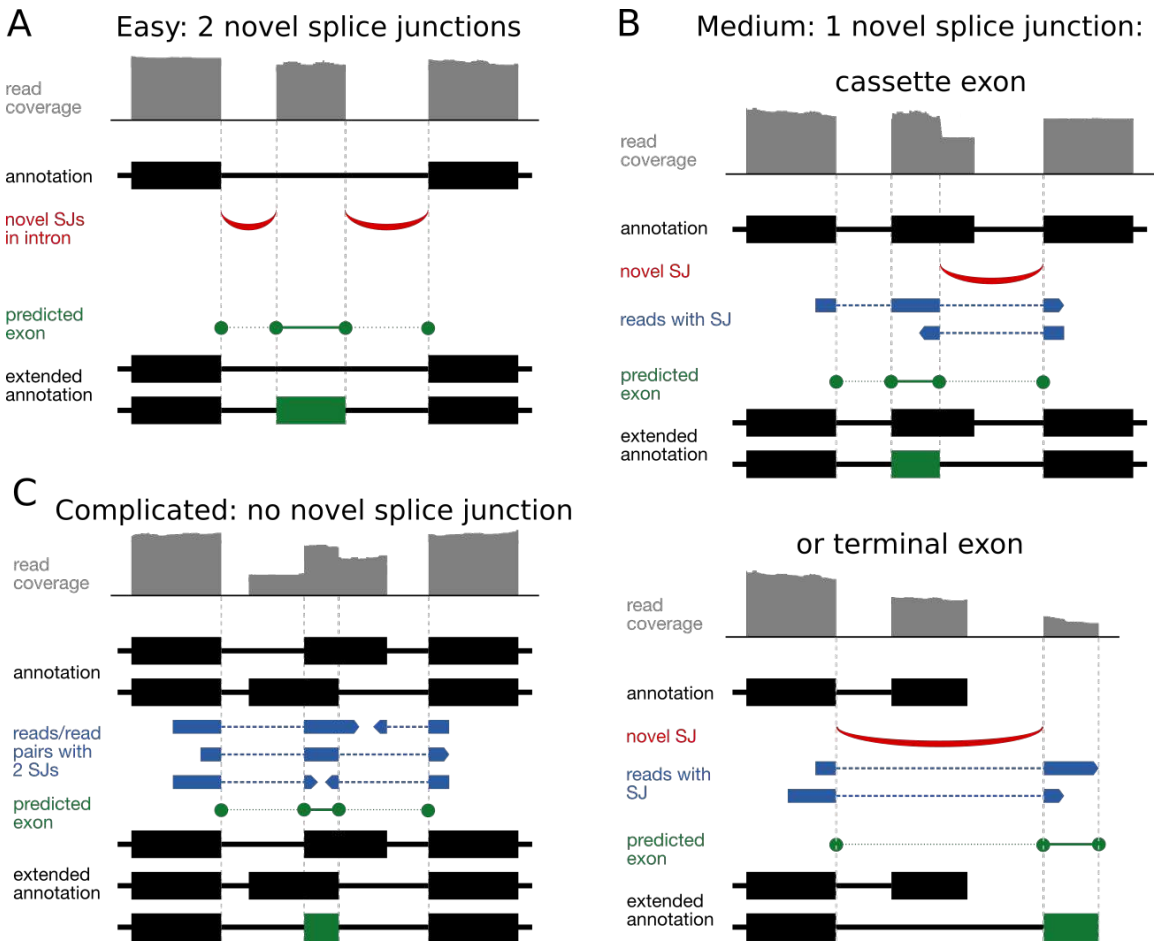


Figure 2: Schematic of the DISCERNS predictions for the three different novel exon classes. (A) Easy: Novel exon with two novel splice junctions. The novel exon (green) is defined by two novel splice junctions (orange) that splice to annotated exons (grey). This class of exons can be predicted with the `find_novel_exons()` function of our R-package and the parameters `single_sj = FALSE, read_based = FALSE`. The function uses the gene annotation and pairs of novel splice junctions to predict novel exons. The output are the coordinates of the novel exon and the up- and downstream connected exons. (B) Medium: Novel exons with only one novel splice junction; can be predicted with parameters `single_sj = TRUE, read_based = FALSE`. The function uses the gene annotation, the single novel splice junction and all reads that support the splice junction to predict the novel exon. For alternative 3' and 5' splice sites, the output are four coordinates and for terminal exons three. (C) complicated: Novel exons without a novel splice junction; can be predicted with parameters `single_sj = FALSE, read_based = TRUE`. Depending on the library type, the function scans all reads with two splice junctions (single-end) or pairs of reads with each one splice junction (paired-end). The splice junction pairs are compared to the gene annotation and novel splice junction combinations are reported as novel exons. The output are four coordinates of the novel exon and the up- and downstream connected exons.

file are compared with the intron annotations. Novel cassette exons are predicted from pairs of novel splice junctions that are located within an annotated intron and share the start and end coordinates of the intron.

- Medium (Figure 2B): Novel alternative 3' or 5' splice sites or terminal exons can have at

most one novel splice junction. We use the unpaired novel splice junction (all splice junction that were not used to predict cassette exons in step 1) and filter all reads from the BAM file that support them. To predict alternative 3' or 5' ends, we search for reads that have a second splice junction in addition to the novel one. The boundaries of the predicted exon are then determined by the mapped read region between the two splice junctions. For terminal exons, we determine if the novel splice junction is located at the end (or start) of an annotated transcript. If there are no annotated exons downstream (or upstream) of the novel splice junction, then the splice junction defines a novel terminal exon. The boundaries of the terminal exon are determined by the novel splice junction and the longest supporting read.

- Complicated (Figure 2C): Novel exons without novel splice junctions because both splice junctions are shared with other overlapping exons. For this type of novel exon, we do not have novel splice junctions that tell us where we should start looking for a novel splicing event. Instead, we analyse all reads or read pairs with two splice junctions to identify novel splice junction combinations. The two splice junctions are compared to all transcripts that cover the genomic region. If there is no transcript that contains the two splice junctions sequentially, they define a potential novel exon. In case of paired-end reads with each one novel splice junction, the potential exons are filtered based on the distance between the two splice junctions. The distance between the end of the first junction and the start of the second junction must not exceed $< 2 * (\texttt{read\_length} - \texttt{overhang\_min}) + \texttt{min\_intron\_size}$, where `read_length` is the length of the reads, `overhang_min` is the minimum overhang length for splice junctions on both sides for canonical-motifs as defined by the `outSJfilterOverhangMin` parameter of STAR, and `min_intron_size` is the minimal required intron length as defined by the `alignIntronMin` parameter of STAR. For example, paired-end reads with a length of 101 nts and a minimal overhang of 6 and a minimal intron length of 21 allow a distance of at most 211 nucleotides between the two splice junctions: $2 * (101 - 6) + 21 = 211$. If the distance between the two splice junctions exceeds the limit, it cannot be guaranteed that the junctions are connected to the same exon. Only the predicted exons that are located within the boundaries of an annotated gene are reported. This filtering step prevents false positive predictions from wrongly mapped reads.

The final output of our prediction method is a table with the genomic coordinates of all pre-

dicted novel exons. Each row is a predicted exon. The columns are the chromosome (`seqnames`), the end of the upstream exon (`lend`) in the transcript, the start and end of the predicted exon (`start` and `end`), the start of the downstream exon (`rstart`) in the transcript and the strand (`strand`). The next three columns are the number of reads supporting each of the two splice junctions that define the predicted exon: `unique_left` is the number of reads supporting the splice junction from `lend` to `start`, `unique_right` is the number of reads supporting the splice junction from `end` to `rstart`, and `min_reads` is the minimum of `unique_left` and `unique_right`. The last column is the ID (`ID`) of the predicted exon, i.e. a number between 1 and the total number of predictions. If the predicted exon is terminal, i.e. it is the first or last exon in a transcript, then `lend` or `rstart` are undefined (`NA`).

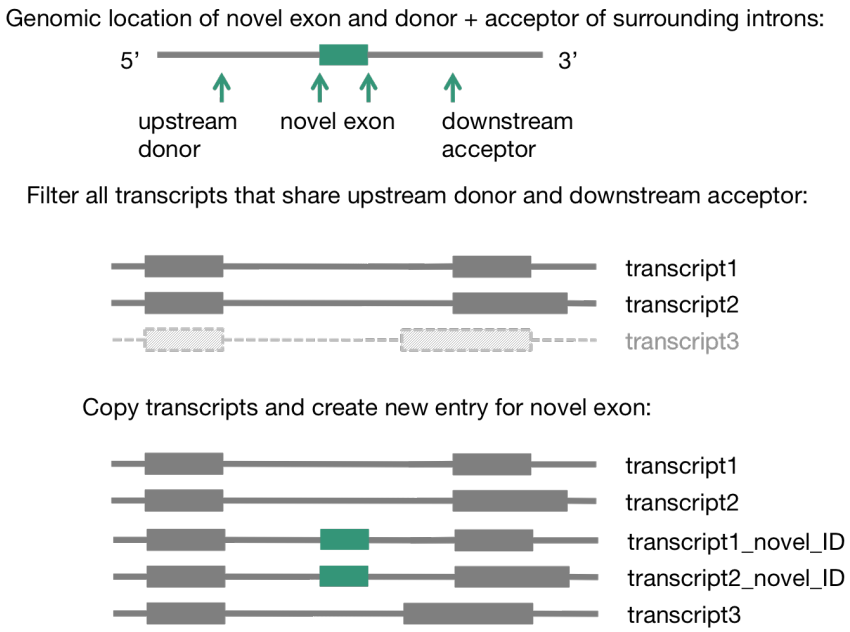


Figure 3: Extension of annotation catalog with predicted exons. The output from our prediction method are the coordinates of the novel exon and the connected exon. For each predicted exon, we filter the gene annotation and keep all transcripts that share the upstream donor and downstream acceptor. We copy the filtered transcript annotations, append the novel exon ID to the transcript name and create a new entry with the start and end coordinates of the predicted exon.

2.7 Extension of existing annotation

DISCERNS can extend the existing annotation by adding the predicted exons (Figure 3) to the corresponding transcripts. For each predicted exon, all transcripts that contain the up- and downstream exons of the predicted exon are copied and given a new transcript ID and name. A new

exon entry is created with the start and end coordinates of the novel exon and the `exon_ID` is suffixed such that the predicted exon can be identified. By copying the transcripts before inserting the predicted exon, we make sure that the original transcripts are not altered.

2.8 Comparison with StringTie

StringTie version 1.3.4d [22] was run separately using the three reduced GTF files and the three corresponding BAM files from STAR (STAR alignment with parameter `outSJfilterOverhangMin6`, see Table 2). We tested different StringTie parameters but we only report the result for the parameter combination with the best precision and recall (`minReadCoverage1_minIsoformAbundance0.05`). All tested parameters are listed in Table 3. The StringTie output is a GTF file with all assembled transcripts. We construct a table with all novel exons predicted by StringTie through a comparison of the StringTie transcript assembly with the corresponding reduced GTF annotation file. All exons from the StringTie assembly that are missing in the annotation are novel. Unfortunately, StringTie does not currently report a confidence score for each transcript and therefore we use the reported average per-base coverage of each exon for ranking.

Table 3: Tested StringTie parameters. The first column (Name) specifies the name of the parameter setting. The second column (Parameter) defines how the parameter was set in the StringTie call. The last column describes the function of the parameter.

Name	Parameter	Description
default		only default parameters
minJuncOverhang6	`-a 6`	Filter out junctions with an overhang on both sides with less than this amount of bases. Default 10
minJuncOverhang3	`-a 3`	see above
minReadCoverage1	`-c 1`	Filter out predicted transcripts with a read coverage below this value. Default: 2.5
minIsoformAbundance0.05	`-f 0.05`	Minimum isoform abundance of the predicted transcripts as a fraction of the most abundant transcript assembled at a given locus. Default: 0.1
minIsoformAbundance0.2	`-f 0.2`	see above
noEndTrimming	`-t`	Disable trimming at the ends of assembled transcripts based on sudden drops in coverage.
minReadCoverage1_ minIsoformAbundance0.05	`-c 1 -f 0.05`	see above for explanation of `-c` and `-f`

2.9 Precision-Recall curve of exon predictions

Precision-Recall curves were computed for both the DISCERNS exon predictions and those from StringTie. All predictions were compared to the list of known novel exons (see 2.2), the truth, based on the `start`, `end`, `lend` and `rstart` coordinates. If all four of the predicted coordinates matched a true novel exon, the prediction was labelled as `True` and `False` otherwise. Predicted terminal exons at the 5' (3') end of a transcript were labelled `True` if their `end` and `rstart` (`lend` and `start`) coordinates matched a true novel exon and `False` otherwise. The table was sorted by the minimal number of supporting junction reads (DISCERNS predictions) or the average per-base coverage of the exon (StringTie predictions) and we use the sorting column as score for the precision-recall plot. The FN value is global and defined by the number of true novel exons that are missing in the set of predicted exons. On the other hand, the TP and FP predictions were counted for all existing score thresholds. Precision and recall were computed for increasing score thresholds from the cumulative sums of TPs and FPs and the global FN value. All true novel exons without a read supporting the splice junction(s) were excluded from this analysis. Precision and recall were separately computed as described for lowly and highly expressed exons (simulated number of reads below or above median) and the three different exon classes (see Section 2.6). This results in 6 separate precision-recall curves: the two expression levels times the three different exon classes.

3 Results

3.1 Mapping comparison

The simulation is outlined in Figure 1. It is based on a real RNA-seq experiment (paired-end, 101 nts length) to make the distribution of gene expression more realistic. We simulated reads only for chromosome 19 and 22 to reduce the runtime of our pipeline. We used the exon boundaries from the original annotation and simulated 1,118,017 reads in 47,897 exons from 4,172 genes. For each of the simulated reads, we know the true genomic location. Additionally, we created three reduced annotation sets that consist of the human gene annotations for chromosomes 19 and 22 from which we removed (1) 100 microexons ($<$ 28nts), (2) 100 exons ($>$ 27nts) or (3) 100 microexons and 100 exons. To compare the performance of the two most commonly used splice-aware alignmnet tool, STAR and hisat2, we aligned the simulated reads using the three reduced annotation sets. We know the sets of exons and splice junctions that are missing from the annotation and we can thus

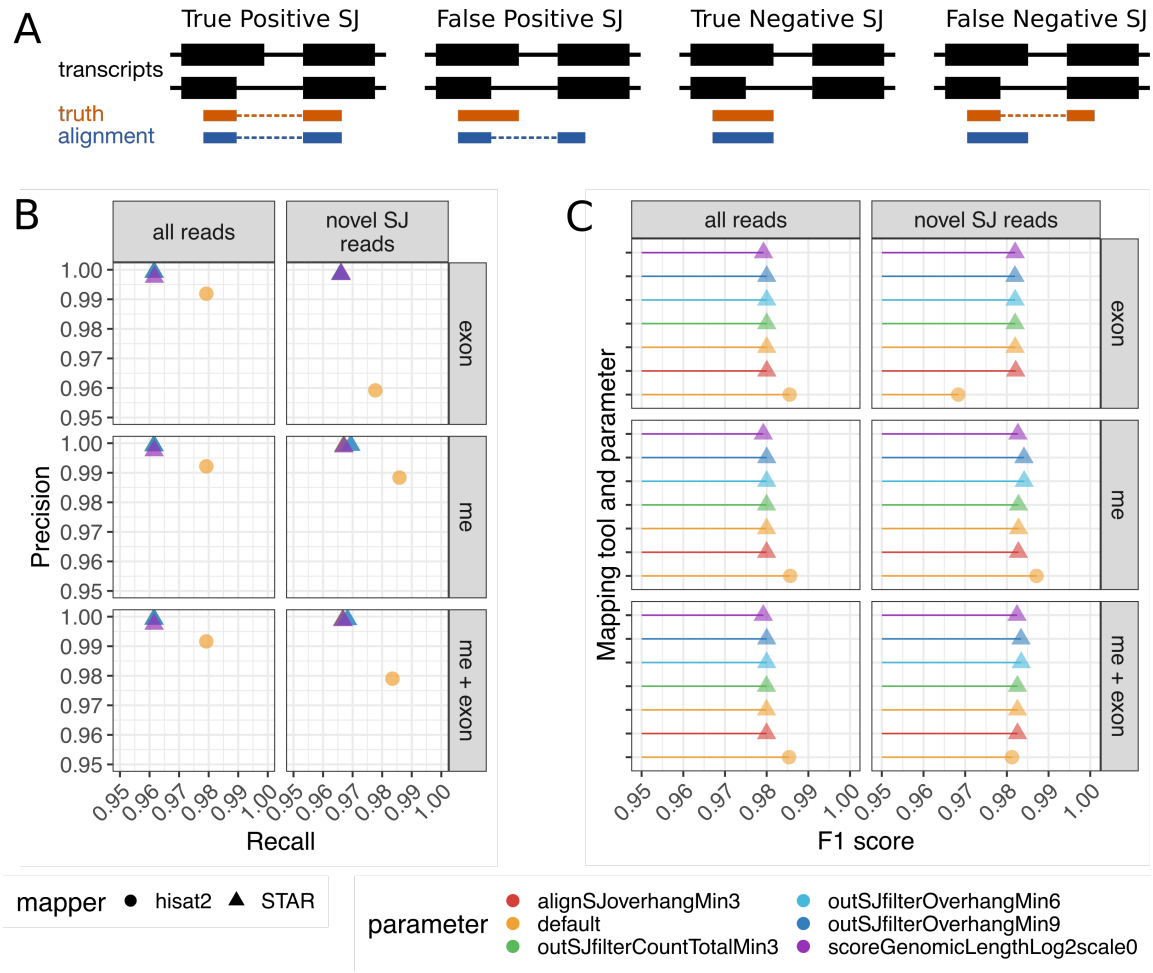


Figure 4: Comparison of STAR and hisat2 splice-junction read alignments (**A**) Schematic of true positive, false positive, true negative and false negative splice junctions. The gene annotation is shown in grey, the true genomic location of the simulated read in green (truth) and the aligned read in blue. (**B**) Precision-Recall plot for hisat2 and STAR alignments (symbol) for different parameter settings (color) as described in Table 2. The left plot shows precision and recall of all simulated reads and the right plot only considers the reads that overlap a novel exon. The first column in each plot considers the first read in the pair and the second column the second read in the pair. The three rows correspond to the sets of missing exons. First row 100 exons (exon); second row 100 microexons (me); third row 100 exons and 100 microexons (me_exon). (**C**) F1 score of the hisat2 and STAR alignments for different parameter settings (color). The plot is outline identical to the plots in (**B**).

evaluate how well each tool finds the missing events. We compared the read mapping of STAR and hisat2 in terms of splice junctions (Figure 4A). Precision and recall of the STAR and hisat2 alignments are very similar (Figure 4B, note the scale of the plot axes). Hisat2 has slightly better recall, but STAR has better precision. We are more concerned about false positive predictions than missing a novel event and thus precision is more important in our opinion. DISCERNS uses the SJ.out.tab file from STAR as input and we thus recommend STAR as alignmnet tool.

We run STAR with different parameter settings (Table 2). To our surprise, we could not detect strong differences between the tested STAR parameters (Figure 4B-C). Only lowering parameter `outSJfilterOverhangMin` increased the number of correctly detected novel splice junctions. The alignments of all reads with novel exons with STAR parameter `outSJfilterOverhangMin6` and `outSJfilterOverhangMin9` had the best F1 scores (Figure 4C right panel). We used STAR with `outSJfilterOverhangMin6` for all further analyses.

3.2 Novel exon prediction

We developed DISCERNS, an R-package for the prediction of novel exons in RNA-seq data using splice-junction reads. DISCERNS takes the gene annotation, a BAM file and the corresponding SJ.out.tab file from STAR as input and generated a table with the coordinates of predicted novel exons, as well as neighbouring exons. We developed specific strategies to detect novel exons with two (easy), one (medium) or zero (complicated) novel splice-junctions. Easy novel exons are predicted based on pairs of novel splice junctions that fall within annotated introns. Medium novel exons are predicted from a single novel splice junction. We take all reads with the novel splice junction and filter the ones with a a second splice junction. For the complicated unannotated exons, we do not have novel splice junctions that tell us where we can expect a novel splicing event. Therefore, we consider all reads that have two splice junctions and compare the splice junctions to the all transcripts that stem from this genomic region. If no transcript supports the pair of splice junctions, the splice junctions define a new splicing event. Finally, DISCERNS can extend an existing annotation catalog with the predicted exons (Figure 3). The predicted exons will be added to all transcripts with the predicted upstream donor and downstream acceptor; existing transcripts will also remain in their original form.

3.3 Performance evaluation

We compared DISCERNS to StringTie, the current best genome-guided transcriptome assembler. In particular, we wanted to evaluate how well the methods are able to recover novel exons in the simulated RNA-seq data set. We aligned the reads to the human genome with STAR and parameter `outSJfilterOverhangMin6` using the three reduced annotation sets. Precision and recall were computed for DISCERNS and StringTie and the reduced annotation with both novel exons and microexons (Figure 5). There were 75 easy, 82 medium and 52 complicated novel exons with at least one read supporting each splice junction. In total, StringTie predicted 1328 novel

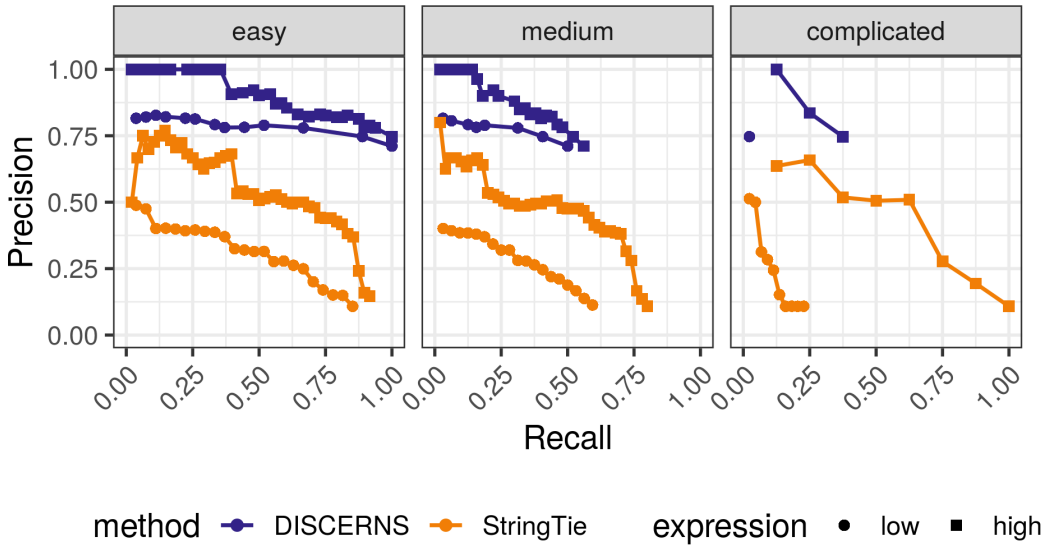


Figure 5: Precision-Recall curves for the exon prediction comparison. DISCERNS (blue) and StringTie (orange) exon predictions based on STAR alignment with the reduced GTF (100 exons and 100 microexons missing) and parameter `outSJfilterOverhangMin6` (see Table 2). Novel exons were divided into two sets, based on the number of simulated reads per exon: circles denote lowly expressed exons ($\leq$ median number of reads) and squared denote highly expressed exons ($>$ median number of reads). The three panels show the curves for each of the three exon classes: easy, medium and complicated.

exons of which 144 were correct and 1184 wrong. DISCERNS predicted 173 novel exons of which 123 were correct and 50 wrong. Stringtie has much lower precision, because it predicted many more novel exons than DISCERNS and had much higher numbers of false positives. DISCERNS has high overall precision, indicating that the predictions with a high number of supporting reads are correct. All easy novel exons were correctly detected by DISCERNS. More than half of the medium novel exons were correctly identified independent of exon expression. 38% of the highly expressed complicated novel exons were identified, but only one of the lowly expressed ones. Overall, DISCERNS' recall is lower than StringTie's recall. However, StringTie had surprisingly low precision (< 0.5) for the lowly expressed novel exons. Compared to DISCERNS, StringTie recall was better for the medium and complicated novel exons, but worse for the easy novel exons. As expected, both DISCERNS and StringTie identified more novel exons with high expression than lowly expressed novel exons. In summary, DISCERNS has good precision and reliably predicts novel exons with enough read coverage, suggesting that it can be used to search for unannotated exons in RNA-seq data. StringTie has good recall, but it also created many wrong predictions. Unfortunatley, the StringTie output does not include a confidence value (other than read depth), thus making it impossible for the user to differentiate between correct and wrong predictions.

3.4 DISCERNS R package

DISCERNS is available as an R package and can be found at `https://github.com/khembach/DISCERNS`. We have tested the package with BAM files of $>$ 100 million reads which resulted in a maximal memory consumption of 25G.

4 Application

To highlight the usage of DISCERNS, we predicted novel splicing events in three ALS-related RNA-seq data sets where we expected to find cryptic exons. First, we show that DISCERNS correctly recovers reported human cryptic exons associated with nuclear loss of TDP-43 function. Then, we show that DISCERNS predicts unannotated exons in the human genome and identifies novel ALS patient specific splicing events, including microexons. Lastly, we report that DISCERNS correctly identifies a known cryptic exon in stathmin-2 and that it predicted multiple novel cryptic exons caused by TDP-43 knockdown or mutant TDP-43.

4.1 DISCERNS recovers known cryptic exons from Ling et al. (2015)

As a first validation step, we applied DISCERNS to a RNA-seq data set from Ling *et al.*[6] that consists of two samples: WT and TDP-43 siRNA-treated HeLa cells. Ling *et al.* reported 41 cryptic exons (see supplemental table 3 of Ling *et al.*[6]). The cryptic where identified by manual screening of novel exons annotated by Cufflinks [24]. The authors searched for novel exons that were highly abundant in the TDP-43 knockout samples but not the control samples.

We mapped the reads to the human genome GRch38 and predicted novel exons with DISCERNS. Ling et al. reported the location of the cryptic exons on the old human genome (hg19). Therefore, we first converted their putative cryptic exon locations to GRch38 (liftOver) and then compared them to our predicted exons (Table 4). We correctly predicted 14 cryptic exons. For 13 cryptic exons, we predicted a novel exon that had the correct start (7) or end (8) coordinates. For two cryptic exons, the DISCERNS predictions were labeled as wrong, because the predicted exon start and end did not match the published coordinates. However, a closer inspection of the genomic region in IGV revealed that our predictions match the read alignments and that the published coordinates of the cryptic exons are either wrong or there was a problem with the liftOver. 12 cryptic exons did not have any predictions because 10 of them were already annotated in GRch38 and the other 2 did not have any read coverage (possibly due to genome conversion

Table 4: The 41 human cryptic exons from Ling *et al.* classified by their corresponding DISCERNS prediction.

DISCERNS prediction	# cryptic exons	explanation
correct	14	identical star and end
partially correct	13	7 correct start; 8 correct end
wrong	2	wrong start and end
missing	12	2 no prediction; 10 are already annotated

errors).

DISCERNS identified many more putative exons in the two samples, but the majority were found in both samples and presumably specific to HeLa cells, which originated from a cervical cancer and we thus expected to find transcriptomic differences to the human reference annotation.

4.2 DISCERNS finds unannotated splicing events and microexons in the human frontal cortex

Next, we reanalysed a data set from Prudencio *et al.*[32], which includes paired-end RNA-seq data set from sporadic ALS (sALS) patients and healthy controls. Prudencio *et al.* sequenced the cerebellum and the frontal cortex of each individual, but we only analysed the frontal cortex samples because the frontal cortex is the brain region where TDP-43 pathology occurs in sALS and where we expected to find splicing errors due to TDP-43 loss of function in the nucleus. We mapped all frontal cortex samples of sALS patients and controls to the latest human genome build (GRCh38) and predicted novel exons with DISCERNS.

The number of aligned reads and the number of predicted exons was comparable between the control and sALS samples. Most samples had 1000-2000 DISCERNS predictions (Figure 6A). Contrary to our expectation, the sALS samples do not have more predicted exons than the control samples. The percentage of identical predictions between pairs of samples is only slightly higher between sALS samples than any pair of control samples (Figure 6B). The TDP-43 pathology in the sALS samples does not seem to cause more novel splicing events in the patients than in the controls.

Altogether, we could identify three distinct types of predictions. The first type are predicted exons that are shared between all samples, which can be explained by missing annotations in the human genome (for example, KIFAP3 in Figure 6C). The second type are predicted exons that are specific for (some of) the sALS samples and they might be caused by TDP-43 pathology. However,

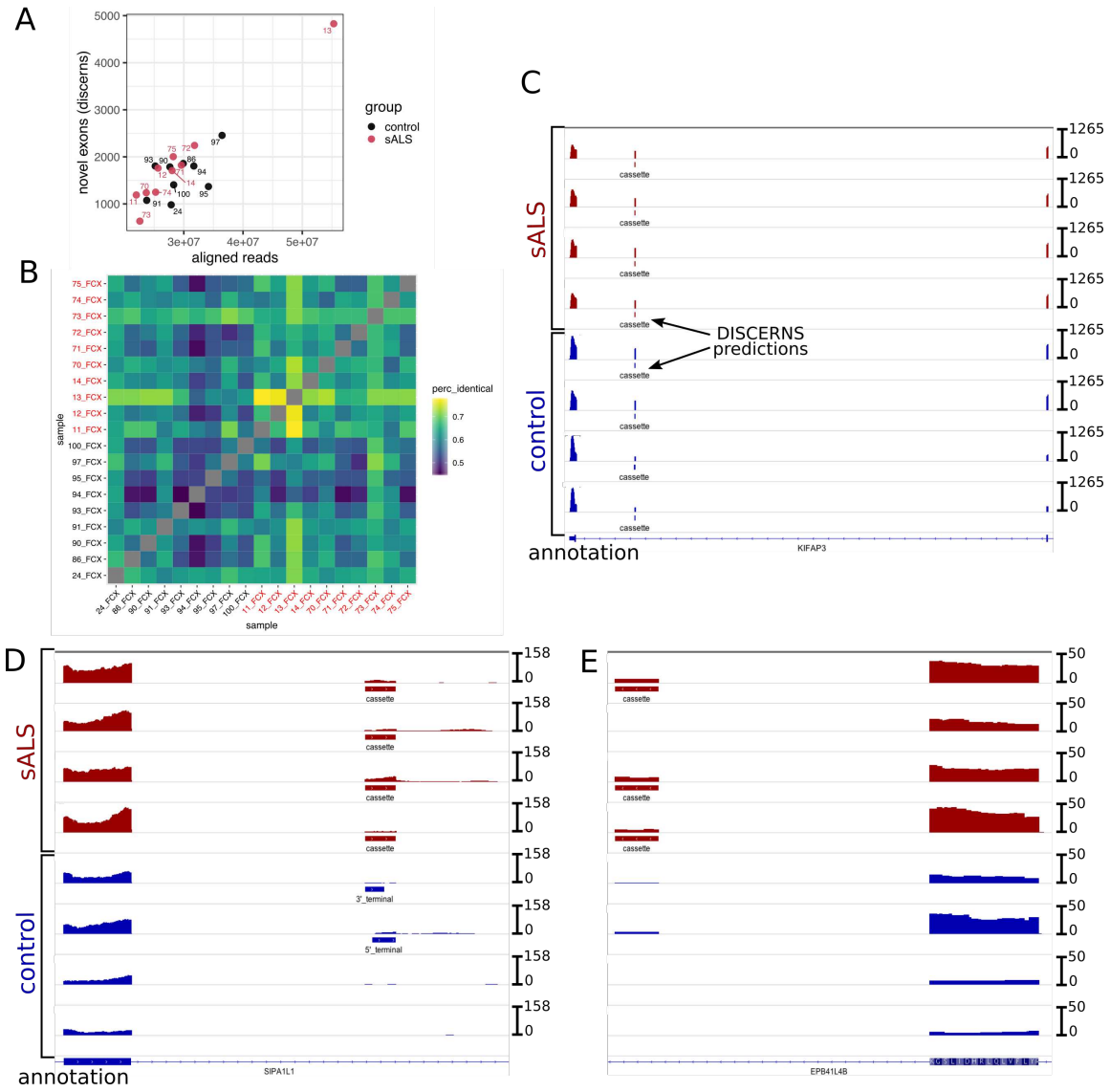


Figure 6: DISCERNS analysis of the frontal cortex samples from Prudencio *et al.*[32]. (A) Number of aligned reads and number of predicted novel exons per sample (control samples in black and sALS samples in red). (B) Heatmap of all analysed samples. The color indicates the percentage of identical predictions between a pair of samples. Control samples are labelled in black and sALS samples in red. (C) IGV screenshot of the gene KIFAP3 and eight representative samples with similar number of aligned reads. The sALS samples are shown in red and the control samples in blue. DISCERNS predictions are indicated as bars below the coverage tracks. The predicted cassette exon is found in all 19 samples and indicates a missing exon annotation in the human genome. (D) IGV screenshot of gene SIPA1L1. The gene contains a novel cassette exon that is specific for the sALS samples. (E) IGV screenshot of gene EPB41L4B where DISCERNS predicted a novel microexon (18 bp).

the number of supporting reads in the sALS samples is often quite low (less than ten reads) and the number of reads in the control samples might simply be below the detection threshold (Figure 6D). The third type of predicted exons are microexons (Figure 6E). Some of the microexons appear to

be specific for the sALS samples, but again, the number of supporting reads is low and microexons might be missed in the control samples due simply to low coverage.

Some of the DISCERNS predictions were obviously caused by wrongly mapped reads (data not shown); of course, other approaches will be affected by such errors. STAR sometimes wrongly creates novel splice junctions if two consecutive exons in a transcript share identical sequences. Wrong novel splice junctions in turn can result in wrong exon predictions from DISCERNS.

In summary, the frontal cortex data set contains putative novel exons that we could identify with DISCERNS. Many of the predictions can be explained by missing exon annotations in the human genome or by wrongly aligned reads. However, we did find a few predictions that could be direct or indirect effects of TDP-43 pathology and the loss of TDP-43 splicing control in the nucleus. The sALS specific splicing changes were located in the genes SMOC1, SIPA1L1, GPR137B, TMEM128, ARFGEF1, EPB41L4B and EVI5. Further experiments and analyses are required to determine if any of the detected novel exons can be reproduced.

4.3 DISCERNS identifies cryptic exons caused by TDP-43 knockdown

The last data set that we analysed with our pipeline is from Melamed *et al.*[33], where the authors report that the STMN2 gene expressed a cryptic exon under TDP-43 knockdown conditions. STMN2 encodes the stathmin-2 protein, a neuronal growth-associated factor. Melamed *et al.* used SH-SY5Y cells and either downregulated TDP-43 with siRNA or they introduced an ALS causing mutation (N352S) in both TDP-43 alleles using CRISPR-Cas. The N352S mutation is known to cause familial ALS.

We reanalysed the samples and predicted novel splicing events with DISCERNS (we required at least 5 supporting reads). In this data set, the samples were sequenced with single-end 51 bp sequencing, which increases the challenge to computationally find novel splicing events.

In total, we predicted 3153 novel exons (Figure 7A). We mostly identified 3' terminal (1439) and 5' terminal exons (1421), and only a few cassette exons (293). The short read length is the reason why DISCERNS could not identify longer cassette exons. The number of predicted exons strongly depends on the required number of supporting reads. We only predict 120 novel exons with at least ten supporting reads per sample (41 3' terminal, 43 5' terminal and 36 cassette exons). Amongst these, seven predicted exons were specific for the three siTDP-43 samples (three cassette, 1 5' terminal and three 3' terminal exons). All seven predicted exons are cryptic exons that showed no or weaker coverage in the control samples (genes ZNF826P, PDXN, MEIS2, ELAVL3, ISL1,

ARHGAP32, AC01522.1). The cryptic exon in ELAVL3 (Figure 7D) has already been reported in a study from Klim *et al.*[34] where the authors sequenced the transcriptome of siTDP-43 treated human motor neurons (hMNs). ELAVL3 has enriched neuronal expression, but is downregulated upon TDP-43 knockdown in hMNs. Some of the predicted novel exons were found in all samples and are most likely common to SH-SY5Y cells, but simply not annotated in the reference (Figure 7B). The reported cryptic exon in STMN2 was identified by DISCERNS, including in two of the control samples (Figure 7C).

We found predicted exons that were shorter versions of already annotated exons. Most of these cases were microexons with 6 or 9 bp. There are two possible explanations for this: First, the short predicted microexons are false positives caused by incorrectly mapped reads because the intronic sequences at the start and end of two consecutive exons are very similar. However, we also found some predictions where we could not identify any wrongly mapped reads. The second explanation is that the predictions are correct. The default STAR parameters require a splice junction overlap of at least 9 bp but we lowered this parameter to 6 bp which could explain why the shorter version of the microexons have not been reported before.

Apart from the cryptic exon in STMN2, we found multiple other novel splicing events that were only detected in the TDP-43 siRNA samples, but not in the control siRNA samples (Figure 7E-F). Surprisingly, we did not find any novel exons that were specific for the N352S cells. All putative N352S specific unannotated exons also had read coverage in other samples, but the number of splice junction reads was below the identification threshold (5 reads).

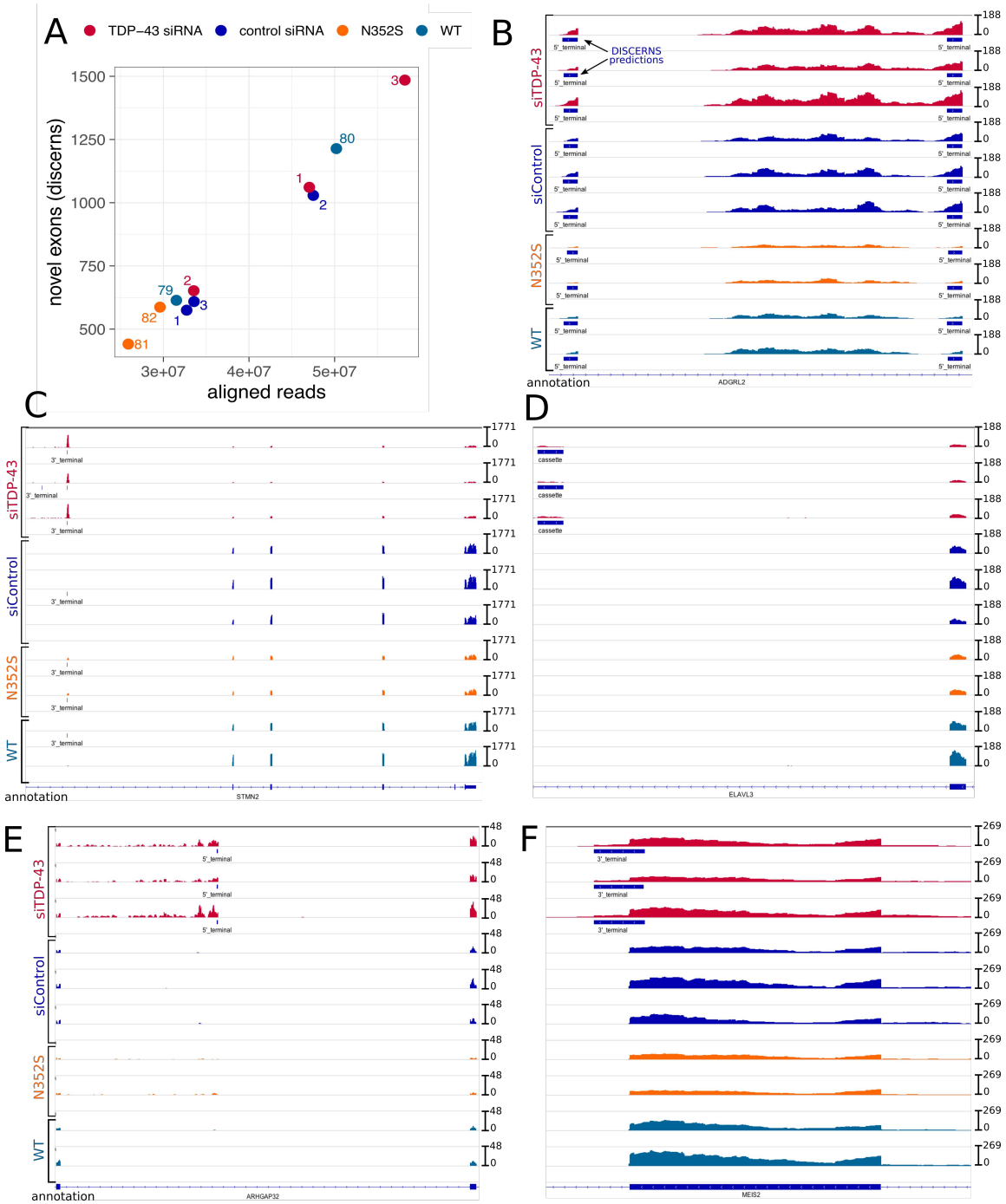


Figure 7: DISCERNS analysis of the Melamed *et al.*[33] data set. (A) Number of aligned reads and number of predicted novel exons per sample (TDP-43 siRNA in red, control siRNA in blue, wild type SH-SY5Y in green and N352S TDP-43 in orange). (B) IGV screenshot of a region in the gene ADGRL2 with two novel exons in all samples. TDP-43 siRNA samples in red, control siRNA in blue, N352S TDP-43 in orange and wild type SH-SY5Y in cyan. The DISCERNS predictions are indicated as blue bars below each coverage track. (C) IGV screenshot of the STMN2 gene. The novel exon was found in all TDP-43 siRNA and N352S TDP-43 samples, as well as in two control samples. (D) IGV screenshot of part of gene ELAVL3. The gene contains a novel cassette exon that is specific for the TDP-43 siRNA samples. (E) IGV screenshot of part of gene ARHGAP32. The TDP-43 siRNA samples contain a cryptic exon. (F) IGV screenshot of the second exon in gene MEIS2. The TDP-43 siRNA samples reveal an exon extension event.

5 Discussion

Given the low cost and accessibility of deep Illumina sequencing of cDNA libraries, the analysis of alternative splicing has become much more accessible. Publicly available databases contain thousands of RNA-seq data sets with raw or already mapped reads. Examples of such databases are the European Nucleotide Archive (ENA) [35] or the Gene Expression Omnibus (GEO) [36]. The large scale analysis of these public data sets requires well designed pipelines that run fast, reliable and produce easily comparable and parsable outputs.

In this paper, we developed a pipeline for the prediction of novel splicing events and integrate them with existing pipelines (e.g., for quantification). We developed DISCERNS, an R package for the identification of alternative splicing in RNA-seq data sets using information from splice-junction reads. We showed that DISCERNS has better precision than StringTie on simulated RNA-seq data with missing (but known) splicing events. StringTie overestimated the number of novel events and predicted many false positives. DISCERNS had better precision, i.e. low number of false positives, but slightly lower recall, i.e. missing predictions.

DISCERNS creates an output table with the coordinates of all predicted events and the number of supporting reads of the novel splice junctions, which can be consolidated across multiples samples. The table is easy to read, understand and process. Moreover, DISCERNS provides a function to augment existing genome annotations in GTF format with the predictions. The extended GTF file from DISCERNS can directly be used for downstream analyses to ensure correct feature quantification, such as exons, transcripts or genes. It has been shown that missing or incomplete annotations can lead to wrong transcript expression estimates [37]. Therefore, it might be beneficial to discover unannotated splicing events and to extend the annotation catalog prior to quantification even for RNA-seq analyses, such as differential gene or transcript expression, where exon prediction is not the objective of the study.

Currently, DISCERNS processes each sample individually. For the application to the three ALS data sets, we post-processed the DISCERNS results to identify predictions that appeared in more than one sample. However, we could implement this functionality in DISCERNS by providing a function that takes multiple output tables input and filters the predictions based on user defined parameters. These parameters could be for example a minimal number of samples, in which the prediction has to appear or a minimal number of supporting reads required in a specific number of samples.

Long read sequencing, also termed third-generation sequencing, generates reads of more than 10 Kbp. At present, the technology still has lower sequencing depth, higher error rates and higher cost per base than RNA-seq, but continues to improve [38]. On the long term, long-read sequencing will most likely replace RNA-seq for the purpose of novel splicing event detection and transcriptome assembly. Particularly, because long-read sequencing can be targeted to genomic regions by amplification of the regions of interest with long-range PCR. Most existing alternative splicing detection methods will also work for long-read data sets, especially DISCERNS, because it takes aligned reads as input and the detection of novel exons will only be improved by longer reads.

We reanalysed ALS related RNA-seq data set with DISCERNS and showed that it correctly predicts known cryptic exons. DISCERNS also found novel splicing events in SH-SY5Y cells. However, the sALS samples from Prudencio et al. did not show higher number of identical predictions withing the patient group than compared to the control samples. An explanation could be that the post-mortem RNA-seq data set might not have the required quality to reliably identify novel exons and splicing events. We did find a few novel events that might be caused by the loss of TDP-43 function in the nucleus, but the exon might simply be missed in the control samples because of too few reads. One reason for the high variance and lower quality is that the post-mortem interval varied from two to 30 hours between the samples. What also complicates the analysis is that we cannot have paired disease and control samples from an ALS patient. The patient and controls are unrelated and greatly vary in their age of death.

We also identified the published STMN2 cryptic exon [33] in one of the sALS patient samples (sample 72) and it had a few reads in a second patient sample (73), but was below the detection limit. It was not detectable in any other sample. This indicates that TDP-43 pathology did not cause a complete loss of TDP-43 splicing function in these patients and cryptic exons were still correctly repressed. In contrast, the stathmin-2 downregualation was very strong in the SH-SY5Y cells in the Cleveland study.

In all data sets, DISCERNS identified many novel exons that were missing in the human Ensembl annotation, but present in the RefSeq annotation of the human genome. The two annotations are mostly comparable, but there are a few genes that are missing transcripts in one of the two annotation. If we would have used the RefSeq annotation instead of Ensembl, we would most likely have identified missing exons as well. However, this highlights the importance of using the correct annotation set for your analyses and that differences in gene annotation can lead to

wrongly estimated transcript/gene expression levels.

Knowing the splicing changes that occur in disease opens new possibilities for therapies or treatment. One example are antisense oligonucleotides that are targeted to mis-spliced transcripts to prevent the protein translation from these mRNAs. ASO therapy for spinal muscular atrophy (SMA) was approved in 2016 [39] and ASOs for ALS therapy are currently being developed [40,41]. Melamed *et al.*[33] showed that restoration of stathmin-2 RNA could rescue axonal regeneration upon loss of TDP-43 function in iPSCs. This highlights the therapeutic possibilities of characterising disease specific changes in alternative splicing and transcription regulation. With the many high quality tools available, we are now able to revisit legacy data sets that have been deposited in public repositories and to analyse them with the goal of characterising and identifying sample or disease/condition specific events that will ultimately help to understand disease progression and to identify potential targets for treatment.

Funding

MDR acknowledges support from the University Research Priority Program Evolution in Action at the University of Zurich.

References

1. Cartegni, L., Chew, S. L. & Krainer, A. R. *Listening to silence and understanding nonsense: Exonic mutations that affect splicing* 2002. doi:10.1038/nrg775.
2. Scotti, M. M. & Swanson, M. S. RNA mis-splicing in disease. *Nature Reviews Genetics* **17,** 19–32. ISSN: 14710064 (2016).
3. Beachy, P. A., Helfand, S. L. & Hogness, D. S. Segmental distribution of bithorax complex proteins during Drosophila development. *Nature* **313,** 545–551. ISSN: 00280836 (Feb. 1985).
4. Irimia, M. *et al.* A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell* **159,** 1511–1523. ISSN: 10974172 (2014).
5. Hoskins, A. A. *et al.* Ordered and dynamic assembly of single spliceosomes. *Science* **331,** 1289–1295. ISSN: 00368075 (Mar. 2011).
6. Ling, J. P., Pletnikova, O., Troncoso, J. C. & Wong, P. C. TDP-43 repression of nonconserved cryptic exons is compromised in ALS-FTD. *Science (New York, N.Y.)* **349,** 650–655. ISSN: 1095-9203 (Aug. 2015).
7. Singh, B., Trincado, J. L., Tatlow, P. J., Piccolo, S. R. & Eyras, E. Genome sequencing and RNA-motif analysis reveal novel damaging noncoding mutations in human tumors. *Molecular Cancer Research* **16,** 1112–1124. ISSN: 15573125 (2018).
8. Climente-González, H., Porta-Pardo, E., Godzik, A. & Eyras, E. The Functional Impact of Alternative Splicing in Cancer. *Cell Reports* **20,** 2215–2226. ISSN: 22111247 (2017).
9. Soneson, C. *et al.* A junction coverage compatibility score to quantify the reliability of transcript abundance estimates and annotation catalogs. *Life Science Alliance* **2,** e201800175 (2019).
10. Rhoads, A. & Au, K. F. *PacBio Sequencing and Its Applications* Oct. 2015. doi:10.1016/j.gpb.2015.08.002.
11. Sessegolo, C. *et al.* Transcriptome profiling of mouse samples using nanopore sequencing of cDNA and RNA molecules. *Scientific Reports* **9,** 1–12. ISSN: 20452322 (Dec. 2019).
12. Gunady, M. K., Mount, S. M. & Bravo, C. Fast and interpretable alternative splicing and differential gene-level expression analysis using transcriptome segmentation with Yanagi. *bioRxiv Bioinformatics,* 1–23 (2018).
13. Katz, Y., Wang, E. T., Airoldi, E. M. & Burge, C. B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods* **7,** 1009–1015. ISSN: 15487091 (Dec. 2010).
14. Shen, S. *et al.* rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proceedings of the National Academy of Sciences of the United States of America* **111,** E5593–E5601. ISSN: 10916490 (Dec. 2014).
15. Trincado, J. L. *et al.* SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biology* **19,** 40. ISSN: 1474-760X (Dec. 2018).
16. Wang, Q. & Rio, D. C. JUM is a computational method for comprehensive annotation-free analysis of alternative pre-mRNA splicing patterns. *Proceedings of the National Academy of Sciences of the United States of America* **115,** E8181–E8190. ISSN: 10916490 (Aug. 2018).
17. Denti, L. *et al.* ASGAL: aligning RNA-Seq data to a splicing graph to detect novel alternative splicing events. *BMC Bioinformatics* **19,** 444. ISSN: 1471-2105 (Dec. 2018).
18. Rogers, M. F., Thomas, J., Reddy, A. S. & Ben-Hur, A. SpliceGrapher: detecting patterns of alternative splicing from RNA-Seq data in the context of gene models and EST data. *Genome Biology* **13,** R4. ISSN: 1465-6906 (Jan. 2012).
19. Goldstein, L. D. *et al.* Prediction and quantification of splice events from RNA-seq data. *PLoS ONE* **11** (ed Xing, Y.) 1–18. ISSN: 19326203 (May 2016).
20. Kahles, A., Ong, C. S. & Rätsch, G. SplAdder: identification, quantification and testing of alternative splicing events from RNA-Seq data. *bioRxiv,* 017095. ISSN: 1367-4803 (2015).
21. Sterne-Weiler, T., Weatheritt, R. J., Best, A. J., Ha, K. C. & Blencowe, B. J. Efficient and Accurate Quantitative Profiling of Alternative Splicing Patterns of Any Complexity on a Laptop. *Molecular Cell* **72,** 187–200. ISSN: 10974164 (Oct. 2018).
22. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology* **33,** 290–295. ISSN: 15461696 (Feb. 2015).

23. Guttman, M. *et al.* Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature Biotechnology* **28,** 503–510. ISSN: 10870156 (May 2010).

24. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology* **28,** 511–515. ISSN: 1087-0156 (2010).

25. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology* **37,** 907–915. ISSN: 15461696 (Aug. 2019).

26. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21. ISSN: 13674803 (2013).

27. Cunningham, F. *et al.* Ensembl 2019. *Nucleic Acids Research* **47,** D745–D751. ISSN: 0305-1048 (Jan. 2019).

28. Li, B. & Dewey, C. N. RSEM : accurate transcript quantification from RNA-Seq data with or without a reference genome (2011).

29. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74. ISSN: 14764687 (Sept. 2012).

30. Grimwood, J. *et al.* The DNA sequence and biology of human chromosome 19. *Nature* **428,** 529–535. ISSN: 00280836 (Apr. 2004).

31. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079. ISSN: 1367-4803 (Aug. 2009).

32. Prudencio, M. *et al.* Distinct brain transcriptome profiles in C9orf72-associated and sporadic ALS. *Nature Neuroscience* **18,** 1175–1182. ISSN: 15461726 (Aug. 2015).

33. Melamed, Z. *et al.* Premature polyadenylation-mediated loss of stathmin-2 is a hallmark of TDP-43-dependent neurodegeneration. *Nature Neuroscience*. ISSN: 1097-6256. doi:10.1038/s41593-018-0293-z. http://www.nature.com/articles/s41593-018-0293-z (2019).

34. Klim, J. R. *et al.* ALS-implicated protein TDP-43 sustains levels of STMN2, a mediator of motor neuron growth and repair. *Nature Neuroscience* **22,** 167–179. ISSN: 15461726 (Feb. 2019).

35. Leinonen, R. *et al.* The European Nucleotide Archive. *Nucleic acids research* **39,** 28–31. ISSN: 1362-4962 (Jan. 2011).

36. Edgar, R., Domrachev, M. & Lash, A. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research* **30,** 207–210. ISSN: 1362-4962 (Jan. 2002).

37. Chabbert, C. D., Eberhart, T., Guccini, I., Krek, W. & Kovacs, W. J. Correction of gene model annotations improves isoform abundance estimates: The example of ketohexokinase (Khk) [version 2; peer review: 3 approved]. *F1000Research* **7,** 1956. ISSN: 1759796X (Apr. 2019).

38. Pollard, M. O., Gurdasani, D., Mentzer, A. J., Porter, T. & Sandhu, M. S. Long reads: their purpose and place. *Human molecular genetics* **27,** R234–R241. ISSN: 1460-2083 (2018).

39. Aartsma-Rus, A. *FDA Approval of Nusinersen for Spinal Muscular Atrophy Makes 2016 the Year of Splice Modulating Oligonucleotides* Apr. 2017. doi:10.1089/nat.2017.0665.

40. Miller, T. M. *et al.* An antisense oligonucleotide against SOD1 delivered intrathecally for patients with SOD1 familial amyotrophic lateral sclerosis: A phase 1, randomised, first-in-man study. *The Lancet Neurology* **12,** 435–442. ISSN: 14744422 (May 2013).

41. Klim, J. R., Vance, C. & Scotter, E. L. *Antisense oligonucleotide therapies for Amyotrophic Lateral Sclerosis: Existing and emerging targets* May 2019. doi:10.1016/j.biocel.2019.03.009.

Part III.

Concluding Remarks

6 Conclusion and Outlook

In this PhD thesis, I presented my work on the establishment of pipelines for high-throughput sequencing data to understand RNA metabolism defects associated with ALS. The majority of transcriptomic studies nowadays uses RNA-seq to assess gene expression across experimental groups. In the last ten years, many methods and analysis pipelines for RNA-seq data have been published. However, there is still room for improvement, especially in reproducible research and the development of easy to use, standardized pipelines. This includes pipelines for specialized discoveries and differential methods, such as for microexons or alternative splicing.

My first collaborative paper presented a comprehensive review of the current state-of-the-art RNA-seq technologies and differential analysis approaches. We explained the RNA-seq protocol and give recommendations for the design of RNA-seq experiments.

My second collaborative paper introduced ARMOR, a snakemake workflow for automated processing and differential analyses of RNA-seq data. We use snakemake [1] as the workflow language for two reasons. Firstly, interoperability with conda [2] makes it platform independent and secondly, snakemake workflows are modular and can easily be adjusted by the user to replace existing modules or to add modules with additional functionality. ARMOR includes modules for read preprocessing and quality control. The user can quantify transcript and gene expression with Salmon and, optionally, align the reads to a reference genome with STAR to visualise them in a genome browser. Differential gene expression analysis is performed with edgeR [3] and differential transcript usage with DRIMseq [4]. As output, ARMOR generates a standardized SingleCellExperiment R object that can readily be imported into R or iSEE [5] for manual inspection and further analyses. In summary, ARMOR facilitates RNA-seq data processing and analysis. I hope ARMOR enables researchers without a strong background in computational biology to engage in data analysis and exploration. Besides, I hope that ARMOR promotes reproducible research and boosts the comparison and reanalysis of existing RNA-seq data sets.

The third paper in my thesis was a close collaboration with Sonu Sahadevan, who performed most of the experimental work in the paper. In this project, we set out to understand the role of FUS at the synaptic site in neurons from mouse frontal cortex. The second goal was to study the effect of increased cytoplasmic levels of FUS in mice with a heterozygous mutation in the FUS NLS that prevents the protein from being imported into the nucleus (FUS-KI mice). We sequenced two CLIP samples of wildtype mouse frontal cortex and synaptoneurosomes (SNS) to identify synapse specific RNA targets of FUS. We developed a filtering strategy to identify the RNAs that were specifically bound by FUS in the SNS sample. We showed that the synaptic FUS targets are mainly located in exons and the 3'UTR region of transcripts suggesting a potential role of FUS in RNA transport, stability and possibly local translation; in the nucleus, FUS binds mainly intronic regions. The genes with synaptic FUS peaks were associated with synapse organisation and plasticity, indicating the importance of FUS in maintaining synapse integrity in the mouse frontal cortex. We identified age-dependent synaptic changes in transcripts that code for essential proteins of the GABAergic and glutamatergic networks, by RNA-seq of FUS-KI mice. Some of these genes were also putatively bound by FUS in the synapse, highlighting the importance of stable FUS levels at the synaptic site. We observed transcriptomic changes as early as 6 months, even before the mice showed mild motor neuron deficits.

Our study highlights early synaptic impairments as a possible cause of ALS caused by mutant FUS. It also highlights a possible link between FTD and ALS because the FUS-KI mice first show cognitive impairment before they develop motor neuron deficits. This suggests that doctors should monitor patients with a family history of ALS to capture memory impairment as early as possible, because this could be an early symptom of ALS even before motor deficits manifest. Optimally, patients can be diagnosed with ALS early on to improve prognosis and therapy. Currently, there is no cure for ALS. However, there are two disease-modifying treatments and more therapies are currently in clinical trials [6].

In the FUS-KI mice RNA-seq experiment, we did not observe many genes with high log2 fold-change: The majority of genes had a log2 fold-change below 2. One possible explanation for this is that FUS is part of the FET family of proteins (FUS, EWS and TAF15). All FET proteins are very similar in domain composition and they are known to have shared functionality and to bind similar RNA targets [7]. It is likely that EWS and TAF15 compensate for the loss of FUS function in the nucleus, caused by decreased

levels of FUS. Interestingly, TAF15 binds a GGUAAGU motif [7] and we identified the same motif in the exonic and 5' UTR peak sequences of our synaptoneurosome FUS CLIP-seq (see Table 1 in the FUS paper). Another explanation for the small observed log2 fold-changes is that the mice still have FUS proteins in the nucleus. The mice are heterozygous for the NLS mutation and the wildtype copy of the gene still produces functional FUS proteins.

The fourth paper in this thesis presents DISCERNS, a novel method for the discovery of unannotated exons in RNA-seq data. This paper was motivated by the lack of methods to extend existing annotation catalogs with novel events. We developed an R package that takes RNA-seq genome alignments and splice junctions from STAR as input and compares novel splice junctions to the reference gene annotation to identify novel exon. The package includes a function to extend gene annotations in GTF format with the predicted exons. We simulated an RNA-seq data set with known novel splicing events that enabled us to evaluate different alignment tools and to test the prediction performance of DISCERNS, which showed high precision.

We applied DISCERNS to three published ALS related RNA-seq data sets. The original publications reported multiple cryptic exons in these data sets and we correctly recovered most of them. In addition, we identified other cryptic events: some of them had been reported before, some were novel. We are planning to experimentally verify some of the predicted cryptic exons from DISCERNS. For this, we could knock down TDP-43 expression with siRNA in human neurons derived from induced pluripotent stem cells. Ideally, we should be able to reproduce cryptic events associated with TDP-43 depletion in these neurons using reverse transcription PCR and primers specifically designed to amplify the cryptic exon splice junctions.

I combined DISCERNS with ARMOR to analyse the three published ALS RNA-seq data sets. Raw reads were preprocessed and aligned to the reference genome with STAR using the ARMOR workflow. Subsequently, the STAR output files were used as input to DISCERNS. We are thus considering to make DISCERNS an optional step in ARMOR. This way, the user would have the possibility to predict and filter novel exons in all or a subset of the samples and to add the predicted events to the reference gene annotation. Gene expression quantification with Salmon can then be performed on the augmented set of genes to improve the gene expression estimates.

An important consideration is that many ALS studies are exclusively conducted in mice. Naturally, we cannot perform most experiments with human individuals and hu-

man ALS research is restricted to cell culture experiments or samples from post mortem brains of ALS patients. On the contrary, we can create genetically modified mouse models or deplete/knockout the expression of a gene of interest in living mice, rendering mouse research very attractive. Valuable discoveries have been made with model organisms, but one needs to be aware that there are limitations, especially in disease research. For example, in the Ling *et al.* [8] study, the cryptic exons identified in TDP-43 knockout mice had no overlap with the human cryptic exons in TDP-43 siRNA treated HeLa cells. Instead, the human cryptic exons were located in different genes than the ones in mice. Notably, TDP-43 is conserved between human and mouse with 96% identity on the amino acid sequence level.

Public repositories, such as the gene expression omnibus (GEO) [9], contain many ALS related RNA-seq data sets that are readily available for download. I hope that my PhD, and especially DISCERNS and ARMOR, contributed to the establishment of standardized workflows for RNA-seq data analysis in the context of ALS research. This will make analysis of public data sets faster and, ultimately, lead to a more complete list of cryptic splicing events associated with ALS. We are just beginning to understand the molecular mechanisms that lead to neurodegeneration in ALS patients and I hope that this PhD thesis made a valuable contribution.

References

[1] J. Köster and S. Rahmann. "Snakemake-a scalable bioinformatics workflow engine". *Bioinformatics* **28**:19 (2012), pp. 2520–2522.

[2] R. Dale, B. Grüning, A. Sjödin *et al.* "Bioconda: Sustainable and comprehensive software distribution for the life sciences". *Nature Methods* **15**:7 (2018), pp. 475–476.

[3] M. D. Robinson, D. J. McCarthy, and G. K. Smyth. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data". *Bioinformatics* **26**:1 (2010), pp. 139–140.

[4] M. Nowicka and M. D. Robinson. "DRIMSeq: A Dirichlet-multinomial framework for multivariate count outcomes in genomics". *F1000Research* **5** (2016), p. 1356.

[5] K. Rue-Albrecht, F. Marini, C. Soneson, and A. T. Lun. "iSEE: Interactive SummarizedExperiment Explorer". *F1000Research* **7** (2018), p. 741.

[6] N. Nowicka, J. Juranek, J. K. Juranek, and J. Wojtkiewicz. "Risk Factors and Emerging Therapies in Amyotrophic Lateral Sclerosis". *International Journal of Molecular Sciences* **20**:11 (2019), p. 2616.

[7] K. Kapeli, G. A. Pratt, A. Q. Vu *et al.* "Distinct and shared functions of ALS-associated proteins TDP-43, FUS and TAF15 revealed by multisystem analyses". *Nature Communications* **7**:1 (2016), pp. 1–14.

[8] J. P. Ling, O. Pletnikova, J. C. Troncoso, and P. C. Wong. "TDP-43 repression of nonconserved cryptic exons is compromised in ALS-FTD". *Science (New York, N.Y.)* **349**:6248 (2015), pp. 650–655.

[9] R. Edgar, M. Domrachev, and A. Lash. "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository". *Nucleic Acids Research* **30**:1 (2002), pp. 207–210.

7 Acknowledgements

I would like to thank everyone who supported me during my PhD in the last four years. Most importantly, I would like to thank Mark D. Robinson and Magdalini Polymenidou for offering me the shared PhD position in their groups. I am deeply grateful for the supervision, support and valuable feedback that you have given me throughout these years.

I also wish to acknowledge my committee members Prof. Mihaela Zavolan and Prof. Frédéric Allain for their input and helpful advice on my PhD during my committee meetings.

The Polymenidou group was always fun to work with and I learned a lot about cell culture, imaging and neurobiology. Thank you so much guys! My special thanks to Sonu Sahadevan for the great and exciting collaboration. We hit many obstacles along the way of the project, both experimental and computational, but our discussions always helped to find a solution.

I want to thank everyone in the Robinson group for the amazing and inspiring working environment. I really appreciate the positive attitude in the group and the scientific and especially non-scientific lunch discussions. Special thanks to Stephany Orjuela, Ruizhu Huang and Charlotte Soneson for the productive collaboration and help in conquering conda and snakemake.

Finally, I am very grateful to my family for their continuous support and believe in me. Visiting you always helped me to relax and to not worry about work so much. Last of all, I want to thank Alex for always cheering me up when I was down and for helping me with stupid writing and programming issues. Without you, I would have given up many times along the way!

8 Appendix

List of Figures